




This is to certify that the
dissertation entitled
ALGORITHMS FOR INTERPRETATION OF MS/MS DATA FOR
CHEMICAL STRUCTURE ELUCIDATION

presented by

Peter T. Palmer

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in Chemistry


Major professor

Date June 29, 1988



RETURNING MATERIALS:
Place in book drop to
remove this checkout from
your record. FINES will
be charged if book is
returned after the date
stamped below.

FEB 18 1993

MAY 12 1993

**ALGORITHMS FOR INTERPRETATION OF MS/MS DATA FOR
CHEMICAL STRUCTURE ELUCIDATION**

By

Peter T. Palmer

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Chemistry

1988

ABSTRACT

ALGORITHMS FOR INTERPRETATION OF MS/MS DATA FOR CHEMICAL STRUCTURE ELUCIDATION

By

Peter T. Palmer

Several pattern recognition and artificial intelligence methodologies have been developed for structure elucidation from MS and MS/MS data. Together, they form an integrated set of software tools called ACES (Automated Chemical structure Elucidation System). Components of ACES include MAPS (Method for Analyzing Patterns in Spectra), which is used for substructure identification, the MFG (Molecular Formula Generator) program, which is used for generation of molecular formulae, and GENOA, which is used for structure generation.

A set of standard conditions has been identified and used to collect a database of MS/MS spectra. This database is used by MAPS to identify spectral feature/substructure relationships. These relationships are utilized to formulate inclusion and exclusion rules, which can be used to predict the presence and absence of substructures in unknowns. The different criteria which exist for predicting the presence and absence of substructures have been identified and employed for the optimization of both rule types. The resulting rules are characterized by high recall and low false positives, and thus can be of great utility in the structure elucidation of unknowns. A heuristic search algorithm has been

developed to identify combinations of individual spectral features for even more reliable substructure identification.

As a further aid for structure elucidation, a methodology has been developed to facilitate the determination of molecular formulae through the use of unit resolution MS and MS/MS data. This methodology exploits MS and MS/MS isotopic ratio data and substructures identified by MAPS to formulate elemental composition constraints. The composition data and molecular weight information are then entered as input to the MFG program which generates all possible formulae consistent with these constraints. Given sufficient constraints, the list of candidate formulae can often be reduced to a single formula.

ACES is the first automated structure elucidation system based on MS/MS data. As opposed to spectral matching methods, ACES is an interpretive approach which does not require the spectra of a given unknown to exist in any library. In principle, an unlimited number of compounds can be identified through the use of a rule base for a few hundred substructures.

**Copyright by
Peter T. Palmer
1988**

If the greatest achievement is incomplete, then its usefulness is unimpaired.

Lao Tzu
Tao Te Ching

The time will come when diligent research over long periods will bring to light things which now lie hidden. ... There will come a time when our descendants will be amazed that we did not know things that are so plain to them ... Many discoveries are reserved for ages still to come, when memory of us will have been effaced. Our universe is a sorry little affair unless it has in it something for every age to investigate ... Nature does not reveal her mysteries once and for all.

Seneca
Natural Questions

We do not ask for what useful purpose the birds do sing, for their song is their pleasure since they were created for singing. Similarly, we ought not to ask why the human mind troubles to fathom the secrets of the heavens. ... The diversity of the phenomena of Nature is so great, and the treasures hidden in the heavens so rich, precisely in order that the human mind shall never be lacking in fresh nourishment.

Johannes Kepler
Mysterium Cosmographicum

The essential point in science is not a complicated mathematical formalism or a ritualized explanation. Rather the heart of science is a kind of shrewd honesty that springs from really wanting to know what the hell is going on!

Saul-Paul Sirag

That which does not kill us makes us stronger.

Nietzsche

ACKNOWLEDGEMENTS

I would like to extend a very special thanks to my advisor and mentor, Dr. Chris Enke. It has been an honor and a pleasure to work with Chris, and his guidance, acumen, and creative insight were invaluable. I would also like to thank the other members of my committee, Dr. J. Throck Watson, Dr. W. R. Reusch, and Dr. C. C. Sweeley, for their time, guidance, and input. I would like to acknowledge Hugh Gregg, Phil Hoffman, Carl Beckner, Ann Giordani, and Kevin Cross, who laid the foundations for this work. I also extend my gratitude to my coworkers in the structure elucidation group, Kevin Hart and Dr. Adrian Wade, for this project was unquestionably a joint effort. I extend special thanks to Adrian, Bruce Newcome, Mark Bauer, Adam Schubert, Mark Victor, and Norm Penix, who represented a wealth of knowledge and experience and were always willing to answer my questions.

I certainly cannot mention all the people who made my stay here more enjoyable, but I would like to thank some really special friends of mine who made these last few years really memorable: Steve Johnson and Mike Werner (may they finally achieve the perfect partial pressure), Paul and Theresa Kraus (and our good friend Gandalf), Mike Kristo (Mr. Gusano), Keiji Asano (the Frisco Kid), Karen Light (no, a Bud Light), Kris Kurtz (the dockcrawler), and Jon Bon Wahl (OYI). I would like to thank the rest of my colleagues in the Enke group for their companionship and support and I wish them the best of luck in their future endeavors. I

also acknowledge El Azteco restaurants for allowing me to willfully destroy my stomach lining, Rick's American Cafe for exposing me to some truly bodacious blues and rock and roll, Labatt's Porter, Captain Morgan, Jose Cuervo, Electric Larry, Danga (the patron god of video games and grad students), and all other deities recognized by the God of the Month Club.

I acknowledge support from grants from the National Institutes of Health, a L.L. Quill memorial fellowship, summer fellowships from BASF and Dow, and a Fire Retardants Chemical Association memorial scholarship. Thanks are due due Finnigan MAT for the loan of and support for the Xerox 1108, which was invaluable for the development of MAPS. Special thanks go to my chemistry professors at Canisius College, especially Dr. Joe Bieron and Dr. Ray Annino, for providing me with the knowledge and motivation to succeed in graduate school. I would like to thank my brother Mike, Pete Chlebek, John Aker, and my relatives for their love and support. Lastly, and most importantly, I would like to thank my parents, Dan and Mary Ann, who were always there, never doubted my abilities, and shared my dream. It has finally become reality.

TABLE OF CONTENTS

LIST OF TABLES.....	xi
LIST OF FIGURES.....	xiv
 CHAPTER 1: STRUCTURE ELUCIDATION FROM MS AND MS/MS DATA.....	 1
Introduction.....	1
Direct Database Methods.....	2
Indirect Database Methods.....	6
Pattern Recognition Approaches.....	6
Artificial Intelligence Approaches.....	9
Spectral Simulation Approaches.....	14
Utility of MS/MS Data for Structure Elucidation.....	15
The Structure Generator (GENOA).....	21
Original System for Structure Elucidation using MS/MS Data.....	22
An Automated Chemical Structure Elucidation System using MS/MS Data.....	38
References.....	43
 CHAPTER 2: MS/MS SPECTRAL LIBRARY ACQUISITION.....	 47
MS Libraries.....	47
MS/MS Libraries.....	51
Instrumental Parameter Effects on Daughter Spectra.....	53
Effects of Collision Gas Pressure and Collision Energy on Daughter Spectra of the Phthalate Anhydride Ion.....	58
Effects of Collision Gas Pressure and Collision Energy on Daughter Spectra of Alkyl Ions.....	71
Criteria for Collecting a Database of MS/MS Spectra....	78
Standard Conditions.....	78
Data Collection.....	84
Transfer of MS and MS/MS data into the MAPS Database...	85
The Multidimensional Database.....	85
The MAPS Database.....	86
Requirements for Automatic Entry of Data into the MAPS Database.....	88
Linking the PDP-11 and the Xerox 1108.....	89

Front End for Transfer of Data from the PDP-11 to the Xerox 1108.....	90
Back End for Transfer of Data from the PDP-11 to the Xerox 1108.....	92
Automatic Entry of Data into the MAPS Database....	92
The Current MAPS Database.....	92
References.....	94
 CHAPTER 3: THE MAPS SOFTWARE.....	 97
Introduction.....	97
Choice of a Development Tool for MAPS.....	100
KEE.....	101
LISP.....	102
Generation of Substructure Identification Rules via MAPS.....	105
Construction of a Training Set.....	106
Construction of Feature and Substructure Bucket Lists.....	107
Correlation of Features with Substructures.....	109
Filtering the Rules to Retain only Relevant Features.....	111
Evaluation of Unknowns via MAPS.....	113
Evaluation of Rule Performance.....	115
References.....	126
 CHAPTER 4: MODIFICATIONS TO THE MAPS SOFTWARE.....	 128
Introduction.....	128
Modifications to the Rule Generation Process.....	130
Use of Correlation and Uniqueness Filters in Rule Generation.....	131
Use of Intensity Data in Rule Generation.....	136
Modifications to the Rule Application Process.....	141
Identifying the Presence and Absence of Substructures.....	141
Development of a Weighting Function which uses Uniqueness Factors.....	144
Exclusion Rule Optimization.....	145
Inclusion Rule Optimization.....	151
The Substructure Hierarchy.....	158
Conclusions.....	161
References.....	164
 CHAPTER 5: AUTOMATED GENERATION OF MS AND MS/MS SPECTRAL FEATURE COMBINATIONS FOR RELIABLE SUBSTRUCTURE IDENTIFICATION.....	 165
Introduction.....	165
Addition of Multiple Daughter Ion and Neutral Loss Feature Types to the Rule Generation Process.....	168

Search Techniques.....	176
Finding Reliable Feature Combinations.....	177
Results and Discussion.....	184
Conclusions.....	192
References.....	194

CHAPTER 6: PROGRAMS FOR MOLECULAR FORMULA DETERMINATION EMPLOYING DATA FROM UNIT RESOLUTION MS/MS SPECTRA..... 195

Introduction.....	195
Experimental.....	198
Algorithms.....	200
Applying Constraints to a Molecular Formula Generator.....	200
Determining the Number of Carbon Atoms.....	201
Identifying Substructures, Ion Structures, and Neutral Losses.....	205
Examples of Molecular Formula Determination.....	206
Example 1 - 1,2-benzene-dicarboxylic acid, di-cyclohexyl ester.....	207
Example 2 - Eicosane.....	211
Example 3 - 1,2-benzene-dicarboxylic acid, di-n-octyl ester.....	215
Conclusions.....	220
References.....	221

CHAPTER 7: APPLICATION OF ACES TO SEVERAL UNKNOWN..... 223

Introduction.....	223
MAPS Results.....	223
MFG Results.....	231
GENOA Results.....	233
Conclusions.....	234

CHAPTER 8: SUGGESTIONS FOR FUTURE WORK..... 236

Introduction.....	236
Expansion of the Training Set.....	237
Improvements to MAPS.....	238
Improvements to the MFG Program.....	251
Improvements to GENOA.....	252
Further Development of ACES.....	252
References.....	258

LIST OF TABLES

1.1	Match factor definitions.....	31
1.2	Match of m/z 105 daughter spectra of 1,2-benzene-dicarboxylic acid, di-n-octyl ester to m/z 105 daughter spectra from the reference database.....	32
1.3	Match of m/z 149 daughter spectra of 1,2-benzene-dicarboxylic acid, di-n-octyl ester to m/z 149 daughter spectra from the reference database.....	32
2.1	Logical device names, typical values, and scan limits on the Extrel TQMS instrument.....	57
2.2	Reaction orders for the production of selected daughter ions from the phthalate anhydride ion on two different TQMS instruments.....	65
2.3	Reaction orders for the production of selected daughter ions from the m/z 71, 57, and 43 ions from n-heptane.....	75
2.4	Record for registry name CS520 in the MAPS database.....	87
3.1	Rule composition for the ethyl substructure at 3 different C values.....	114
3.2	Results from application of selected substructure rules to the training set at several match values.....	119
3.3	Partial results from cross correlation of the rules.....	122
3.4	Unreliable rule lists at three different match values.....	123
4.1	Exclusion rules for the phthalate-ester substructure generated at 3 different C values: a) C = 70%, b) C = 85%, and c) C = 100%.....	132

4.2	Inclusion rules for the phthalate-ester substructure generated at 3 different U values: a) U = 33%, b) U = 50%, and c) U = 100%.....	135
4.3	Exclusion rules for the ethyl substructure generated at a C value of 50%. Parts a and b represent the ethyl exclusion rule generated with and without use of intensity data, respectively.....	137
4.4	Inclusion rules for the benzyl substructure generated at a U value of 100%. Parts a and b represent the benzyl inclusion rule generated with and without use of intensity data, respectively.....	140
5.1	Inclusion rules for the carbonyl, chloro, and ethyl substructures generated at C = 75%.....	172
5.2	Inclusion rule for the carbonyl substructure generated at U = 75%.....	174
5.3	Inclusion rules for the chloro substructure generated at U = 75%.....	175
5.4	Inclusion rules for the bromo substructure a) generated at C = 33% and b) using feature combinations derived from this rule.....	187
5.5	Number of rule clauses, number of feature combinations identified, and overall recall for sets of feature combinations when applied to the training set using rules generated at several C values.....	189
5.6	Number of rule clauses, number of feature combinations identified, and overall recall for sets of feature combinations when applied to the training set using rules generated at several U values.....	190
6.1	Comparison of experimental to predicted isotopic ratios for several compounds.....	208
6.2	Molecular formula generator output for 1,2-benzene-dicarboxylic acid, di-cyclohexyl ester.....	210
6.3	Molecular formula generator output for eicosane...	216
6.4	Molecular formula generator output for 1,2-benzene-dicarboxylic acid, di-n-octyl ester...	219

