Generation of Substructure Identification Rules Using Feature-Combinations from Tandem Mass Spectra

K. J. Hart,* P. T. Palmer,‡ D. L. Diedrich, and C. G. Enke
Department of Chemistry, Michigan State University, East Lansing, Michigan, USA

Software to interpret tandem mass spectra, entitled Method for Analyzing Patterns in Spectra (MAPS), has been developed to provide substructure information for an automated compound identification system. This software consists of several program modules which manipulate databases of tandem mass spectra and substructure information, generate substructure identification rules, and apply these rules to the tandem mass spectra of unknown compounds to identify components of their structure. The MAPS rule generation program has been modified to generate rules based on specific combinations of spectral features that occur concertedly. False positives are drastically reduced by searching for "feature-combinations" that have 100% uniqueness with respect to a reference database of compounds. Recall is increased by the determination of multiple feature-combinations indicative of the presence of a given substructure. Strategies were developed in the algorithm for the discovery of feature-combinations that avoid the computation "explosion" that occurs when working with a large number of spectral features. The rules developed have the form: "IF feature-combination a (FC a) or FC b, . . . , or FC x, THEN substructure SSn is present." *(J Am Soc Mass Spectrom 1992, 3, 159–168)*

The tandem mass spectrometry (MS/MS) technique has been widely used to provide increased structural information for solving structure elucidation problems [1]. Evidence of this trend may be seen simply by examining a few issues of any mass spectrometry journal. Unfortunately, there has been a lack of user-friendly software to assist mass spectrometrists in evaluating the large numbers of tandem mass spectra that can be acquired for even relatively simple compounds. Thus, the Method for Analyzing Patterns in Spectra (MAPS) software was developed to provide the capability of correlating the presence of substructures in known compounds with spectral features observed in the tandem mass spectra of these compounds [2–4]. This software is a principal program module of the Automated Chemical structure Elucidation System (ACES), which is being developed to provide molecular structures for unknown compounds from tandem mass spectra [5, 6].

Two major approaches to automated structure elucidation utilizing tandem mass spectra have been explored [2]. The first of these approaches is the traditional spectral matching method where an unknown tandem mass spectrum is compared to a library of tandem mass spectra. The major problems in utilizing this method with tandem mass spectra are that it does not take full advantage of the extra dimension of structural information that MS/MS affords, and that it relies on the existence of a database of reproducible tandem mass spectra [2]. The second approach is to automate the interpretation of the tandem mass spectra to deduce the presence of structural features in organic compounds, and is presented here.

The MAPS system utilizes the complete MS/MS data sets of reference compounds to provide a list of characteristic features for use by a rule generation program. A complete MS/MS data set for a given compound consists of daughter spectra for all primary scan ions with an intensity greater than a specified threshold value. The spectral features that can be extracted from the MS/MS data set are primary ions, daughter ions, neutral losses, and specific parent ion-daughter ion pairs. Intensity information is not used by the current software and thus avoids the problem of irreproducible daughter ion intensities that has heretofore hindered the application of spectral matching routines to this problem. Standardization of MS/MS instrumentation and operating conditions has been explored by Martinez [7] but has yet to yield the requisite database of daughter spectra. The advantage

*Present address: Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, TN 37831-6365.
‡Present address: NASA-Ames, M/S 242-2, Moffett Field, CA 94035-1000.
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

of the MAPS approach is that any instrument or conditions can be utilized so long as they are used consistently within a database and a data importation routine is written to accommodate the data format of the instrument. A listing of the substructures present in the reference compounds is also required and can be automatically generated using a structure generation or manipulation program [8, 9]. Once the rules have been generated, a rule application program is used to identify the presence of substructures in an unknown compound based on the MS/MS data set of the unknown. This discussion focuses on the generation of the MAPS rules. A companion paper provides a discussion of the content of the MAPS rules and the application of the rules to unknown compounds.

Evolution of the MAPS Software

The MAPS software automatically generates the substructure identification rules for ACES and has been recently modified to provide rules with greater reliability and recall through the use of "feature-combinations" [8]. The reliability and recall of a rule are given by eqs 1 and 2. The goal of the ACES system is to obtain a single, definitive structure for an unknown or, at the very least, a set of candidate structures which is consistent with the structural information contained in the tandem mass spectra of the unknown. While 100% rule reliability is required to ensure that this goal is achieved, 100% recall is not an absolute requirement. A subset of all the possible substructure identifications can lead to a single structural candidate for an unknown. However, a larger overall recall for a set of rules can increase the probability that only a single structure will be obtained for an unknown.

$$\mathrm{REL}(\%) = \frac{\text{number of correct predictions} \times 100}{\text{total number of predictions made by a rule}} \quad (1)$$

$$\mathrm{REC}(\%) = \frac{\text{number of correct predictions} \times 100}{\text{total number of possible correct predictions}} \quad (2)$$

where REL is the rule reliability and REC is the rule recall.

In previous versions of MAPS, a substructure identification rule had the form:

> "IF the fraction of the features listed in this rule that are represented in the sample spectrum exceeds the value MF, THEN substructure 'SSn' is present."

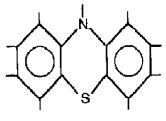
The "match factor" (MF) specifies the minimum fraction of rule clauses that had to "fire" (i.e., the minimum number of spectral features that had to be found in the tandem mass spectra of an unknown) for a substructure prediction to be made by the system. A rule for the **phenothiazine** substructure obtained from the previous version of MAPS is provided in Figure 1 and consists of nine clauses [5]. (Note that names appearing in bold type refer to a substructure definition and not a specific compound. For example, the compound phenothiazine consists of 12 carbons, one nitrogen, and one sulfur bonded as shown in Figure 1 with nine hydrogens occupying all the free valences shown in the **phenothiazine** substructure. The **phenothiazine** substructure definition shown in Figure 1 does not include specific substituents or a specific substitution pattern.) Each of the rule clauses in the rules generated by this version of MAPS is composed of a single tandem mass spectral feature. The spectral features that can be found in MAPS rules include primary ions ("P m/z"), daughter ions ("D m/z"), neutral losses ("NL u"), and parent-daughter pairs ("PD m/z m/z"). In this case, the parent ion-daughter ion pairs correspond to a single neutral loss in that single-collision conditions were used to acquire the daughter spectra. Daughter spectra were also acquired under multiple-collision conditions but kept in a separate database. In general, the PD features provided the highest degree of specificity for a substructure of all the spectral features used by the MAPS software (i.e., parent ions, daughter ions, neutral losses, and parent ion-daughter ion pairs). 1

Two database parameters, uniqueness (U) and correlation (C), were calculated to quantify the specificity and frequency of occurrence of a feature in the database with respect to a particular substructure. These parameters are listed for each feature in the rule shown in Figure 1 in the square brackets. Equations 3 and 4 were used to calculate these values.

```
IF   "D  (211.0)              [40,76] "  {F1}
and  "D  (209.0)              [40,76] "  {F2}
and  "D  (198.0)              [41,92] "  {F3}
and  "PD (198.0 -> 171.0)     [76,76] "  {F4}
and  "PD (198.0 -> 154.0)     [92,92] "  {F5}
and  "PD (197.0 -> 196.0)     [68,84] "  {F6}
and  "PD (197.0 -> 153.0)     [90,76] "  {F7}
and  "PD (196.0 -> 152.0)     [90,76] "  {F8}
and  "PD (70.0 -> 27.0)       [45,84] "  {F9}

THEN substructure phenothiazine is present.

(Umin = 40%, Cmin = 70%)
```



phenothiazine substructure

Figure 1. The initial MAPS rule obtained for the **phenothiazine** substructure with the uniqueness and correlation of each spectral feature shown in square brackets. Ion structures for the parent ions in this rule can be found in the following article [10].

MAPS rules were generated by filtering the set of all features in the reference database so that only those features which had an initial feature uniqueness (Ui) and initial feature correlation (Ci) values greater than specified minimum values (e.g., Ui = 40% and Ci = 70% as in Figure 1) were included in a particular substructure identification rule.

$$Ux(\%) = \frac{\text{number of compounds with SS}x \text{ and F}x \times 100}{\text{number of compounds with F}x} \quad (3)$$

$$Cx(\%) = \frac{\text{number of compounds with SS}x \text{ and F}x \times 100}{\text{number of compounds with SS}x} \quad (4)$$

where Ux is the spectral feature uniqueness, Cx is the spectral feature correlation, SSx is substructure X, and Fx is the spectral feature x.

The reliability and recall with respect to the reference database obtained for the **phenothiazine** rule are shown in Table 1 for a variety of match factors [5]. There is a clear trend in these data; increased reliability can be had only at the expense of recall using the early versions of MAPS. Specifically, 100% reliability can only be achieved for this **phenothiazone** rule if a match factor of 70% is used (and a resulting recall estimate of 77%). The incorporation of feature-combinations in MAPS, on the other hand, has provided a rule for **phenothiazine** that has both a reliability and a recall of 100% with respect to the reference database. The reliability and recall of this rule was also tested against 20 test compounds and 100% reliability and recall were obtained [10]. Although these figures of merit do not mean that a MAPS rule will predict the presence of the **phenothiazine** substructure with absolute certainty in a true unknown, they do show that the new MAPS software is utilizing the MS/MS data more effectively than previous versions of the software. It is expected, however, that the reliability and recall estimates will more accurately portray the effectiveness of the MAPS rules in analyzing true unknowns when a larger tandem mass spectral database is available to generate the rules. A detailed evaluation of the **phenothiazine** feature-combination rule is provided in the following article [10].

Table 1. Reliability and recall estimates obtained for the **phenothiazine** substructure identification rule generated by using the previous version of MAPS at several different match factors

MF (%)	Reliability (%)	Recall (%)
100	100	38
90	100	54
80	100	69
70	100	77
60	80	92
50	75	92
40	65	100

A "feature-combination" is a set of features which collectively have a uniqueness (with respect to the reference database) greater than or equal to any individual feature in the set for a particular substructure. A rule based on feature-combinations has the general form:

"IF ⟨all spectral features $f_{a1}, f_{a2}, \ldots$⟩ are present
or ⟨all spectral features $f_{b1}, f_{b2}, \ldots$⟩ are present
or ⟨all spectral features $f_{n1}, f_{n2}, \ldots$⟩ are present
THEN substructure SSn is present."