7.1	Registry names, IUPAC names, molecular formulae, and molecular weights for 16 test compounds.....	224
7.2	Summary of MAPS inclusion results for 16 test compounds.....	227
7.3	Summary of MAPS exclusion results for 16 test compounds.....	228
7.4	Correct and incorrect inclusions for 16 test compounds.....	230
7.5	MFG results for 16 test compounds.....	233
8.1	Fragment formulae and correlation factors for clauses common to both the phenol and t-butyl inclusion rules (generated at a C value of 33%)...	241

LIST OF FIGURES

1.1	Classification of automated structure elucidation techniques into direct and indirect database methods.....	3
1.2	Schematic of DENDRAL's approach to structure elucidation using plan, generate, and test stages.....	12
1.3	Three-dimensional fragmentation map of n-heptane.....	17
1.4	Schematic of learning mode of original structure elucidation system.....	25
1.5	Schematic of identification mode of original structure elucidation system.....	27
1.6	Structures of 1,2-benzene-dicarboxylic acid, di-n-octyl-ester (I), the benzoyl substructure (II), and the phthalate substructure (III).....	28
1.7	Electron impact mass spectrum of 1,2-benzene-dicarboxylic acid, di-n-octyl-ester.....	30
1.8	Parent spectrum of m/z 149 from 1,2-benzene-dicarboxylic acid, di-n-octyl-ester.....	34
1.9	Daughter spectrum of the ¹³ C-containing protonated molecular ion from 1,2-benzene-dicarboxylic acid, di-n-octyl-ester.....	36
1.10	Schematic of ACES (Automated Chemical structure Elucidation System).....	39
2.1	Schematic of the Extrel 400 series Triple Quadrupole Mass Spectrometer.....	56
2.2	Structure of the phthalate anhydride ion.....	59
2.3	Daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate at three different collision gas pressures.....	61

2.4	Daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate at three different collision energies.....	62
2.5	Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the phthalate anhydride ion from di-n-octyl-phthalate on the Extrel TQMS instrument.....	63
2.6	Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the phthalate anhydride ion from di-n-octyl-phthalate on the LLNL TQMS instrument.....	64
2.7	Plot of relative intensity versus collision energy for several daughter ions of the phthalate anhydride ion from di-n-octyl-phthalate on the Extrel TQMS instrument.....	67
2.8	Plot of relative intensity versus collision energy for several daughter ions of the phthalate anhydride ion from di-n-octyl-phthalate on the LLNL TQMS instrument.....	68
2.9	Plots of total ion count versus collision gas pressure at several collision energies for daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate.....	69
2.10	Plots of daughter ion count versus collision gas pressure at several collision energies for daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate.....	70
2.11	Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the m/z 71 ion from n-heptane.....	72
2.12	Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the m/z 57 ion from n-heptane.....	73
2.13	Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the m/z 43 ion from n-heptane.....	74
2.14	Plot of relative intensity versus collision energy for the m/z 71 ion from n-heptane.....	75

2.15	Plot of relative intensity versus collision energy for the m/z 57 ion from n-heptane.....	76
2.16	Plot of relative intensity versus collision energy for the m/z 43 ion from n-heptane.....	77
2.17	Daughter spectra of the phthalate anhydride ion (m/z 149) from several different compounds.....	81
2.18	Daughter spectra of the phenylethyl ion (m/z 105) from several different compounds.....	82
2.19	Daughter spectra of the benzoyl ion (m/z 105) from several different compounds.....	83
3.1	Schematic of rule generation process.....	106
3.2	Abbreviated fragmentation tree for 1,2-benzene-dicarboxylic acid, diethyl ester (CS525).....	108
3.3	Schematic of rule validation process.....	117
3.4	Performance graphs for rules generated by MAPS at a C value of 75% showing the effect of exclusion of unreliable rules.....	124
4.1	Schematic of rule validation process.....	143
4.2	Recall versus match value for exclusion rules (generated at a C value of 100%) applied against the training set, with and without use of intensity data in the exclusion rules.....	146
4.3	Recall, false positives, and the reliability factor versus match value for exclusion rules generated at three different C values applied to the training set.....	147
4.4	Recall versus match value for inclusion rules (generated at a U value of 100%) applied to the training set, with and without use of intensities.....	152
4.5	Recall, false positives, and the reliability factor versus match value for inclusion rules (generated at a U value of 50%) applied to the training set, with no weighting and uniqueness weighting.....	153

4.6	Recall, false positives, and the reliability factor versus match value for inclusion rules generated at three different U values applied to the training set.....	155
4.7	Portion of the substructure hierarchy defining inheritance relationships between substructures...	159
4.8	Portion of the substructure hierarchy defining inheritance relationships between substructures...	160
5.1	Feature types obtained from MS and MS/MS data for use in rule generation via MAPS.....	169
5.2	Daughter spectrum of m/z 135 from 4-t-butyl phenol.....	171
5.3	State space for a set of 4 features comprising 15 possible combinations of individual features...	179
5.4	Schematic of procedure for identifying feature combinations for substructure identification.....	180
5.5	Portion of the state space for feature combinations indicative of the bromo substructure.....	183
5.6	Plot of computation time versus the number of nodes examined by the search algorithm.....	185
6.1	Information available from TQMS data.....	197
6.2	Schematic of molecular formula determination process.....	199
6.3	Daughter spectrum of the (M+1) ⁺ ion from eicosane.....	212
6.4	Selected daughter spectra of the m/z 149 ion from several different compounds.....	218
7.1	Registry names and structures for the set of 16 test compounds.....	225
8.1	An intelligent TQMS instrument incorporating a feedback loop from expert interpretive tools to the instrument.....	255

CHAPTER 1

STRUCTURE ELUCIDATION FROM MS AND MS/MS DATA

Introduction

The determination of molecular structures is a fundamental problem posed to chemistry, biology, and many other disciplines. Mass spectrometry (MS) has long been recognized as a powerful tool for structure elucidation. It is one of the oldest analytical techniques and was first employed in 1913 by J. J. Thomson to demonstrate that neon consists of more than one isotope (1). However, mass spectrometry did not become a general tool for structure elucidation until the 1960's. A key to this development was the growth of GC/MS as an automated mixture analysis technique. In recent years there has been tremendous growth in the so-called hyphenated techniques (2). Mass spectrometry has played no small role in this growth, and the success of GC/MS has spearheaded the "marriages" of other techniques with mass spectrometry such as LC/MS, SFC/MS, MS/MS, MS/MS/MS, and recently GC/IR/MS.

There has been considerable interest over recent years in advancing the state of automated structure elucidation. This has to some extent been a necessary reaction to the growth of hyphenated techniques, improvements in data collection speeds, and the ever-

increasing ability of instrumentation to generate large quantities of multidimensional data. The sheer volume of data produced by such techniques mandates some automated method for extracting structural information. With such multidimensional instrumentation becoming more common, one can expect that traditional structure elucidation tools and human experts will fail to extract all the valuable analytical information within a reasonable time interval. Thus, the development of new automated structure elucidation procedures has become a priority.

This chapter begins with an overview of the various approaches for automated structure elucidation using mass spectral data. These approaches can be categorized as direct or indirect database methods (3), as shown in Figure 1.1. Although it is useful to distinguish between spectral matching, pattern recognition, artificial intelligence, and spectral simulation approaches, specific structure elucidation systems may straddle these categories, as will be seen later. The advantages and disadvantages of these different approaches are discussed. Next, the utility of MS/MS data is evaluated and contrasted to MS data for structure elucidation. Lastly, the structure generator GENOA and two separate systems developed in our laboratories for automated structure elucidation from MS/MS data are described.

Direct Database Methods

Direct database methods, commonly referred to as library searching or spectral matching methods, are widely used for spectral interpretation. They require a database or library of reference spectra

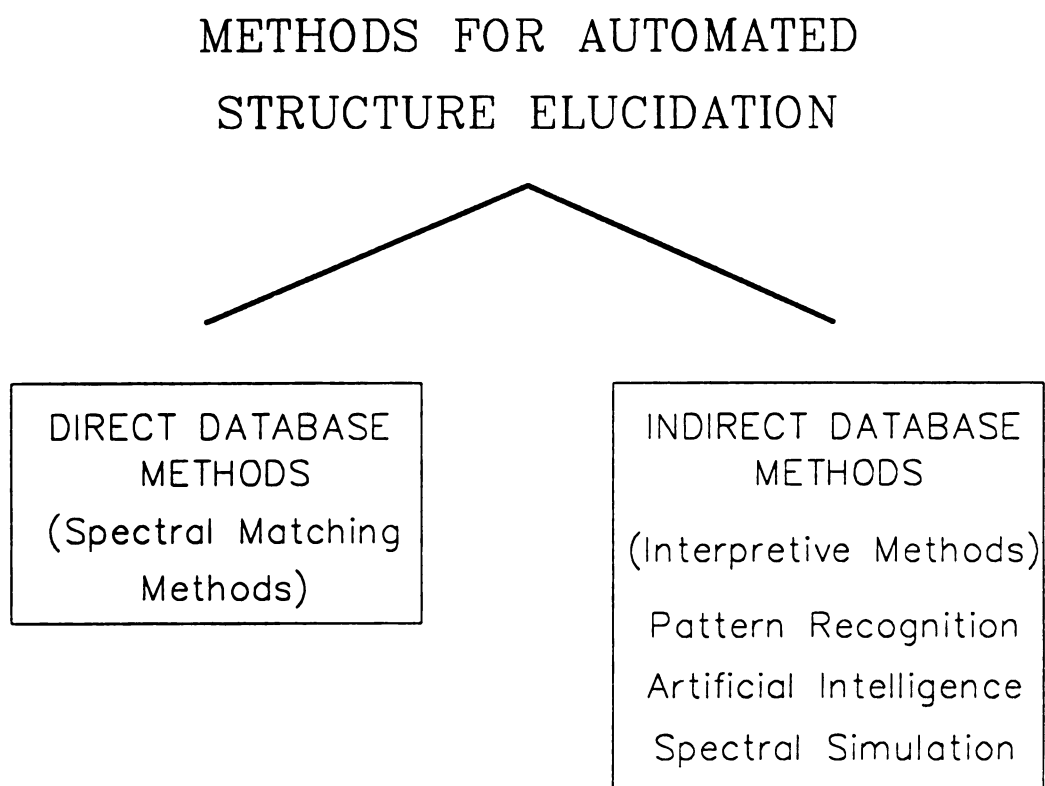


Figure 1.1 Classification of automated structure elucidation techniques into direct and indirect database methods.

along with some means for comparing sample and reference spectra. Many different approaches to spectral matching of MS data have been discussed in the literature, including the Blemann method (4), the SISCOM system developed by Henneberg and coworkers (5,6), and the LIANEL and MSSMET systems developed by Sweeley's group (7,8). Perhaps the most well-known matching system for MS data is the probability-based matching (PBM) system developed by McLafferty's group (9).

The available methods for spectral matching differ in their algorithms for data encoding and comparison. The method used for encoding determines which data from reference spectra are used in the search. The Blemann method, for example, retains only the two most intense peaks in each 14 u window for the search (4). The comparison algorithm defines the criteria used to match the sample and reference spectra and determines the degree of correspondence between them. The PBM system, for example, weights the significance of peaks in inverse proportion to their frequency of occurrence in the database. Thus, peaks which occur less frequently are more significant in distinguishing between spectra. Comparison algorithms can be classified as forward (10,11) or reverse search (12). The objective of a forward search is to retrieve the reference spectra which are most similar to the sample spectrum. In a reverse search, the matching process is driven entirely by the data in the sample spectrum which are also present in reference spectra. Thus, reverse search algorithms tend to ignore spurious peaks in sample spectra due to impurities and other spectral distortions. There

are several excellent reviews which compare and contrast different mass spectral matching schemes (6,13,14).

Considering that Chemical Abstracts Services currently recognizes *over seven million different organic compounds*, the most serious drawback to any spectral matching method is that no library can ever be complete. Other disadvantages associated with spectral matching methods are that (i) in certain cases, large libraries of mass spectra are required, (ii) the search time increases as the library size increases, and (iii) incorrect identifications and ambiguous hit lists become more common as the library size increases. These faults may be due to the nature of the matching algorithm itself, the use of an abbreviated database, contamination in either the sample or reference spectra, the use of different instrumental parameters or instruments for sample and reference spectra, or the inadequacy of even perfect spectra to distinguish between closely related structures.

Spectral matching has become a mature technique for spectral interpretation and further improvements may be limited in scope. The increasing use of fourier transform mass spectrometry and double focusing instruments has led to the generation of high-resolution mass spectra on a routine basis. Spectral matching methods using such data have shown improved performance (15). The intelligent application of prefilters, improvements in computing speeds, and advances in storage media technology will continue to improve the performance of these methods. Spectral matching has proved to be of great utility for the analysis of the large amounts of spectra generated by such techniques as GC/MS. However, a high degree of match does not necessarily prove the

identity of the unknown. The spectral differences observed for closely related compounds are often too small to be distinguished by spectral matching. In addition, few spectral matching systems attempt any interpretation of their results. It is the user who must evaluate the match list, the significance of the match factors, and the relevance of the corresponding structures. While spectral matching methods are valuable aids for limited domain problems involving known compounds, for true unknowns (compounds whose spectra may not exist in *any* library), one must often resort to indirect database or interpretive methods.