A substructure identification is made by using this type of rule when any of the rule clauses "fire" (i.e., the features in a feature-combination are found in the tandem mass spectra of an unknown) in that each feature-combination has 100% uniqueness, and therefore 100% reliability, for the specified substructure.

An example of a feature-combination rule generated for a substructure labeled "SS118" is shown in Figure 2. Substructure **SS118** is defined in Figure 3 along with several examples of reference database compounds that contain the **SS118** substructure. Note that three different drug classes are represented (i.e., opioids, barbiturates, and amphetamines). It was not realized that the **SS118** substructure was represented

IF	(PD 78.0 77.0 P2) and (PD 115.0 65.0 P1) and (D 71.0 P1)	100 37
OR	(PD 117.0 90.0 P2) and (PD 117.0 115.0 P2) and (PD 115.0 65.0 P1) and (NL 17.0 P1) and (D 31.0 P2)	100 37
OR	(D 119.0 P1) and (D 91.0 P1) and (NL 92.0 P1) and (NL 52.0 P1) and (D 63.0 P2) and (NL 77.0 P2) and (D 31.0 P2) and (D 42.0 P1)	100 37
OR	(D 119.0 P1) and (D 91.0 P1) and (NL 92.0 P1) and (NL 52.0 P1) and (D 63.0 P2) and (NL 77.0 P2) and (NL 27.0 P1) and (D 42.0 P1)	100 37
OR	(PD 78.0 77.0 P2) and (PD 117.0 115.0 P2) and (D 71.0 P1)	100 35
OR	(PD 78.0 77.0 P2) and (PD 117.0 115.0 P2) and (D 71.0 P2)	100 35
OR	(PD 78.0 77.0 P2) and (PD 115.0 65.0 P1) and (D 43.0 P1)	100 35
OR	(PD 117.0 90.0 P2) and (PD 117.0 115.0 P2) and (PD 115.0 65.0 P1) (NL 118.0 P2) and (D 31.0 P2)	100 35
OR	(PD 117.0 115.0 P2) and (PD 115.0 65.0 P1) and (D 119.0 P1) and (D 63.0 P1) and (D 31.0 P2)	100 35
OR	(D 119.0 P1) and (D 91.0 P1) and (NL 92.0 P1) and (NL 52.0 P1) and (D 63.0 P2) and (NL 77.0 P2) and (NL 31.0 P1) and (D 42.0 P1)	100 35

(...)

THEN substructure *SS118* is present.

REL = 100% and REC = 97% with respect to reference database

37 / 86 of the training set compounds have SS118
Total of 49 rule clauses with 100% reliability out of 9,437,681 checked
U_i = 40%, C_i = 40%, Cc = 30%, mass filter enabled, 406 initial features

NOTE: P1 refers to single collision conditions and P2 refers to multiple collision conditions as noted in the experimental section

Figure 2. MAPS feature-combination rule for the **SS118** substructure generated using reference tandem mass spectra acquired under both single- and multiple-collision conditions.

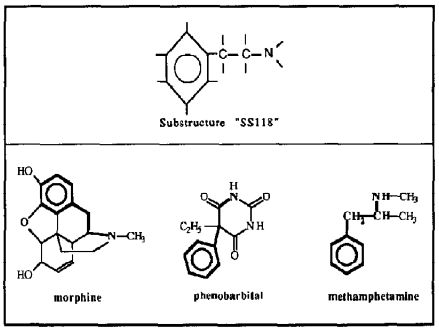


Figure 3. The "SS118" substructure definition with several examples of reference database compounds with this substructure highlighted with thick lines.

in three different structural environments until after a substructure search program was used to analyze the structures of the reference database compounds. This rule demonstrates several capabilities of the new MAPS program. First, note that the spectral features that comprise each feature-combination (or rule clause) are tagged to identify the collision conditions (i.e., P1, single collisions; or P2, multiple-collision conditions) utilized in acquiring the MS/MS reference spectra. Second, there are no primary mass spectral features in this rule because they were excluded in generating the training set. This feature may be useful if a database of fast atom bombardment (FAB) tandem mass spectra were available to eliminate some of the FAB matrix features from the training set. However, it should be noted that, if spectral features appear in the spectra of all the reference database spectra, the Ui values that the MAPS software calculates will be quite low. Third, only those neutral loss features that are tagged with "P1" are bona fide neutral losses because the P1 tag indicates single-collision conditions. The possibility that neutral losses with the "P2" tag were the result of multiple (and possibly different) neutral losses must be considered. One aspect of neutral losses that the MAPS software does not yet address is the automatic recognition of neutral losses from doubly charged parent ions. This limitation can lead to some unlikely neutral losses such as a neutral loss of 8 u. This neutral loss is, in fact, a neutral loss (i.e., with charge retention) of 16 u from a doubly charged parent ion.