Indirect Database Methods

Indirect database or interpretive methods for structure elucidation attempt to distill the information contained in a spectral database to a more general form which can then be employed independently of the database. These methods can be classified into three categories: pattern recognition, artificial intelligence, and spectral simulation approaches (3). The relative merits and disadvantages of each of these approaches are discussed below along with appropriate examples of systems which exploit them.

Pattern Recognition Approaches. Humans are naturally adept at pattern recognition. It has been said that "To compete with the human's ability to recognize patterns in only two or three dimensions is a fairly difficult task" (16). However, when dealing with very large quantities of data or several dimensions of data, the human's ability to recognize

patterns is poor. In the field of chemistry, automated data interpretation techniques are necessitated by the large amount of multidimensional data routinely generated by modern instrumentation. Thus, pattern recognition techniques, and in particular, supervised learning techniques have found a wide variety of applications in chemistry (17-19). "Supervised learning refers to a suite of techniques in which *a priori* knowledge (or assumptions) about the category membership of a set of samples is used to develop a classification rule. The purpose of the rule is usually to predict the category membership for new samples" (16).

Spectral matching systems can be classified as a rudimentary form of pattern recognition. Zupan and coworkers have used hierarchically organized databases or "trees" of mass spectra to achieve clustering of spectra whose compounds contain similar structural features (20). This method offers several advantages over conventional spectral matching approaches. Average retrieval time is logarithmically proportional to the number of objects in the tree. When an unknown passes through the tree, substructural features can be predicted on the basis of the properties of the nodes encountered.

Pattern recognition methods have been widely used for structure elucidation. Several papers have concentrated on using pattern recognition techniques for molecular structure descriptions (17,21,22). Pattern recognition approaches have been applied to mass spectral data to identify specific compound classes (23-25). In an insightful paper, MacLagan and Mitchell compared the capacity of four different pattern recognition techniques for predicting 21 different structural features from mass spectra of nucleosides (26). Recently, principal component

analysis has been applied to MS/MS spectra to differentiate between different alkylbenzene isomers (27).

The self-training interpretive and retrieval system (STIRS) is a pattern recognition technique which deduces substructural information in unknowns through the application of 26 different classes of mass spectral data (9). These data classes correspond to combinations of masses or mass differences which are known to have structural significance, such as neutral losses, characteristic ions, and ion series. STIRS extracts these data classes from an unknown mass spectrum and matches them against the corresponding data classes from library spectra. The structures of the best matching library spectra in each class are then analyzed for common structural moieties. If a large portion of these contain a specific substructure, then that substructure is likely to be present in the unknown. In addition to predicting substructures, STIRS also can predict molecular weight and chlorine and bromine composition from isotopic ratio data. In tests of unknown spectra, STIRS gave consistently better results than a k-nearest neighbor approach (28), identified an average of 49% of the substructures present in these unknowns at a reliability of 98% (29), and correctly identified three substructures in each unknown along with perhaps one incorrect prediction (30). The STIRS system is not restricted to analysis for a standard set of substructures. Given the spectrum of an unknown, STIRS will retrieve the most similar spectra, which can then be analyzed for common substructural features to identify specific substructures. This technique directly utilizes information from all available library spectra without resorting to predefined spectrum/substructure

correlations, hence the "self-training" appellation. Unfortunately, STIRS has not found wide applicability in the real world and lacks a structure generator to assemble complete structures.

Analytical instrumentation can now provide several dimensions of information. GC/IR/MS, for example, provides five dimensions of data: retention times, infrared frequencies and absorption values, m/z values, and peak intensities. Data interpretation is the bottleneck for many techniques. Pattern recognition methods will be increasingly employed to derive the most useful information from large quantities of data. However, most pattern recognition approaches are capable of only binary decisions (e.g., the sample is or is not in class A). Obviously, a structure elucidation scheme utilizing pattern recognition would require many such decisions. Most pattern recognition classification systems do not assign any measure of reliability to their predictions. In addition, pattern recognition systems, unlike artificial intelligence based approaches, cannot justify their conclusions.

Artificial Intelligence Approaches. Artificial intelligence has been defined as "the scientific discipline which attempts to endow computers and computer-controlled machinery with the ability for actions which, if carried out by a human being, would be thought to require intelligence" (31). Artificial intelligence research currently comprises several different areas, including cognitive science, natural language processing, automated learning, robotics, and expert systems (32).

Expert systems, also referred to as knowledge-based systems, are finding a broad range of applications in chemistry (33-36). To become an

expert in a particular domain requires years of experience and the capability to deal with many problems over a narrow range. Expert systems draw upon the collective knowledge of humans to solve problems which would normally require human intelligence. Expert systems usually consist of three parts: a knowledge base, an inference engine, and a user interface. The knowledge base is a collection of heuristics, or "rules-of-thumb", which are usually in the form of rules. An inference engine is used to apply the rules to reach intelligent conclusions.

There are substantial differences between conventional programs and knowledge-based systems. Conventional programs reach decisions using data and prescribed algorithms in a manner which is opaque. Modifying these programs is often difficult and tedious because the knowledge for solving the problem is scattered throughout the code. Expert systems make judgements based on rules and knowledge. They can justify their conclusions and thus their reasoning is transparent. The performance of an expert system can be extended and otherwise modified by changing the rules or adding new rules. Expert systems can work with uncertain and incomplete data and can contemplate multiple competing hypotheses. Many can interact with humans using natural language. Although LISP and Prolog are the preferred languages for expert system development, expert systems can be written in more conventional languages, such as FORTRAN and C.

One of the most famous applications of artificial intelligence is the DENDRAL project (37-39), which began at Stanford University in 1965. This project represents the culmination of many man years of work and

the collective knowledge of many mass spectroscopists. DENDRAL is a classical approach to a solution of a problem with a large state space. It employs three stages: plan, generate, and test. The plan stage (Heuristic DENDRAL) derives constraints on the unknown structure. Empirically derived fragmentation rules are used to determine the molecular fragments which are present or absent in the unknown. These fragmentation rules are automatically inferred from the mass spectra of known compounds by Meta-DENDRAL (38). The generate stage (GENOA) generates all possible structures consistent with the constraints (40). The test stage ranks the resulting list of structures by simulating their mass spectra and comparing them to the unknown spectrum. This process is depicted schematically in Figure 1.2.

DENDRAL has been applied to several problems and its performance has been said to equal or exceed the performance of a human expert in structure elucidation. DENDRAL's power is not derived from "knowing" more than any human expert, but from effective application of constraints and a systematic search through the "state space" of all possible structures. However, in many cases, mass spectral data alone were insufficient to determine the complete structure of an unknown. In these cases, DENDRAL required additional substructural constraints derived from i.r. and n.m.r. data (41).

An increasing number of applications of knowledge-based systems for structure elucidation are being reported in the literature. Most of these systems utilize spectral feature/substructure relationships in the form of rules. The simplest and often the most effective systems exploit spectral correlation charts (42). The CHEMICS system developed by Sasaki and

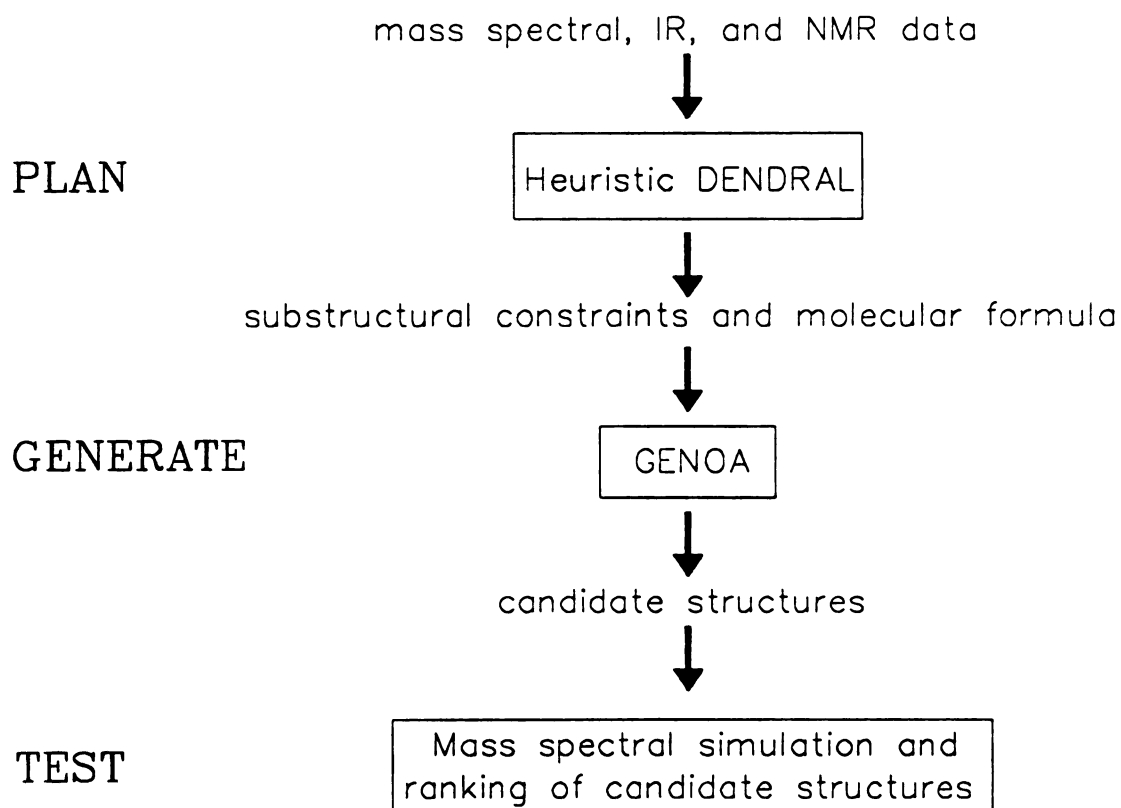


Figure 1.2 Schematic of DENDRAL's approach to structure elucidation using plan, generate, and test stages.

coworkers is one of the more well-known rule-based systems for structure elucidation, and utilizes data from i.r., mass, proton n.m.r., and ^{13}C n.m.r. spectroscopy correlation charts (43-46). Combining data from different spectroscopic techniques can often yield complementary information and confirmatory evidence. It is also possible to tailor such expert systems for specific structure elucidation applications.

Expert systems can solve problems which are often impossible to solve via conventional programming. They can perform as well as or better than human experts. Expert systems convert the private or sometimes inactive knowledge of experts into an active, inspectable form. They are useful for training young experts since they can reveal the reasoning behind their conclusions. However, expert systems have been somewhat "overhyped" in recent years and it is important to understand their limitations. Construction of a knowledge base is laborious and often requires a knowledge engineer. Expert systems exhibit unstable behavior near the limit of their knowledge. The explanations for the conclusions drawn are stereotyped. More importantly, the expertise of such systems are confined to a narrow range. While expert systems have yet to find broad applicability in the real world, they hold much promise for the future. In summary, the current status of computer-assisted structure elucidation systems is best described by Hippe: "the actual performance of program systems for structure elucidation by artificial intelligence at the present time is generally at the level of performance of a post-doctoral spectroscopist. Their performance is good, not because they use know more than an experienced spectroscopist, but because they use most of the rules applied by a spectroscopist to solve structure

problems, because they apply the same set of rules to every problem, and because they apply systematically the whole set of rules each time, without mistakes or loss of memory" (47).

Spectral Simulation Approaches. Spectral simulation is often used to determine which of a set of candidate structures is most likely the unknown, and thus represents only a part of any structure elucidation technique. DENDRAL uses mass spectral simulation to rank a set of candidate structures. In this approach, the mass spectra of candidate structures are simulated using empirically derived fragmentation rules (38). The simulated spectra are then matched against the unknown spectrum and ranked on the basis of their similarity to indicate the most likely candidate structure. Jurs used pattern recognition techniques to correlate individual fragment masses to structural, geometrical, and topological features; these correlations were then used to generate an artificial spectrum for a given structure (22). Quasi-equilibrium theory, which can provide data on the relative abundance of specific ions with respect to energy, can be used to calculate the mass spectra of simple molecules. Accurate spectral simulation requires either a large database of fragmentation rules (in the case of DENDRAL) or enormous amounts of mass spectral data (in the case of pattern recognition based techniques). Spectral simulation is an excellent means of "closing the loop" for any interpretive method to verify the candidate structure(s). However, a complete and accurate simulation of mass spectra for all molecules under various experimental operating conditions is currently unobtainable.

Utility of MS/MS Data for Structure Elucidation

There are many reasons for the prominence of mass spectrometry among techniques for structure elucidation. It is an extremely sensitive technique; mass spectra can be obtained from nanomole quantities or less. Common fragment ions and neutral losses (which are postulated from the mass difference between two peaks) from mass spectra have been recognized as fairly specific indicators for certain structural features. These have been tabulated and are widely used in spectral interpretation (48,49). Over the last three decades, a tremendous amount of research has been invested into developing automated techniques for structure elucidation from MS data.

A major difficulty in the interpretation of conventional mass spectra is that the products of all the fragmentation processes are overlapped in a mass spectrum. Electron impact (EI) ionization imparts ions with excess energy. These ions then undergo fragmentation within the ion source, and the subsequent ion-molecule reactions (which depend on the source pressure) and decompositions can give a wide variety of products. Rearrangements further complicate interpretation. Mass spectra indicate only the presence of ions and give no certain information on which ions are precursors of other specific ions or which ions are formed from fragmentation of other specific ions.