Rule Generation Using Feature-Combinations

The rule generation process, which provides the substructure identification rules for the ACES system, is outlined in the schematic diagram provided in Figure 4. The GENT (GENerate Training set) program preprocesses several types of data to create a spectral feature and substructure array which is referred to as

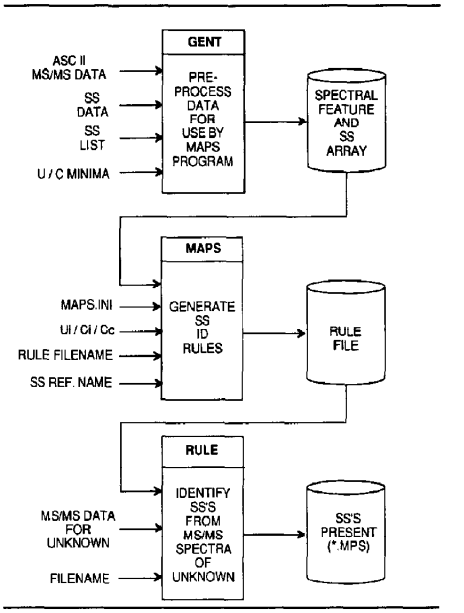


Figure 4. Important input and output from the programs comprising the MAPS software.

the training set. The MAPS program then generates substructure identification rules from the training set by using a number of program parameters to control the generation of rules. The MAPS program writes rules to a file for subsequent use by the RULE program. The RULE program can then be used at any time to apply the rule(s) to the MS/MS data of an unknown compound and display the names of the substructures identified as present in the unknown. The substructures thus identified are also written to a results file for subsequent use by an automated structure generator within the ACES system [8]. The new version of MAPS has been written in C and is currently running on a VAXstation 3200 computer (Digital Equipment Corp., Marlboro, MA).

Creation of the Training Set

There are several useful program parameters that control how the GENT program creates the training set. These parameters are useful in eliminating features of low intensity that may not be reproducible, controlling the tandem mass spectra that are included in the training set by the intensity of the parent ion, and excluding features that have small uniqueness and correlation for any of the defined substructures. The latter features are of little value for the MAPS program. A brief description of the functioning of the GENT program is provided here to ensure that it is

clear how the training set, which is the basis of the MAPS rules, is formed.

The program begins by inputting the primary mass spectrum and daughter spectra associated with all the reference compounds. Each spectral feature found in the collection of spectra is followed by a list of compound names that contain the feature so the final result is very much like an inverted database. The tandem mass spectral features include primary ions "P (m/z)", daughter ions "D (m/z)", neutral loss "NL (u)", and parent-daughter pairs "PD (m/z m/z)". Three user-definable intensity thresholds, PTHRESH (primary mass spectra), DTHRESH (daughter spectra), and PDTHRESH (parent ion selection), can be used to exclude spectral features with weak intensities from the training set. For example, a 1% PDTHRESH threshold was used during data acquisition to select ions from the primary mass spectrum for collision-induced dissociation and subsequent acquisition of daughter spectra. Typical values for PTHRESH, DTHRESH, and PDTHRESH are 0.1%, 1%, and 1%, respectively. Another program parameter, MINWF, defines a minimum number of compounds, typically three compounds, in which a spectral feature must be found before the feature is included in the training set. The current reference database contains 100 compounds, many of which are regulated drugs.

After the spectra have been loaded, the GENT program prompts for the name of a file containing a list of substructures contained in each reference compound. The Automated Structure Library Search (ASLS) program [8] creates the substructure list required by the GENT program. This program is a modified version of the STRCHK program originally developed at Stanford University as part of the DENDRAL project [9]. The ASLS program automatically checks the structures for all the reference compounds (contained in a library of connectivity tables) against a substructure library [8] and writes the results to a file used by the GENT program. The connectivity tables can be extracted from libraries of chemical structures or input manually using a program such as GENOA, an interactive structure generator [9]. The current substructure library contains 161 substructure definitions. Of these substructures, 121 are represented in at least one compound in the reference database.

Another set of parameters also limits the spectral features included in the training set. These parameters are a minimum feature uniqueness (U) and a minimum feature correlation (C). Typical values for U and C are 10%. Thus, all spectral features in the training set have at least 10% uniqueness and correlation with respect to at least one substructure in the substructure library.

An excerpt from a training set file generated using 40 compounds and two substructures is provided in Figure 5. The file begins with a list of the defined substructures and the mass of each substructure. The

```
SS18 80.0
SS50 96.0

GMR10    SS18 SS50
GMR11    SS18 SS50
GMR12    SS18 SS50
...
(and so on for a total of 40 compounds)

(PD 207.0 192.0)      0001000010000000011000100000000100000000
(PD 190.0 189.0)      1010001000000000000000000000010100010000
(PD 185.0 184.0)      0011000000100000000000000000000100100000
(PD 207.0 179.0)      0001001010000000010000000000000100000000
...
(NL 77.0)  0100000101000010111000000000100000011111
(D 77.0)   1111101111111011111000111000111110111111
(P 77.0)   1111111111111011111111111111111111111111
...
(and so on)
```

Figure 5. An excerpt from the GENT output file produced by using 40 reference database compounds and two substructure definitions.

next section of the file contains lists of the name of each reference compound and the substructures found in the compound. The last section is a list of spectral features that meet the threshold, MINWF, and U/C criteria. Each spectral feature is followed by a series of bits which identify the reference compounds which yield the feature. In that 40 compounds were used for this example, there are 40 bits in the string following each spectral feature in Figure 5. GENT required ~ 30 min of central processing unit time to process 100 primary and 5749 daughter spectra of 100 reference compounds. The GENT output file contains all of the spectral and substructure data required by the MAPS program.

Search Strategies for Generation of Feature-Combinations

The current MAPS program includes several enhancements not found in previous versions of MAPS. The most important of these enhancements is the ability to rapidly generate feature-combination rules. Two important strategies have been implemented to avoid the "combinatorial explosion" problem that is often observed when a combinatorial method is used. For example, exhaustive generation of all combinations of the 964 spectral features that have at least 10% uniqueness and correlation for the **barbiturate** substructure would result in $2^{964} - 1$ different feature-combinations, each of which would have to be tested against the training set. The computation time required to complete this operation would be measured in years. Obviously, strategies had to be developed to limit the size of the feature-combination search space.

Pruning the Feature List. The first strategy used in the new software is to limit the initial number of spectral features passed to the feature-combination generator by specifying a minimum initial uniqueness (Ui) and correlation (Ci) value. The training set is already re-

duced to those spectral features with at least 10% uniqueness and correlation for at least one substructure. However, higher values of Ui and Ci are necessary for reliable rule generation. The U/C values used by GENT were kept low to allow several different sets of Ui and Ci values to be explored within MAPS without having to regenerate the training set. The number of initial features obtained for any substructure decreases with increasing Ui and Ci values as shown in Figure 6. For example, 56 features were obtained for the **barbiturate** substructure using a value of 30% for Ui and Ci (down from the 964 features with at least 10% Ui and Ci). The number of features selected for the **barbiturate** substructure can be further reduced with higher values of Ui and Ci.

Different sets of Ui and Ci values can also change the nature of the spectral features included in the initial feature list. For example, there are two sets of Ui/Ci values that yield 11 spectral features for the **barbiturate** substructure (i.e., 20%/80% and 70%/30%). In the first set of values, Ui is low as Ci is high while the opposite is true for the second set of values. Almost all (10 out of 11) of the spectral features selected using the low Ui and high Ci values are primary, daughter, and neutral loss features. These features tend to be more general than parent ion–daughter ion pairs and thus have lower Ui values. However, the features selected when Ui was set high and Ci low were only parent ion–daughter ion pairs which tend to be more specific than the other types of spectral features observed in tandem mass spectra and thus tend to have high U values.

Pruning Feature-Combination Search Branches. A second strategy used in the new software removes nonproductive search branches as early as they can be detected during rule generation. A feature-combination is constructed by adding features from the list of initial features until no false positives are observed (i.e., U = 100%) or the recall for the combination falls below a specified minimum value, Cc. A search branch is terminated when the correlation of a candidate feature-combination falls below the Cc value or when the compound list associated with a candidate feature-combination is composed of compounds already identified a specified number of times (called the "HITS" limit) by previously generated feature-combinations. Additional computational efficiency is obtained using the new MAPS algorithm which eliminates false positives (an integer value) rather than increasing uniqueness (a floating-point value). Integer operations can be performed on the computer system used for this work more rapidly than floating point operations.

The new MAPS algorithm is illustrated in Figure 7 which shows the generation of candidate feature-combinations for the **barbiturate** substructure from a list of 23 initial spectral features. These spectral features were selected using Ui = 35% and Ci = 30%. The minimum feature-combination correlation, Cc, was set to 30%. The first feature-combination shown in Figure 7 was simply the first feature in the feature list. The second feature-combination was composed of two features from the single-feature list with a combined uniqueness of 100% and correlation of 30%. This feature-combination was included in the **barbiturate** rule because it had 100% uniqueness for the **barbiturate** substructure and met the Cc value. An overall recall value was also provided by the software for feature-