MS/MS, as its name suggests, provides another dimension of mass spectrometry to conventional MS. In a typical MS/MS instrument, the sample is ionized and ions selected by the first mass analyzer enter a

collision chamber filled with an inert gas where they may undergo further fragmentation. This process is referred to as collision-induced dissociation (CID) or collisionally-activated dissociation (CAD). The products of such fragmentation processes may then be analyzed using a second stage of mass spectrometry.

Figure 1.3 shows a partial MS/MS spectrum or three-dimensional fragmentation map for n-heptane. Individual MS/MS spectra may be either daughter, parent, or neutral loss spectra. Examples of daughter and parent spectra are shown in this figure. The daughter spectrum represents the fragmentation products formed from collisional activation and subsequent dissociation of the parent ion (m/z 85). The parent spectrum represents the precursor ions which fragment to form a specific daughter ion (m/z 71). A neutral loss spectrum, which is not shown in this figure, represents the parent ions which undergo a defined neutral loss. A *complete* MS/MS spectrum or fragmentation map would contain a daughter spectrum for *each* ion appearing in the conventional mass spectrum of a compound.

MS/MS has several advantages over conventional MS for structure elucidation, the most obvious of which is the second dimension of information. Neutral losses and parent-daughter relationships can be determined directly, rather than inferred as in conventional MS. From Figure 1.3 it is obvious that a tremendous amount of information can be obtained from even small molecules using MS/MS. Due to the increased data dimensionality, MS/MS spectral features are generally more specific and selective than MS features. The identity of specific fragment ions can often be deduced from their characteristic daughter spectra. Thus,

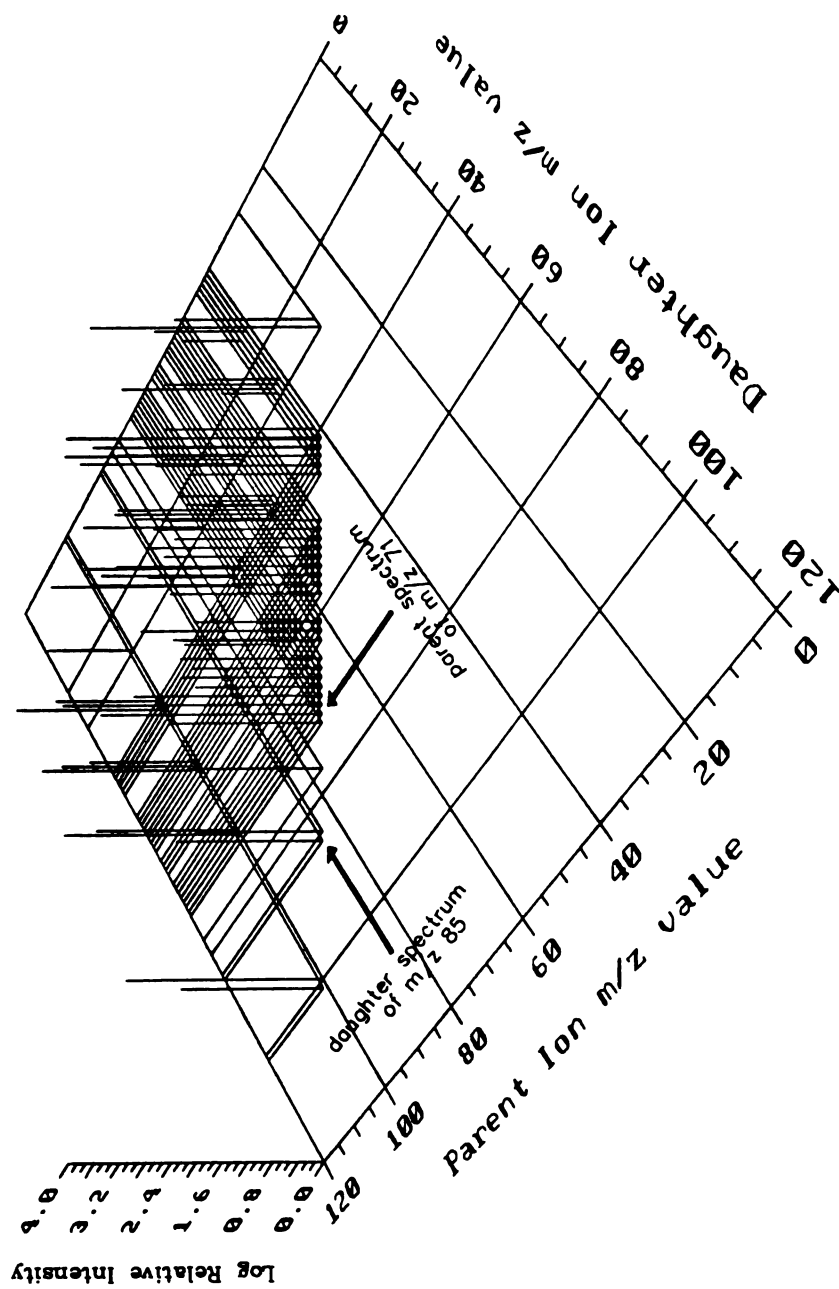


Figure 1.3 Three-dimensional fragmentation map of n-heptane.

MS/MS data can often be used to differentiate between structural isomers which are not distinguishable using MS data alone. Also, MS/MS spectra have a lower background ion current than MS spectra. The noise reduction resulting from the second dimension of mass spectrometry in MS/MS makes the absence of specific peaks significant. Instrumental parameters such as collision energy and collision gas pressure provide additional dimensions of information in an MS/MS experiment.

The use of MS/MS as a principal technique for structure elucidation was proposed in 1978 (50). In this insightful paper, Beynon and coauthors demonstrated the following advantages of MS/MS for structure elucidation using data from a mass-analyzed ion kinetic energy spectroscopy (MIKES) instrument:

- 1) molecular formulae can be deduced without resorting to high-resolution mass spectrometry,
- 2) substructures can be identified from their characteristic daughter spectra,
- 3) identification of multiple occurrences of substructures in a given structure is possible, and
- 4) molecular structures which differ in a minor structural feature can be differentiated.

These advantages have some very important implications for structure elucidation. The utility of daughter spectra for identifying substructural features has been demonstrated on several occasions

(50,51). A database of daughter spectra can be used to identify the presence of substructures in an unknown *regardless of whether its MS/MS spectra have ever been taken before*. Whereas the number of possible organic compounds is far greater than the number of mass spectra which can be included in a library, a relatively small library of daughter spectra can aid in the structure elucidation of a virtually unlimited number of compounds. Furthermore, substructure information obtained from daughter spectra constrains the number of possible molecular structures consistent with a known molecular formula and thus greatly aids the structure elucidation process. If sufficient substructural information is obtained, the complete structure of an unknown can often be "pieced together". Thus, *structure elucidation from MS/MS data can be achieved through substructure identification*.

This paper proposed some exciting, novel ideas for structure elucidation (50). For various reasons, very little follow-up work has been done since it was published. Since then, MS/MS has developed significantly; MIKES has been supplemented and in many ways eclipsed by other tandem mass spectrometric techniques such as tandem quadrupole mass spectrometry (TQMS). TQMS as opposed to MIKES can provide unit or better mass resolution in both mass analyzers, has better transmission characteristics, and represents a significant improvement in MS/MS instrumentation for structure elucidation (51).

The first TQMS instrument was developed in our labs in the late 1970's by Rick Yost and Dr. Enke (51-53). Later, this instrument was completely computerized using a Newcome-Enke 8085-based

microcomputer (54,55). Instrument control software developed in the FORTH language by Carl Myerholtz allowed the operator to flexibly and conveniently control the TQMS instrument's five scan modes and 30 parameters (56-58). Recently, a multimicroprocessor control system and a control system based on the Novix NC4000 Forth processor (59,60) were implemented to improve the efficiency of data collection for the TQMS instrument (61). A substantial amount of software has been developed in our labs for managing mass spectral data, including a multidimensional database for MS and MS/MS spectra (62,63), software for MS and MS/MS spectral matching (64,65), and software for extracting several planes of data from such a database (63). Thus, it can be seen that Dr. Enke's group has played a large role in not only developing the TQMS instrument itself, but in developing TQMS control systems and software for management of mass spectral data. The next logical step to this research was the development of the full capabilities of MS/MS for structure elucidation.

Currently, *no integrated, interpretive systems for structure elucidation using MS/MS data exist outside this laboratory.* Our approach to structure elucidation from MS/MS data has continually evolved over the course of several years and sundry graduate students. Later on in this chapter, this evolution is traced to show two distinct interpretive systems for structure elucidation which have been developed. Both are based on the *identification of substructures* in unknowns. The first approach achieves this using a direct database or spectral matching method (66). The second approach achieves this through the use of empirically derived substructure rules (67). Each of

these systems require a structure generator, specifically GENOA, which is described next.

The Structure Generator (GENOA)

A crucial part of any interpretive structure elucidation system is a structure generator to assemble the various pieces of information about the unknown into a structure or set of structures. Arguably one of the best is GENOA, a constrained structure generator developed during the course of the Stanford DENDRAL project (40). This program generates molecular structures based on substructural constraints provided by the user. These constraints usually take the form of a substructure and a range of occurrence (i.e., "octyl at least one"). GENOA accepts overlapping substructures as constraints, and thus each substructural constraint does not have to encompass a unique portion of a molecule. In addition, the presence of alternate substructures (i.e., either substructure A or substructure B) can also be specified, and thus even ambiguous substructure information can be used. Negative information (substructures known to be absent) can also be utilized to constrain the structure generation. These substructural constraints may be derived from any source. In elucidating the structure of several compounds, GENOA users have derived constraints from MS, i.r., and n.m.r. data (41). For the purposes of this research, the substructural constraints are derived from MS and MS/MS data only.

An additional and essential piece of information required by GENOA is the molecular formula of the unknown. Given this along with

substructural constraints, GENOA exhaustively generates *all* plausible candidate structures. This capability eliminates the possibility that a human chemist might overlook any possible structure. Most importantly, the structure of the unknown will always be contained within the set of structures produced by GENOA using *correctly* identified substructures as constraints.

The commercial GENOA software package also includes a routine called STRCHK which performs substructure searching. Given the structure of a compound and a library of predefined substructures, this procedure provides a list of substructures contained in the compound. These data are used by our structure elucidation system for developing spectral feature/substructure correlations and its use is described in Chapter 3.

The original version of GENOA was written in the BCPL programming language. A portion of this code which had been translated to the C language has been obtained by this research group. This code is undergoing several modifications in-house to better suit the purposes of our structure elucidation system.

Original System for Structure Elucidation using MS/MS Data

The original system developed by this research group for structure elucidation from MS/MS data was completed in 1985. This system is based on correspondence of daughter spectra and substructures. Later, we realized that this concept was unnecessarily limited. Over the last few years, Kevin Hart, Adrian Wade, and I have developed more

sophisticated and powerful software tools for structure elucidation which have addressed many of the limitations of the original system. Thus, this original system has since been abandoned. A discussion of this system is worthwhile for understanding the evolution of our approach to structure elucidation from MS/MS data.

The original system utilizes several software tools and databases which represent the work of various individuals (66). A multidimensional database, designed and developed by Hugh Gregg, is used to hold immediate experimental data from the Extrel TQMS instrument (62,63). A reference spectra database, designed and implemented by Phil Hoffman, is used to archive MS and MS/MS spectra (65,68). A structure/substructure database, designed and developed by Carl Beckner and Kevin Cross, is used for storage of both molecular structures and substructures (65). These last two databases include logical links to provide for correlations between daughter spectra and substructures. Spectral matching software (the MATCH program), developed by Kevin Cross, can be used to match either MS or MS/MS spectra (64,65). A molecular formula generator (MFG) developed by this author is used to obtain molecular formulae from unit resolution MS and MS/MS data (69). The STRCHK routine of GENOA is used for substructure searching and GENOA itself is used for structure generation (40). My contributions to this system also include studies on the effects of various instrumental parameters on daughter spectra, development of a set of criteria for collecting a database of such spectra, collection of daughter spectra for the reference database, and extensive

studies for identification and confirmation of daughter spectra/substructure relationships.

This structure elucidation system requires two separate modes to exploit MS/MS data from known and unknown compounds. A TQMS instrument is used to collect MS and MS/MS data for known and unknown compounds. In the learning mode, experimental daughter spectra and structures from known compounds are used to develop daughter spectrum/substructure correlations. In the identification mode, these correlations are used along with other databases and software tools to aid in the structure elucidation of unknown compounds. Both of these modes of operation will now be described.