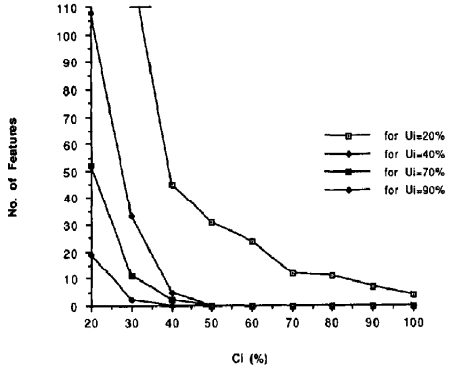


Figure 6. Number of initial features versus Ci for different values of Ui.

```
1  (PD 98.0 80.0 P1)  U=85% C=46% more
2  (PD 98.0 80.0 P1) (PD 169.0 97.0 P1)  U=100% C=30% 30% recall
3  (PD 98.0 80.0 P1) (PD 98.0 27.0 P1)  U=100% C=38% 46% recall
4  (PD 98.0 80.0 P1) (PD 98.0 28.0 P1)  U=100% C=38% 46% recall
...
5  (PD 98.0 80.0 P1) (D 16.0 P1)  U=100% C=7% low
6  (PD 98.0 80.0 P1) (PD 98.0 70.0 P1)  U=85% C=46% worse
7  (PD 98.0 80.0 P1) (PD 160.0 104.0 P1)  U=0% C=0% low
...
8  (PD 98.0 27.0 P1)  U=83% C=38% hits
9  (PD 98.0 28.0 P1)  U=75% C=46% more
10 (PD 98.0 28.0 P1) (PD 169.0 126.0 P1)  U=100% C=23% low
11 (PD 98.0 28.0 P1) (PD 98.0 44.0 P1)  U=100% C=38% hits
...
12 (PD 54.0 27.0 P1) (NL 111.0 P1)  U=71% C=38% more
13 (PD 54.0 27.0 P1) (NL 111.0 P1) (PD 97.0 55.0 P1)  U=80% C=30% hits
14 (PD 54.0 27.0 P1) (NL 111.0 P1) (PD 69.0 39.0 P1)  U=100% C=38% back

KEY:
more:  feature-combination requires more features to achieve 100% U
low:   feature-combination correlation lower than allowed minimum;
       terminate search branch
worse: adding last feature to feature-combination had no effect on U or
       lowered U; terminate search branch
hits:  all compounds identified by this feature-combination have been
       identified a sufficient number of times by other feature
       combinations
back:  a superfulous feature has been detected (i.e. removal of a previously
       added feature in a feature combination with 100% U does not affect
       U); terminate search branch
```

Figure 7. Candidate feature combinations generated for the barbiturate substructure by using a set of 23 initial spectral features selected using Ui = 35% and Ci = 30%.

combinations that were included in a rule as shown for the second feature-combination in Figure 7.

Several labels were used to tag candidate feature-combinations. These labels indicated how the feature-combination search was progressing. The definitions of these labels are listed in the figure. Also notice that each feature in the combinations listed in Figure 7 has an associated database label. This label identifies the database from which a feature originates. For example, the user-defined label "P1" indicates that all features associated with this label were derived from the database of single-collision spectra. This convention allows the MAPS software to generate and utilize rules composed of spectral features derived from spectra acquired under different operating conditions.

Using MAPS with Feature-Combinations

Although the MAPS software was developed for use in the ACES system, it has several features that are useful whenever a series of tandem mass spectra is examined for spectral feature/substructure correlations. Several of these features are accessed through the program commands and parameters discussed below.

The MAPS program begins with a prompt for the training set filename. The program then provides a summary of the training set that includes the number of compounds, substructures, and spectral features in the training set. The primary and daughter scan conditions are also provided. These conditions can vary for different training sets (e.g., electron ionization or negative chemical ionization for primary scan features and single or multiple collision for daughter scan features). The substructure reference names are also provided as they are used to identify the substructures in the program.

The MAPS software is command-line driven. The first step in using the code is the creation of a feature list. This operation is accomplished using the SINGLEU, SINGLEC, and SINGLE commands (i.e., a single feature uniqueness and correlation of 40% and 70%, respectively, for the SS132 substructure). The MAPS software then reports the number of spectral features which meet these values. The feature list can be further manipulated by sorting the list by uniqueness (USORT), correlation (CSORT), or mass (MSORT). There are also several user-selectable mass-to-charge ratio masks implemented in the new MAPS program. The feature list can be displayed using the print command.

The minimum acceptable Cc is also required by the MAPS program. Once the program parameters and the features list have been set, the COMBINATION command is used to generate feature-combinations. A summary of the feature-combinations generated is displayed to the terminal after each COMBINATION command is executed. In this example, six feature-combinations (or rule clauses) were obtained for the substructure labelled "SS132" (phenothiazine) with an overall recall of 100% with respect to the reference database. Only those feature-combinations that have 100% uniqueness for the indicated substructure are written to the rule file.

Applying MAPS Rules to Unknowns

The MAPS rules have two uses. They can be used to identify the most important features or combinations of features that are associated with particular compound classes (i.e., substructures). In some cases, individual feature-combinations can lead to the identification of tandem mass spectral features that are indicative of combinations of substructures [10]. The MAPS rules can also be applied to the MS/MS data sets of unknown compounds to identify the presence of substructures within the molecular structure of the unknowns. This latter task is accomplished using the RULE program.

The RULE program compares the feature-combinations found in a rule file to those present in the MS/MS data set of an unknown. Three filenames are required to analyze an unknown with the current version of MAPS (i.e., the rule, primary scan, and daughter scan filenames). The RULE program lists the number of rule clauses (feature-combinations) that "hit" for the unknown and the total number of clauses in the rule. For example, the RULE program was used to check for the presence of the **phenothiazine** and **barbiturate** substructures in 20 test compounds (i.e., compounds not present in the reference database). The total time required to load the tandem mass spectra of the test compounds and to apply the two MAPS rules was approximately 3 min. When the VERBOSE parameter is enabled in the MAPS initialization file, the rule clauses that hit on an unknown are displayed to the terminal so the user can examine the spectral features in the rule clauses. A RULE results file is written which contains a list of substructures identified as present in an unknown for use by the automated structure generator in ACES.

The criterion for an identification to be made by the RULE program is the matching of at least one rule clause with spectral features in the tandem mass spectra of an unknown. One rule clause is sufficient to make an identification because each rule clause has 100% reliability for the indicated substructure with respect to the reference database. The size of the reference database and the number of spectral features in a feature-combination affect the validity of this method of rule application. These considerations are addressed in the next section.

Results and Discussion

The ability to monitor the false positives, the correlation, and the list of reference compounds associated

with candidate feature-combinations and to terminate unproductive search branches significantly reduces the feature-combination search space. For example, 18 spectral features were selected for the **barbiturate** substructure using Ui = 30%, Ci = 35%, and the current training set. Exhaustive generation of all combinations of these features would result in 262,143 feature-combinations being checked for 100% uniqueness and inclusion in a rule for the **barbiturate** substructure. The new MAPS program with the HITS parameter disabled and the Cc value set to 1%, checked only 927 feature-combinations (~ 0.4% of the possible combinations). A Cc value of 1% was used to prevent any search branches from being terminated due to low, but nonzero, correlation. The rule generated under these conditions contained 192 clauses and had 76% recall for the barbiturate reference compounds. Rules generated using the HITS and Cc parameters further reduces the number of feature-combinations checked by the MAPS software without reducing recall as shown in Table 2. The MAPS program is quite efficient in checking feature-combinations. The number of feature-combinations checked per second ranges from 300 to 600 on a DEC VAXstation 3200 minicomputer (single user).

It is noteworthy that the only point at which useful information could be lost using the HITS and Cc parameters is when a feature-combination is not included in a rule because these limits were exceeded. Although the excluded feature-combination would not have improved the identification of the substructure in any of the reference compounds, it is possible that it might enhance recall with an unknown compound. This situation was not observed for the **phenothiazine** and **barbiturate** rules that were applied to 20 test compounds [10]. These rules had 100% recall with respect to test compounds so no loss of information was observed using the HITS and Cc parameters. Recall with respect to the reference database is also maintained using these parameters as shown in Table 2. The main advantage of using the HITS and Cc parameters is to provide more compact rules (fewer rule clauses) for use with tandem mass spectra of unknown compounds.

Table 2. Feature-combination generation results for various values of the HITS and Cc MAPS program parameters

MAPS parameters	Number of rule clauses	Recall (%)	Number of feature-combinations	Time (s)
Cc = 1% HITS disabled	192	76	927	6
Cc = 30% HITS disabled	81	76	561	2
Cc = 1% HITS = 10	43	76	397	1
Cc = 30% HITS 10	30	76	348	1

Performance

Two evaluations of the MAPS rules have been performed to investigate the performance of the new MAPS software. These include a comparison of the spectral features selected for the **phenothiazine** and **barbiturate** substructures using the Ui/Ci criteria and application of the resulting feature-combination rules to a series of test compounds, as reported in the next article [10]. One of the major conclusions of these evaluations was that optimal rule reliability was achieved using a MINF value > 2. The MINF parameter is an artificial means to increase the number of spectral features in a feature-combination. For example, all feature-combinations (rule clauses) in a rule generated using a MINF value of 4 must contain at least four spectral features even if feature-combination with two or three features has 100% uniqueness with respect to the reference database. Consequently, any substructure prediction made by these rules will be based on the presence of at least four spectral features. The inherent reliability of such predictions is higher for these rules because each prediction is based on an increased amount of spectral information and therefore, has greater specificity for the indicated substructure.

Effect of Size and Composition of the Reference Database

If possible, rule reliability should be increased by increasing the number and variety of compounds in the reference database. A reference database with a greater variety of substructure combinations will ensure that more false correlations will be eliminated from the rules and each feature-combination will have greater statistical validity when employed against unknowns. As the variety of compounds in the database increases, the number of spectral features in a feature-combination should also increase for many substructure rules since these feature-combinations will have to be more specific to achieve the necessary 100% uniqueness value to be included in MAPS rules. The number of environments that a substructure is represented in will also tend to increase to a maximum value with increasing numbers of reference compounds. Thus, the MAPS rules will also be more complete when generated using a larger database.

A reference database with a greater variety of compounds will also reduce potential errors due to cross-correlations [10]. The empirical nature of the MAPS algorithm demands that a substructure be represented in the reference database in a number of different structural environments so the substructure of interest can be effectively isolated from all other substituents in the reference compounds. Fortunately, substructure cross-correlation can be detected in advance by calculating cross-correlation coefficients for all substructures represented in the reference database

with respect to a reference substructure. These coefficients are obtained by issuing the SSCROSS command to the MAPS prompt and specifying the name of the reference substructure. It is expected that rule reliability and recall will asymptotically approach a maximum as more compounds are added to the reference database because more examples of the same combination of features will not contribute new information to the rules. When this maximum is observed, the identification of substructures in compounds not yet in any database will be practical.