The learning mode of this structure elucidation system is shown in Figure 1.4. MS/MS data from known compounds are collected, comprising daughter spectra for every major ion appearing in their conventional mass spectra. These daughter spectra are stored in the experimenter's database and may be archived in the reference database. Individual daughter spectra can be matched against spectra from other reference compounds by the spectral matching program. High match factors indicate that the test spectrum and the matching spectra share an identical or closely related ion structure. Next, the molecular structures of compounds producing similar daughter spectra are compared using the substructure searching capabilities of GENOA (the STRCHK routine) to identify the substructural features they have in common. The common substructures are then identified as likely precursors of the common ion. Additional compounds with common substructures are studied until clear daughter spectrum/substructure

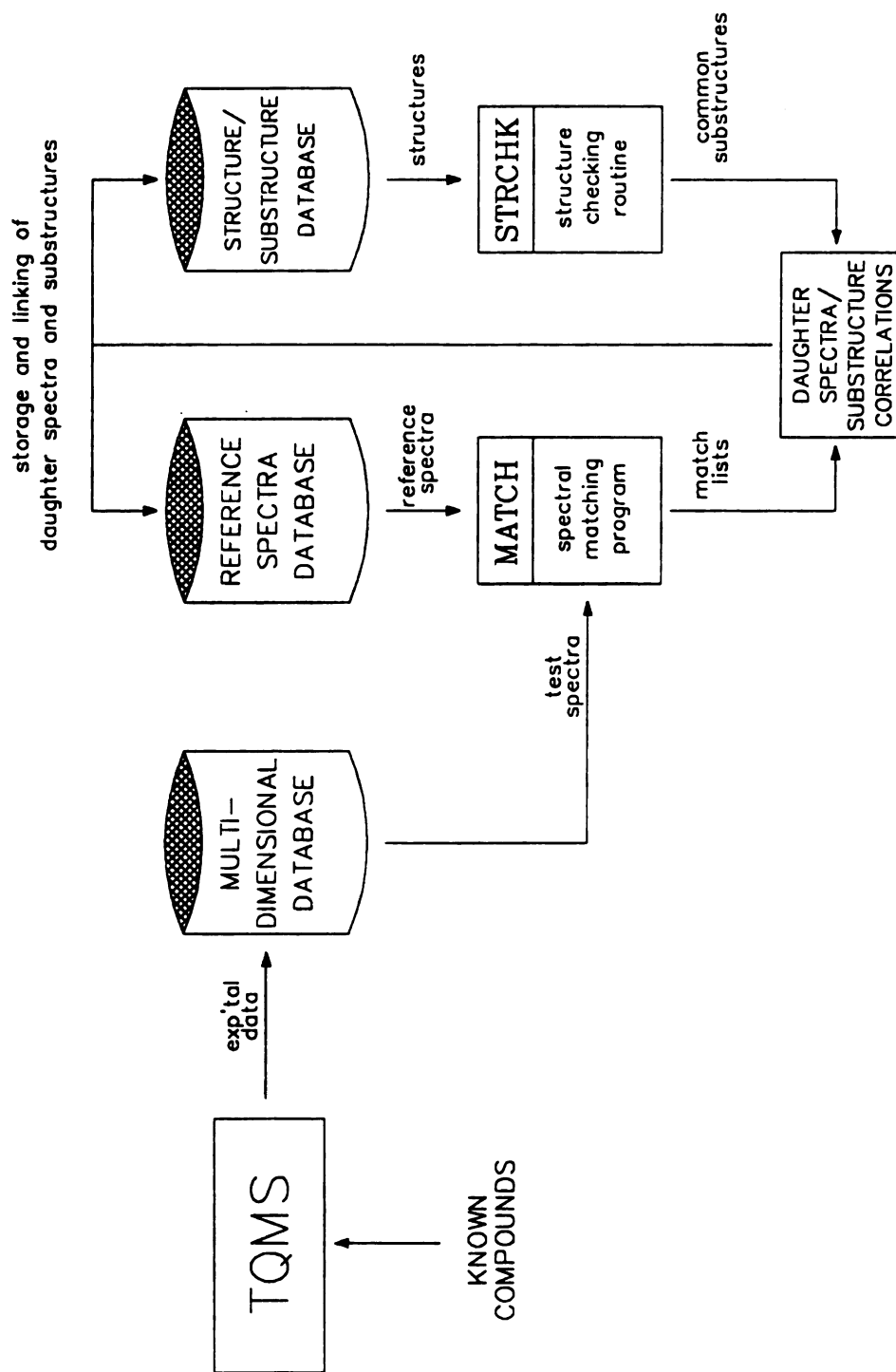


Figure 1.4 Schematic of learning mode of original structure elucidation system.

correlations are produced. Once these correlations are made, the daughter spectra are archived in the reference spectra database, the associated substructures are stored in the structure/substructure database, and the appropriate daughter spectra and substructures are logically linked.

The identification mode of this structure elucidation system is shown in Figure 1.5. Daughter spectra for every major ion appearing in the conventional mass spectrum of an unknown are collected and stored in the multidimensional database. The spectral matching program then compares these daughter spectra against those in the reference database. High match factors suggest the presence of the corresponding substructure in the unknown. The best matches from the match lists produced by this program are then used to extract the corresponding substructures from the structure/substructure database. By acquiring a daughter spectrum for every major ion in the conventional mass spectrum of the unknown, many of the substructures present in the unknown may be identified. This process may produce redundant and overlapping substructures, which may serve as confirmatory evidence. The MFG program is then invoked to generate molecular formulae from the molecular weight and elemental composition constraints. Once given a molecular formula and substructural constraints, GENOA then exhaustively generates all candidate structures.

An example of the use of the identification mode of this system is demonstrated here using 1,2-benzene-dicarboxylic acid, di-n-octyl ester (structure I in Figure 1.6) as a test compound. Daughter spectra were acquired for every ion whose abundance was greater than 1% of the base

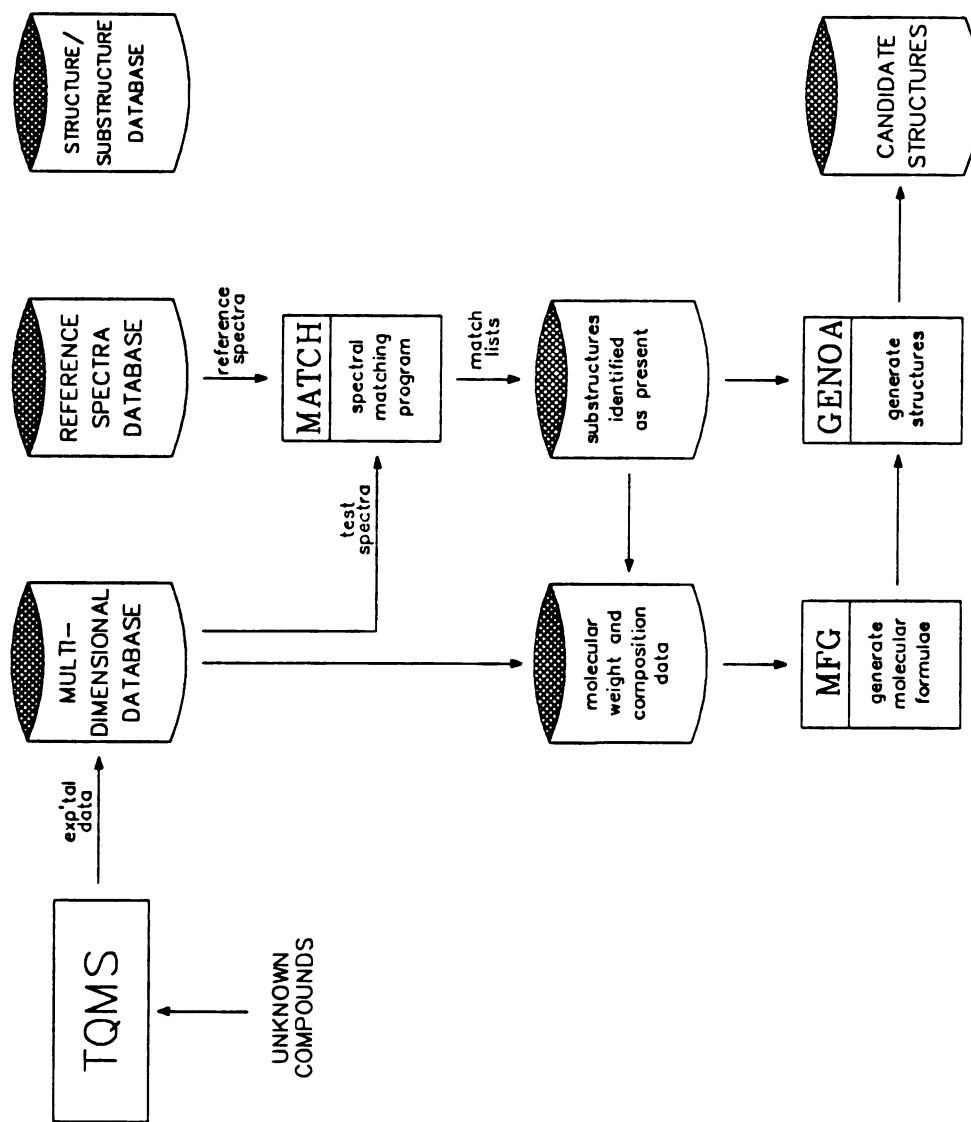


Figure 1.5 Schematic of identification mode of original structure elucidation system.

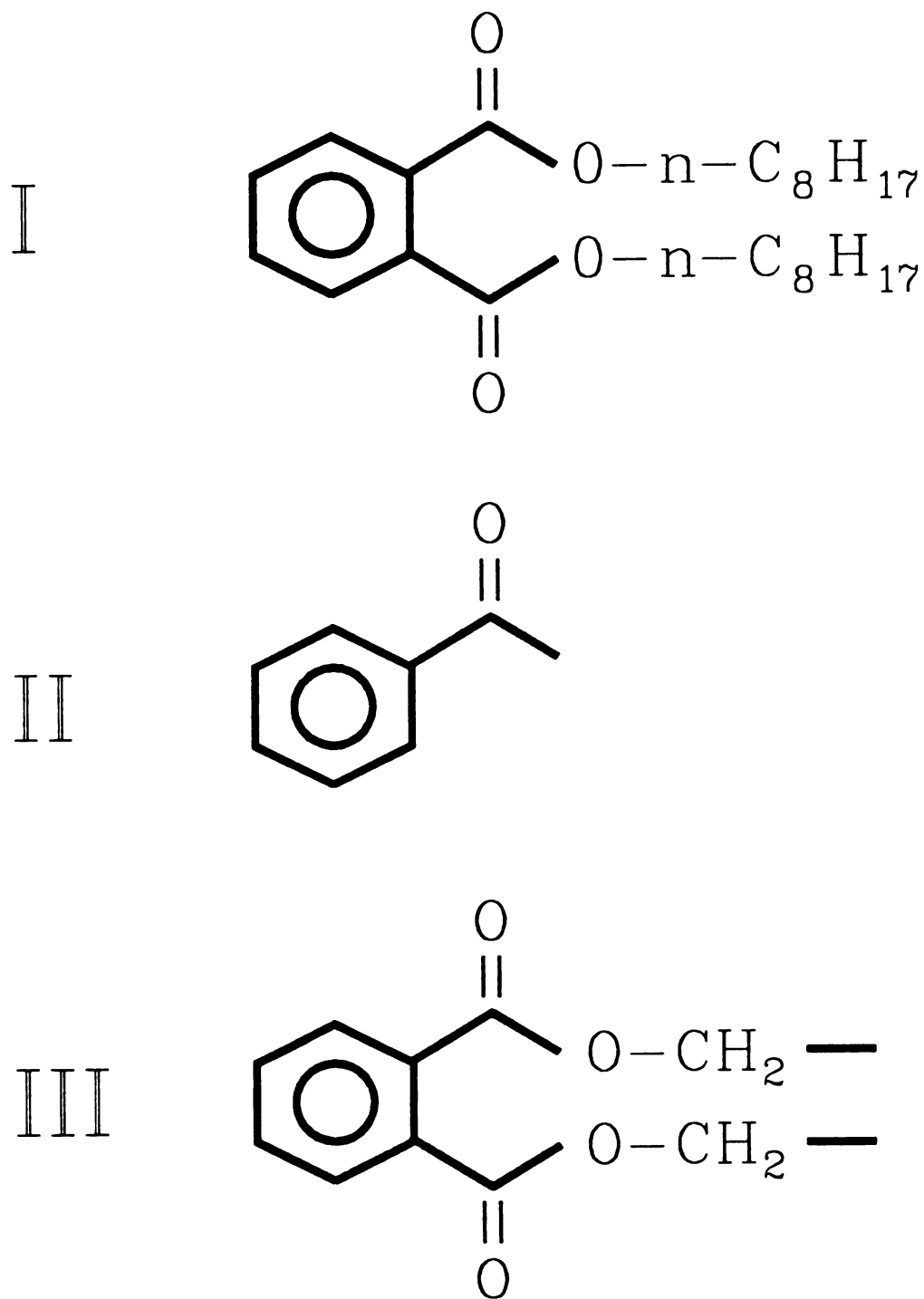


Figure 1.6 Structures of 1,2-benzene-dicarboxylic acid, di-n-octyl-ester (I), the benzoyl substructure (II), and the phthalate substructure (III).

peak in the conventional mass spectrum of this compound (Figure 1.7). These daughter spectra were then matched against daughter spectra of the same parent mass from the reference database. The results of some of these matches are described below. The various match factors calculated by the MATCH program are described in Table 1.1 (64-66). The overall match factor (PT) combines information from both forward and reverse search techniques. The pattern correspondence match factor (PC) takes into account the differences in intensities between the sample and reference spectra for peaks in common to both spectra. NC, NS, and NR give an indication on the number of matching peaks between spectra. IS and IR indicate the magnitude of intensity that is unmatched in the forward and reverse directions.

The match of the daughter spectrum of m/z 105 from the test compound to m/z 105 daughter spectra in the reference database is presented in Table 1.2. The top four matching spectra correspond to the benzoyl substructure (structure II in Figure 1.6). Note that the top four matching spectra are very similar; all contain the same peaks, only the intensity patterns are different. There also is a large difference in match factors for daughter spectra representing the correct substructure and the next best match.

The results of the match of the m/z 149 daughter spectrum of the test compound against m/z 149 daughter spectra from other compounds in the reference library are given in Table 1.3. The top four spectra all correspond to the phthalate substructure (structure III in Figure 1.6). It is important that the daughter spectra are correctly grouped by MATCH program and that there is a substantial difference between the overall

Table 1.1 Match factor definitions.

PT	An overall match factor that indicates how well the intensities of all the peaks between the two spectra match.
	$PT = 100 * (\sum Y_S + \sum Y_R - 2 * (Y_R - Y_S)) / (\sum Y_S + S Y_R)$ <p>where $Y_i = \log_2$ (Intensity/Total Ion Count) and Y_S and Y_R correspond to the adjusted abundances at each mass in the sample and reference spectra respectively</p>
PC	A pattern correspondence factor that indicates how well the intensity of the peaks in common match.
	$PC = 100 * (\sum Y_S - (Y_R - Y_S)) / (\sum Y_S)$
NC	The number of peaks common to both the candidate and unknown spectrum.
NR	The number of peaks remaining unmatched in the unknown spectrum.
NR	The number of peaks remaining unmatched in the reference spectrum.
IS	The percent total ion current of the unknown spectrum that was unmatched in the comparison due to NS.
IR	The percent total ion current of the reference spectrum that was unmatched in the comparison due to NR.

Table 1.2 Match of m/z 105 daughter spectra of 1,2-benzene-dicarboxylic acid, di-n-octyl ester to m/z 105 daughter spectra from the reference database.

PT	PC	NC	NS	NR	IS	IR	COMPOUND
100	100	2	0	0	0	0	DI-N-OCTYLPHTHALATE
99	99	2	0	0	0	0	DI-N-PENTYLPHTHALATE
98	98	2	0	0	0	0	DI-N-BUTYLPHTHALATE
98	98	2	0	0	0	0	DI-ETHYLPHTHALATE
66	93	2	0	2	0	31	4-N-BUTYL-1,2-BENZENEDIOL
60	85	2	2	0	2	20	2-N-BUTYL-4-METHYLPHENOL
38	50	1	1	3	42	29	P-T-BUTYLBENZYL ALCOHOL
36	50	1	1	3	42	52	2-T-BUTYL-6-METHYLPHENOL

Table 1.3 Match of m/z 149 daughter spectra of 1,2-benzene-dicarboxylic acid, di-n-octyl ester to m/z 149 daughter spectra from the reference database.

PT	PC	NC	NS	NR	IS	IR	COMPOUND
100	100	4	0	0	0	0	DI-N-OCTYLPHTHALATE
96	96	4	0	0	0	0	DI-N-BUTYLPHTHALATE
87	86	4	0	0	0	0	DI-N-PENTYLPHTHALATE
87	86	4	0	0	0	0	DI-ETHYLPHTHALATE
54	57	3	1	7	3	2	2-N-BUTYL-4-METHYLPHENOL
44	56	3	1	10	9	15	P-T-BUTYLBENZYL ALCOHOL
42	35	1	3	1	19	29	P-T-AMYL-PHENOL
35	61	3	1	10	3	26	2-T-BUTYL-6-METHYLPHENOL

match factors for compounds containing the correct substructure and those corresponding to unrelated substructures. Once again, this is demonstrated in the match results where only the relative intensities differ between the top matching spectra.

Due to the limited number of the daughter spectra in the reference database, only two substructures were identified: the benzoyl and phthalate substructures. At the time this analysis was done, the reference database did not contain any daughter spectrum/substructure correlations for alkyl groups. Thus, we were forced to resort to manual methods of spectral interpretation to elucidate the remaining portion of the structure. This involved analysis of the neutral losses leading to the formation of the m/z 149 ion from the test compound. The parent spectrum of this ion, shown in Figure 1.8, has four major nonisotopic peaks at m/z 167, 261, 279, and 391. These ions correspond to neutral losses of 18 (167-149), 112 (261-149), 130 (279-149), and 242 (391-149). The neutral losses may be tentatively identified as H_2 (neutral loss of 18), C_8H_{16} (neutral loss of 112), $C_8H_{17}OH$ (neutral loss of 130), and $C_8H_{17}OC_8H_{17}$ (neutral loss of 242), and suggest the presence of an alkyl chain in the molecule. The low mass ion series in the conventional mass spectrum of the test compound (Figure 1.7) also suggests the presence of an alkyl chain which is most likely unbranched.

The molecular formula is required by GENOA for structure generation. Information from isotopic daughter spectra is very useful for determining molecular formulae from unit resolution MS/MS data. Chemical ionization (CI) mass spectra of the test compound identified its molecular weight to be 390 u. The daughter spectrum of the ^{13}C -

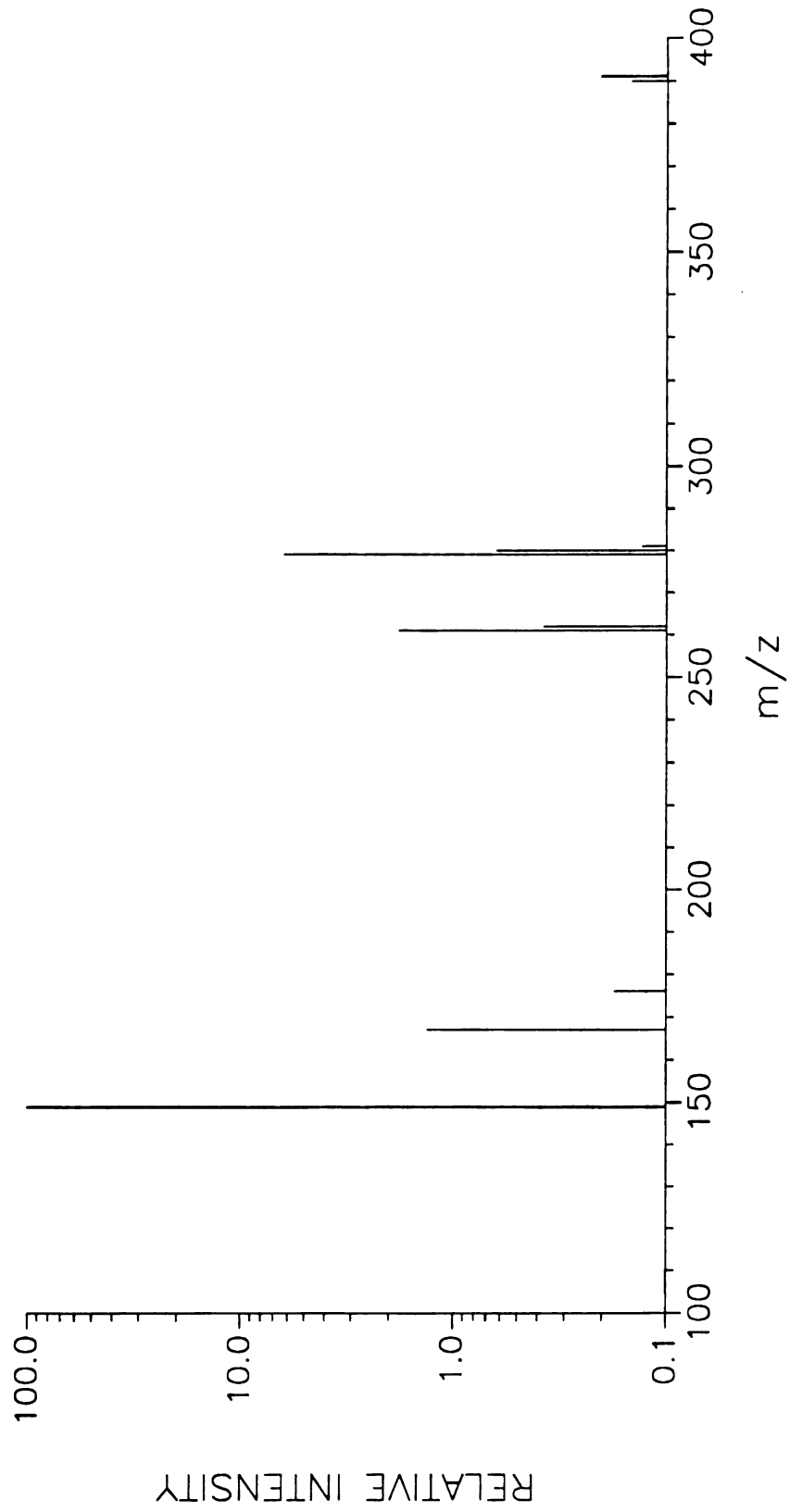


Figure 1.8 Parent spectrum of m/z 149 from 1,2-benzenedicarboxylic acid, di-n-octyl-ester.

containing protonated molecular ion, shown in Figure 1.9, shows peak pairs at adjacent masses which represent the loss or retention of the ^{13}C atom in the fragment ions formed. The relative intensities of these peak pairs depend on the ratio of carbon atoms lost to carbon atoms retained in forming that ion. The ratio of intensities of m/z 149 to 150 in Figure 1.9 was two to one. This indicates that the ^{13}C -containing molecular ion is twice as likely to lose a ^{13}C atom than retain it when fragmenting to form the m/z 149 ion. The m/z 149 ion was identified as the phthalate anhydride ion which contains 8 carbon atoms. Thus, it follows that the compound must contain 24 carbons. When this information is supplied to the MFG program along with the substructural constraints identified above, only one formula results: $\text{C}_{24}\text{H}_{38}\text{O}_4$.

Given the phthalate and benzoyl substructures identified by the matching software, the two *n*-octyl substructures whose presence was inferred manually, and the molecular formula identified by the MFG program, GENOA was then used to generate all possible structures. Only one structure results: the correct structure of 1,2-benzene-dicarboxylic acid, di-*n*-octyl ester.

It is obvious that there are several disadvantages associated with this original structure elucidation system. The process of identifying the daughter spectrum/substructure correlations was found to be time consuming and operator intensive. It required a mass spectroscopist to invoke several software tools including the matching software to identify similar daughter spectra from different compounds, the STRCHK routine of GENOA to identify common substructures in the structures of these compounds, and additional software to logically link the daughter

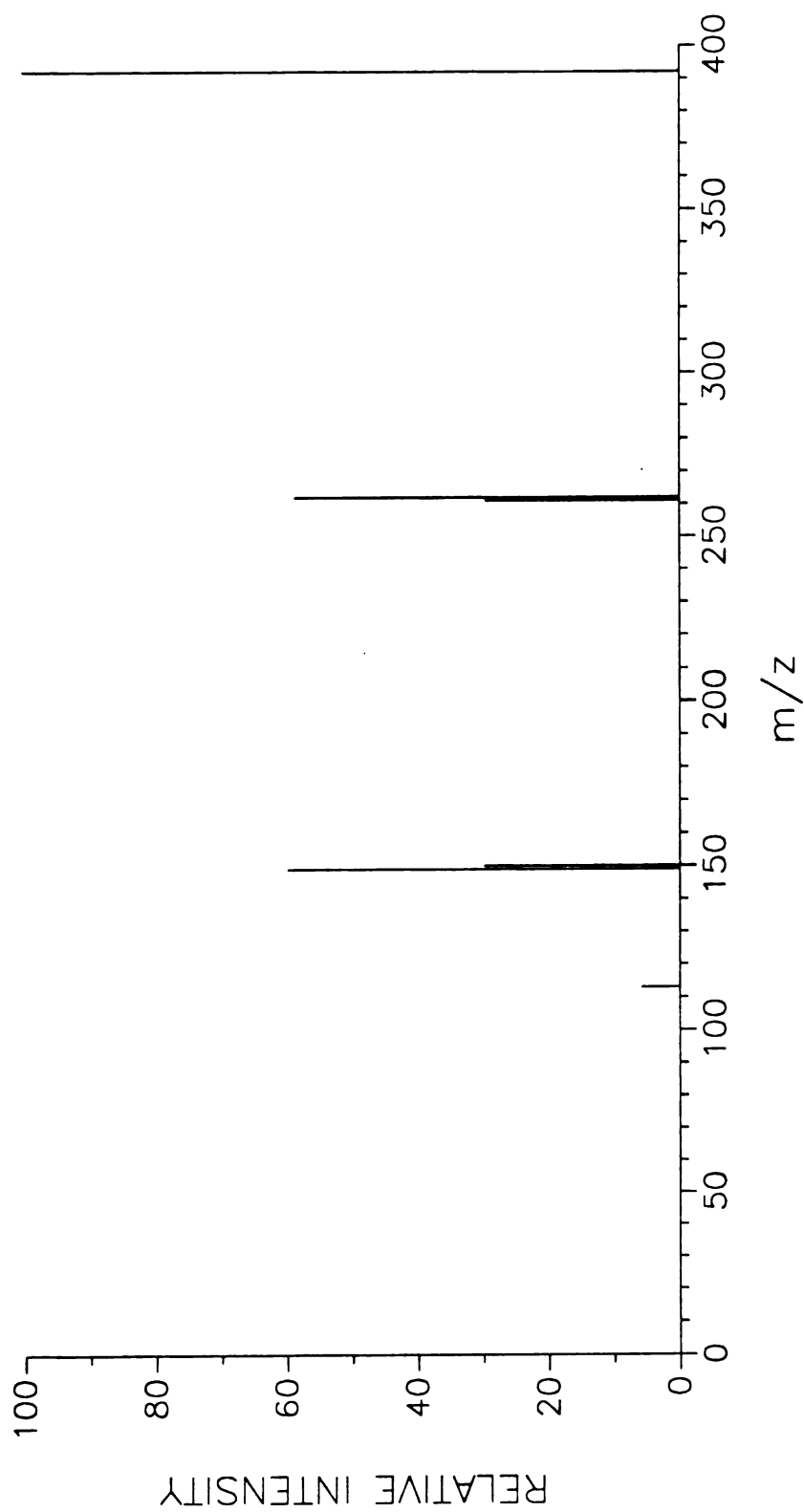


Figure 1.9 Daughter spectrum of the ^{13}C -containing protonated molecular ion from 1,2-benzenedicarboxylic acid, di-n-octyl-ester.