Cross-correlation Between Substructures

Substructure cross-correlation is detrimental to rule reliability only when cross-correlated substructures are unrelated. For example, the **phenothiazine** substructure is 100% cross-correlated with the **thiolphenyl** substructure. This cross-correlation does not adversely affect the generation of a phenothiazine rule because the **thiolphenyl** substructure is an integral part of the **phenothiazine** substructure. It should be expected, therefore, that some fragments that are due to the **thiolphenyl** substructure might appear in the **phenothiazine** rule. However, the large cross-correlation between the **t-butyl** and **phenol** substructure (i.e., 93%) and between the **t-butyl** and the **benzyl** substructure (i.e., 100%) in our reference database are examples of detrimental cross-correlation. These cross-correlations are due to the composition of the reference database. All of the **t-butyl** containing compounds are phenols or contain a benzylic carbon. The only way to reduce this cross-correlation is to add reference compounds that contain the **t-butyl** substructure and not the **phenol** or **benzyl** substructures. Detrimental cross-correlation can also be detected by using the FCROSS command which lists the uniqueness and correlation values for each feature in the feature list with respect to a specified substructure. Inasmuch as detrimental cross-correlation can be detected using the cross-correlation coefficients, new reference compounds can be identified and added to the reference database to reduce the cross-correlation. Alternatively, spectral features can be manually deleted (i.e., using the MAPS DELETE command) from the list of initial features using the cross-correlation coefficients or "chemical intuition" as a guide.

Another cross-correlation coefficient can be calculated that can be quite useful for supplying the structure generator with more specific substructure constraints. It has been observed that the reference compound list for some feature-combinations generated for the **barbiturate** substructure has a high degree of correlation with other substructures [10]. The **barbiturate** substructure provides a good example of this phenomenon because the fragmentation of barbiturates often incorporates much of the **barbiturate** substructure and a side chain (e.g., 5-phenyl barbiturates). Specifically, several spectral features (in this case, ions) that incorporate a phenyl side chain with the barbiturate substructure can combine with other **barbiturate** features to provide a feature-combination specific for 5-phenyl barbiturates. Thus, when this feature-combination is found in the tandem mass spectra of an unknown, not only can the **barbiturate** substructure be inferred to be present in the unknown but the **phenyl** substructure as well.

Limitations

The limitations of the MAPS software include the lack of "self-optimizing" algorithm and the range and purity of reference compounds that can be analyzed by using current MS/MS instrumentation. In the first case, the MAPS code itself is limited because optimal program parameters vary among substructures. Thus, it is difficult to establish a set of parameters that can be used to generate an entire rulebase. Currently, rules must be generated and evaluated on an individual basis. In the second case, the acquisition time required to obtain a MS/MS data set of the reference compounds limits the separation techniques that can be used to purify the reference standards.

An additional limitation involves the range of unknowns that can be analyzed by the ACES system (which includes the MAPS software). This limitation derives from the assumption which must be made by the automated structure generator that all substructure identifications made by the RULE program be assigned to one component. This assumption is valid for pure unknowns but not for the more often encountered mixture. If complete structure elucidation is not required, however, the MAPS rules can be used for substructure screening of mixtures. These limitations can be substantially reduced with further research. First, a set of heuristics can be added to the code to allow self-optimization. Several sets of program parameters may be tried using heuristics as a guide to provide optimal rules for all substructures. Second, development of a MS/MS instrument capable of acquiring the complete MS/MS data set on the chromatographic time scale will assist in the acquisition of reference spectra and allow the ACES system to properly assign substructure predictions to the specific components of a mixture.

Conclusion

The MAPS software provides a number of significant advantages for structure elucidation by using tandem mass spectra. These advantages include a comprehensive set of software tools for manipulating MS/MS and substructure data, the ability to generate substructure identification rules based on feature-combinations with greatly increased reliability and recall as shown for the **phenothiazine** substructure, and the ability to readily analyze the substructure content of an unknown by using the RULE program. Although

the current software is not self-optimizing, which would allow rulebases containing many rules to be easily generated, the addition of heuristics to the MAPS program in order to guide the selection of optimal program parameters should resolve this problem. Further expansion of the current reference database and creation of ancillary databases (e.g., using negative chemical ionization so rules based on negative ions can be generated) will provide a robust expert system for structure elucidation using tandem mass spectra. Advances in MS/MS instrumentation to provide full MS/MS data sets on the chromatographic time scale should expand the range of unknowns that the ACES system can readily analyze.

Acknowledgments

The authors wish to thank Adrian Wade and Chris Weaver for their contributions to this project. This work was 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. 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.
2. Palmer, P. T., Ph.D. thesis; Michigan State University: East Lansing, MI, 1988.
3. Palmer, P. T.; Wade, A. P.; Hart, K. J. Enke, C. G. *Talanta* **1988**, *36*, 107-116.
4. Wade, A. P.; Palmer, P. T.; Hart, K. J.; Enke; C. G. *Anal. Chim. Acta* **1988**, *215*, 169-186.
5. Hart, K. J.; Enke, C. G. In *Proceedings of the Symposium on Chemometrics and Intelligent Laboratory Automation*, Canadian Chemical Conference, Victoria, BC. *Chemometrics Intelligent Lab. Syst.* **1990**, *8*, 293-302.
6. Enke, C. G.; Wade, A. P.; Palmer, P. T.; Hart, K. J. *Anal. Chem.* **1987**, *59*, 1363A-1371A.
7. Martinez, R. I. *Rapid Commun.* Mass Spectrom. **1989**, *3*, 127-129.
8. Hart, K. J. Ph.D. thesis; Michigan State University: East Lansing, MI, 1989.
9. Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. *J. Org. Chem.* **1981**, *46*, 1708-1718.
10. Hart, K. J.; Wade A. P.; Nourse, B. D. Enke, C. G. *J. Am. Soc. Mass Spectrom.* **1992**, *3*, 169-180.