spectra and substructures. Most importantly, it required the mass spectroscopist to have some knowledge of the *ionic structure(s)* represented by the parent ions of daughter spectra to *confirm the reliability of the daughter spectrum/substructure relationships*. At the time this approach was abandoned, the reference database had few daughter spectrum/substructure correlations. Thus, only a limited number of substructures could be identified in unknowns. The example above certainly proves this point, as complete structure elucidation hinged on the identification of not one but two n-octyl substructures. Identification of these substructures was achieved manually through standard methods of spectral interpretation. Another inherent disadvantage to this system involved the use of traditional MS spectral matching algorithms for matching daughter spectra. Normal spectral matching routines weigh intensities rather heavily in the calculation of an overall match factor. However, intensities from daughter spectra are dependent on a large number of instrumental parameters and are certainly not as important as the presence or absence of a particular spectral feature for identifying a substructure. Perhaps the most serious drawback to correlating substructures with daughter spectra is that it does not take *full advantage of the extra dimension of information that MS/MS affords*. It uses the patterns in only one of the several types of data available from MS/MS spectra or three-dimensional fragmentation maps.

The "next generation" of this structure elucidation system had to address these deficiencies to achieve superior performance. A crucial need for this next generation system was some methodology for using *all*

the information available from MS/MS data for deducing spectral feature/substructure correlations. It must be emphasized that information characteristic of a given substructure is not limited to only the daughter spectrum of its closely related ionic structure, but can also appear as neutral losses and daughter ions in daughter spectra of other ionic structures which may contain part or all of that substructure. In addition, a goal of the next generation structure elucidation system was to include some methodology whereby spectral feature/substructure relationships (as opposed to daughter spectrum/substructure relationships) could be automatically deduced and verified. Also, the number of spectral feature/substructure correlations had to be expanded to cover a wider variety of substructures.

An Automated Chemical Structure Elucidation System using MS/MS Data

The next generation of this structure elucidation came to be known as ACES (Automated Chemical structure Elucidation System). It addresses many of the deficiencies in the original system. The individual components and data pathways of ACES are shown in Figure 1.10 (67). A TQMS instrument is used as the source of MS and MS/MS data. Several software tools are utilized by this system, including MAPS (Method for Analyzing Patterns in Spectra), the MFG program, and GENOA. ACES also requires two databases, the MS and MS/MS database and the rule base. Dr. Adrian Wade was responsible for the development of the original version of the MAPS code. Kevin Hart has

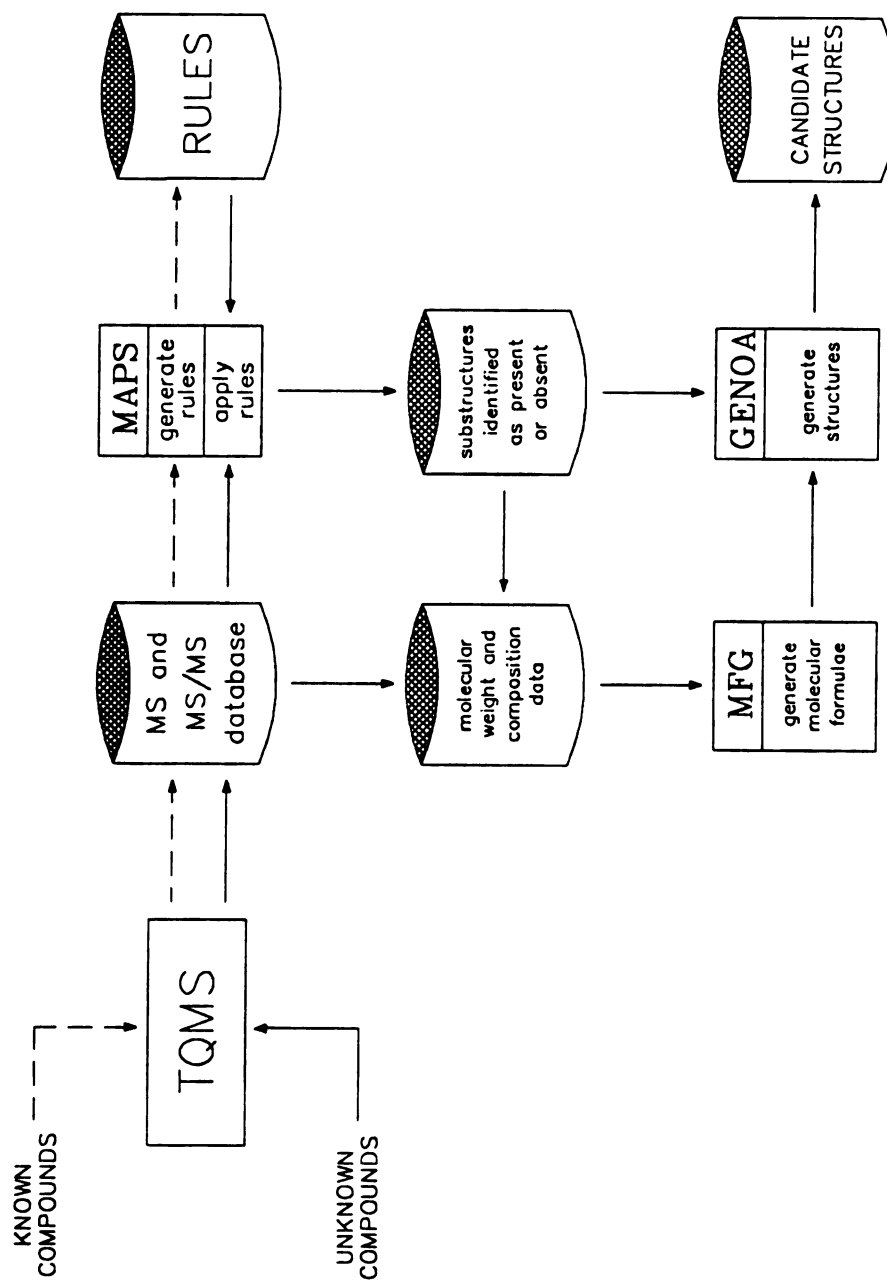


Figure 1.10 Schematic of ACES (Automated Chemical structure Elucidation System).

been involved in modifying the source code to GENOA to better suit the purposes of this structure elucidation system (70). My contributions to this system include the database of MS and MS/MS spectra from which the MAPS rules were derived, and the MFG program and other software for molecular formula determination from unit resolution MS and MS/MS data. I have made substantial modifications to the MAPS code to predict the absence as well as the presence of substructures. I have also made many additional modifications which have greatly improved the predictive capabilities of the rules.

The "heart" of ACES is the MAPS software. As shown in Figure 1.10, this software operates in two different modes: the learning mode (dashed line) and the identification mode (solid line). In the learning mode, MAPS uses data from known compounds to identify the relationships between substructures and their characteristic MS and MS/MS spectral features (71). These features originally included neutral losses, parent ions, daughter ions, and parent-to-daughter transitions. The relationships discovered are stored in the form of rules, which relate each substructure to its distinctive MS/MS spectral features. In the identification mode, these rules are applied to MS and MS/MS spectra from an unknown to predict the presence and absence of substructures. MAPS is an indirect database or interpretive method for substructure identification. It is a totally empirical scheme which assumes no knowledge about the fragmentation process, ion structures or rearrangements. MAPS uses heuristics to remove false correlations from the rules and improve their predictive capabilities. As opposed to the original structure elucidation system, MAPS *automatically derives*

spectral feature/substructure relationships and uses all of the information available from MS/MS data for substructure identification.

In the application of the ACES tools to an unknown, the MAPS rules are applied to the unknown's MS and MS/MS data to identify the presence and absence of as many substructures as possible. The MFG program is then invoked to generate all possible formulae consistent with the molecular weight and elemental composition constraints derived from MS data, MS/MS data, and the substructures identified as present and absent. These formulae and substructural constraints are then supplied as input to GENOA which generates all possible structures.

The original MAPS code as developed by Dr. Wade is discussed in Chapter 3. The modifications that I have incorporated into this code are discussed in Chapter 4. An approach for reliable substructure identification which involves the use of specific combinations of MS and MS/MS spectral features is described in Chapter 5. The methodology for molecular formula determination from unit resolution MS and MS/MS data, the MFG program, and additional software tools for assisting in this process are described in Chapter 6. The utility of ACES and the interaction of these software tools for structure elucidation of several test compounds is demonstrated in Chapter 7. Chapter 8 contains suggestions for future work for improving the capabilities of ACES for structure elucidation.

Up to this writing, MS/MS has been mainly used for identification of target compounds in complex mixtures and has not been fully exploited for structure elucidation. ACES has addressed a substantial deficiency in this area of research, combining techniques of pattern

recognition and artificial intelligence with the power of MS/MS for structure elucidation. The potential utility of ACES, and in particular, the MAPS and MFG programs for structure elucidation will be demonstrated in this dissertation.

References

1. Thomson, J.J., "Rays of Positive Electricity and Their Application to Chemical Analyses", Longmans, Green and Co., London, 1913.
2. Hirschfeld, T., Anal. Chem., **52**, 297A (1980).
3. Small, G.W., Anal. Chem., **59**, 535A (1987).
4. Hertz, H.S., Hites, R.A., Biemann, K., Anal. Chem., **43**, 681 (1971).
5. Damen, H., Henneberg, D., Wiemann, B., Anal. Chim. Acta, **103**, 289 (1978).
6. Henneberg, D., Adv. Mass Spectrom., **8**, 1511 (1980).
7. Blaisdell, B.E., Sweeley, C.C., Anal. Chim. Acta, **117**, 17 (1980).
8. Blaisdell, B.E., Gates, S.C., Martin, F.E., Sweeley, C.C., Anal. Chim. Acta, **117**, 35 (1980).
9. McLafferty, F.W., Stauffer, D.B., J. Chem. Inf. Comput. Sci., **25**, 245 (1985).
10. Wangen, L.E., Woodward, W.S., Isenhour, T.L., Anal. Chem., **43**, 1605 (1971).
11. Knock, K., Venkataraghavan, R., McLafferty, F.W., J. Am. Chem. Soc., **95**, 4185 (1973).
12. Abramson, F.P., Anal. Chem., **47**, 45 (1975).
13. Rasmussen, G.T., Isenhour, T.L., J. Chem. Inf. Comput. Sci., **19**, 179 (1979).
14. Martinsen, D.P., Song, B.H., Mass Spectrom. Rev., **4**, 461 (1985).
15. Spencer, R.B., Stauffer, D.B., 32nd Annual Conference on Mass Spectrometry and Allied Topics, 1984, p. 658.
16. Sharaf, M.A., Illman, D.L., Kowalski, B.R., "Chemometrics", Chemical Analysis Series, Vol. 82, Wiley & Sons, New York, 1986.
17. Isenhour, T.L., Kowalski, B.R., Jurs, P.C., CRC Crit. Rev. Anal. Chem., **4**, 1 (1984).
18. Jurs, P.C., Isenhour, T.L., "Chemical Applications of Pattern Recognition", Wiley & Sons, NY, 1975.

19. Wolff, D.D., Parsons, M.L., "Pattern Recognition Approach to Data Interpretation", Plenum Press, NY, 1983.
20. Zupan, J., Novic, M., in Zupan, J. (Ed.), "Computer-Supported Spectroscopic Databases", Wiley & Sons, NY, 1986.
21. Schecter, J., Jurs, P.C., Appl. Spect., 27, 255 (1973).
22. Zander, G.S., Jurs, P.C., Anal. Chem. 47, 1562 (1975).
23. Lohninger, H., Varzuma, K., Anal. Chem., 59, 236 (1987).
24. Wilkins, C.L., Isenhour, T.L., Anal. Chem., 47, 1849 (1975).
25. Abe, H., Kumazawa, S., Tajl, T., Sasaki, S., Biomed. Mass Spectrom., 3, 151 (1976).
26. MacLagan, R.G.A.R., Mitchell, M.J., Aust. J. Chem., 33, 1401 (1980).
27. Weber, J.J., Van Thuijl, J., DeJong, H.J., Anal. Chim. Acta, 188, 195 (1986).
28. Lowry, S.R., Isenhour, T.L., Justice, J.B., McLafferty, F.W., Dayringer, H.E., Venkataraghavan, R., Anal. Chem., 49, 1720 (1977).
29. Dayringer, H.E., Pesyna, G.M., Venkataraghavan, R., McLafferty, F.W., Org. Mass Spectrom., 11, 529 (1976).
30. Haraki, K.S., Venkataraghavan, R., McLafferty, F.W., Anal. Chem., 53, 386 (1981).
31. Graham, N., "Artificial Intelligence - Making Machines Think", TAB Books Inc., Blue Ridge Summit, PA, 1979.
32. Barr, A., Feigenbaum, E.A., "The Handbook of Artificial Intelligence", Heuristech Press, Stanford, CA, 1982.
33. Dessy, R.A., Anal. Chem., 56, 1200A (1984).
34. Dessy, R.A., Anal. Chem., 56, 1236A (1984).
35. Wade, A.P., Crouch, S.R., submitted to Anal. Chim. Acta.
36. Pierce, T.H., Hohne, B.H., "Artificial Intelligence Applications in Chemistry", ACS Symposium Series No. 306, American Chemical Society, Washington, D.C., 1986.
37. Barr, A., Feigenbaum, E.A., "The Handbook of Artificial Intelligence", Vol. II, Heuristech Press, Stanford, CA, 1982.

38. Buchanan, B.G., Smith, D.H., White, W.C., Gritter, R.J., Feigenbaum, E.A., Lederberg, J., Djerassi, C., J. Am. Chem. Soc., **98**, 6168 (1976).
39. Lindsay, R.K., Buchanan, G.B., Feigenbaum, E.A., Lederburg, J., "Applications of Artificial Intelligence for Organic Chemistry - The Dendral Project", McGraw-Hill, New York, 1980.
40. Carhart, R.E., Smith, D.H., Gray, N.A.B., Nourse, J.G., Djerassi, C., J. Org. Chem., **46**, 1708 (1981).
41. Smith, D.H., Gray, N.A.B., Nourse, J.G., Crandell, C.W., Anal. Chim. Acta, **133**, 471 (1981).
42. Moldoveanu, S., Rapson, C.A., Anal. Chem., **59**, 1207 (1987).
43. Sasaki, S., Abe, H., Hirota, Y., Ishida, Y., Kudo, Y., Ochiai, S., Saito, K., Yamasaki, T., J. Chem. Inf. Comput. Sci., **18**, 211 (1978).
44. Sasaki, S., Fujiwara, I., Abe, H., Yamasaki, T., Anal. Chim. Acta, **122**, 87 (1980).
45. Oshima, T., Ishida, Y., Saito, K., Sasaki, S. Anal. Chim. Acta, **122**, 95 (1980).
46. Yamasaki, T., Abe, H., Kudo, Y., Sasaki, S., in Smith, D.H. (Ed.), "Computer-Aided Structure Elucidation", American Chemical Society, Washington, DC, 1977, p. 108.
47. Hippe, Z., Anal. Chim. Acta, **150**, 11 (1983).
48. McLafferty, F.W., "Interpretation of Mass Spectra", University Science Books, Mill Valley, CA, 1980.
49. McLafferty, F.W., Venkataraghavan, R., "Mass Spectral Correlations", American Chemical Society, Washington, DC, 1982.
50. Bozorgzadeh, M.H., Morgan, R.P., Beynon, J.H., Analyst, **103**, 613 (1978).
51. Yost, R.A., Enke, C.G., Anal. Chem., **51**, 1251A (1979).
52. Yost, R.A., Enke, C.G., McGilvery, D., Smith, D., Morrison, J.D., Int. J. Mass Spectrom. Ion Proc., **30**, 127 (1979).
53. Yost, R.A., Enke, C.G., Org. Mass Spectrom., **16**, 171 (1981).
54. Newcome, B.H., Enke, C.G., Rev. Sci. Instrum., **55**, 2017 (1984).
55. Newcome, B.H., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1984.

56. Myerholtz, C.A., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1982.
56. Myerholtz, C.A., Schubert, A.J., Kristo, M.J., Enke, C.G., Instruments and Computers, **6**, 11 (1985).
58. Myerholtz, C.A., Schubert, A.J., Kristo, M.J., Enke, C.G., Instruments and Computers, **6**, 13 (1985).
59. Novix Inc., Cupertino, CA.
60. Golden, J.H., Moore, C.H., Brodie, L., Electronic Design, March 21, 1985.
61. Kristo, M.J., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1987.
62. Gregg, H.R., Hoffman, P.A., Enke, C.G., Crawford, R.W., Brand, H.R., Wong, C.M., Anal. Chem., **56**, 1121 (1984).
63. Gregg, H.R., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1987.
64. Cross, K.P., Enke, C.G., Comput. and Chem., **10**, 175 (1986).
65. Cross, K.P., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1986.
66. Cross, K.C., Palmer, P.T., Giordani, A.B., Beckner, C.F., Hoffman, P.A., Gregg, H.R., Enke, C.G., in Pierce, T.H., Hohne, B.H. (Eds.), "Artificial Intelligence Applications in Chemistry", ACS Symposium Series No. 306, American Chemical Society, Washington, D.C., 1986, p. 321.
67. Enke, C.G., Wade, A.P., Palmer, P.T., Hart, K.J., Anal. Chem., **59**, 1363A (1987).
68. Hoffman, P.H., Enke, C.G., 31st Annual Conference on Mass Spectrometry and Allied Topics, 1983, p. 556.
69. Palmer, P.T., Enke, C.G., submitted to Int. J. Mass Spectrom. Ion Proc.
70. Hart, K.J., Palmer, P.T., Enke, C.G., to be presented at 36th Annual Conference on Mass Spectrometry and Allied Topics, 1988.
71. Wade, A.P., Palmer, P.T., Hart, K.J., Enke, C.G., accepted for publication in Anal. Chim. Acta.

CHAPTER 2

MS/MS SPECTRAL LIBRARY ACQUISITION

MS Libraries

The prominence of computer-controlled instrumentation has greatly enhanced the use of spectral matching methods for interpretation. Of the three main techniques currently used for structure elucidation, mass spectrometry, infrared spectroscopy, and nuclear magnetic resonance spectroscopy, the development of spectral databases and search algorithms has occurred most rapidly for mass spectrometry. This can be attributed to the popularity of GC/MS for mixture analysis. Given that several hundred mass spectra can be generated during a single GC/MS run, automated techniques are essential to alleviate the bottleneck in spectral interpretation. A substantial amount of research has been invested in improving the performance of mass spectral matching techniques over the last three decades. There are a variety of systems for MS spectral matching. Some of these systems and their concepts have been described in Chapter 1. This section provides a description of the current status of MS databases along with their advantages and limitations.

There are two types of libraries of mass spectra: small, specific libraries and large, comprehensive libraries. Small libraries usually

consist of up to a few thousand spectra and are usually tailored for specific compound classes or applications. Large libraries represent collections of mass spectra of arbitrary compounds obtained from a variety of sources. There are two large libraries of mass spectra currently in use: the Wiley and NBS databases. The Wiley database contains more than 123,000 spectra of 108,000 different compounds (1). This database is available on a single CD-ROM disk which includes McLafferty's PBM/STIRS search and retrieval software (2). The NBS database, developed at the National Bureau of Standards in conjunction with the Environmental Protection Agency (EPA) and the National Institutes of Health (NIH), contains more than 80,000 nonredundant mass spectra (3,4). This database is part of the Mass Spectral Search System (MSSS) and can be accessed through the Chemical Information System (5).

Large databases are usually created from several smaller databases. Generally, these smaller databases differ in quality, resulting in a larger database which is non-homogeneous. Shelley states that "Although many spectral databases have been created, few are of high quality and many are useless" (6). The larger mass spectral databases are being employed in a variety of applications in industry and academia and are certainly useful. It is obvious that the results obtained from spectral matching methods depend on the quality of the reference spectra. However, the quality of spectra in these databases is often questionable. Possible problems include unnecessarily low or inconsistent lower mass limits of spectra, truncated intensities, poor

dynamic range, and contamination. Henneberg has demonstrated the effects of these problems on mass spectra in a recent review article (7).

Several methodologies have been developed to address this problem of low quality mass spectra. McLafferty's group has developed an algorithm for calculating a quality index which may be used to identify dubious spectra or select the best spectrum from a set of duplicate spectra (8). This quality index is calculated from seven factors which include the source of the spectrum, ionization conditions used, high molecular weight impurities, illogical neutral losses, isotopic abundance accuracy, number of peaks, and the lower mass limit. Milne and coworkers have combined this quality index with procedures for ensuring sample purity and stringent control of instrumental parameters for generating high quality mass spectra (10,11). This modified quality index has been applied to the spectra in the NBS database and duplicate, lower quality spectra have been removed (11). Heller cautions that a quality index "is not really an indicator that a spectrum is a good one. Rather it is an indication of problems with a spectrum" (12). Dillard and coworkers have reported some of the criteria for obtaining high quality mass spectra. These criteria include the operating conditions of the MS instrument, the parameters to be included along with the spectra, and the factors required for evaluation of spectra (13). This report concludes that although class III spectra (analytical reference spectra) and class II spectra (research reference spectra) can be obtained using current instrumentation and methodologies, *class I spectra, which are independent of instrumentation, are not yet attainable.*

A serious shortcoming of many mass spectral databases is that they do not include representations of each compound's chemical structure. The representation of a chemical structure is the only way to uniquely identify a compound. Structure representations are required for retrieving similar structures and searching for specific substructures in a database. They also allow spectrum/structure or spectrum/substructure correlations to be made. These correlations are often of great utility for structure elucidation. There are several different methodologies for representation of chemical structure, most of which can be classified as linear or connectivity table notations (14). Linear notations, which represent structures using alphanumeric strings, are more efficient in terms of storage space. Wiswesser line notation (WLN), which is used in the Wiley database for structure representation, is one of the more well-known linear notation systems (15). Unfortunately, WLN is not a canonical notation scheme for chemical structures; that is, a given WLN may give rise to more than one structure. Other linear notation schemes for unique, nonambiguous representation of chemical structures have been developed (16). Linear notation schemes are often very difficult to manipulate for performing substructure searching. Gray notes that "Linear notations were developed for information retrieval, not computer-based structure analysis" (14). Connectivity tables are generally more suitable for structure manipulations although they require more storage space. These represent structures by describing the connections between each atom using a unique, nonambiguous notation scheme. The more complex connectivity table notation methods can represent configurational and stereochemical isomers (17).

Connectivity tables are used in GENOA for structure and substructure representation. The NBS database includes the Chemical Abstracts Service registry number and the associated connectivity table for each compound.

In the past, many databases were abbreviated and did not use all the spectral data due to the high cost of storage media. With inexpensive mass storage devices currently available, this should no longer be a problem. A serious shortcoming of many databases is the lack of structural diversity, which can lead to erroneous spectrum/substructure correlations. Although redundant spectra in databases are often considered to be a drawback, McLafferty and Stauffer have noted that the reliability of spectral matching results is improved when multiple spectra are included in the database (18). The present status of MS databases and their associated problems is best described by Shelley, who stated that "The spectral databases that have been created have been plagued with problems. It is hoped that future database developers will pay more attention to quality than quantity" (6).

MS/MS Libraries

The utility of a database of collisionally activated dissociation (CAD) mass spectra for structure elucidation was first proposed by Beynon's group in 1978 (19). A database of such spectra could be used for comparison with unknown CAD spectra in a manner analogous to the use of EI mass spectral libraries. Furthermore, a CAD spectral database could be used to identify substructures and/or the ionic structures of

fragment ions from unknown CAD spectra. Although several groups have recently created or proposed the creation of databases of reference CAD mass spectra (20,21), there are no applications of these databases in the literature. Our research group proposed collecting a database of such spectra at the ASMS meeting in 1983 (22).

Shelley has noted that "The lack of good databases ... will, at the least, slow the development of expert systems for computer-assisted structure elucidation" (6). This statement certainly applies to MS/MS, where the development of automated structure elucidation techniques has been hampered by the lack of databases of MS/MS spectra. There are several reasons for this. MS/MS is still a relatively new technique and the development of CAD spectral databases is in its infancy. The few databases that do exist are special purpose, proprietary, or not generally available. In addition, these databases are relatively small compared to the number of CAD spectra reported in the literature. Maintenance, certification, and expansion of spectral databases is extremely time-consuming and expensive work. *Most importantly, CAD spectra are strongly dependent on target gas thickness, parent ion energy, a variety of additional instrumental parameters, and the instrument itself* (23). Currently, the mass spectrometry community has still not reached an agreement on a set of practical, standard conditions for collecting CAD spectra (24). The effects of these instrumental parameters on CAD spectra are described next.

Instrumental Parameter Effects on Daughter Spectra

In Dawson's interlaboratory evaluation of the effects of operating conditions on CAD spectra, widely varying results were obtained from different TQMS instruments following a standardized procedure for collecting daughter spectra of specific parent ions (25). These results demonstrated that even standardized operating conditions do not adequately control the dynamics of the CAD process between different instruments.

Martinez has extensively studied the effects of the various kinetic and instrumental parameters which influence CAD spectra with the eventual goal of identifying the criteria for collecting a database of instrument-independent CAD spectra (24,26,27). The key instrumental parameters which cause the relative intensities of various daughter ions to differ significantly between different TQMS instruments include:

- 1) the type of ionization technique used;
- 2) the type of collision or "target" gas used;
- 3) the number of collisions undergone by a parent ion within the collision chamber, a parameter usually characterized in terms of "target thickness", which is equivalent to the effective number density of the target gas multiplied by the actual path length traversed by the parent ion through the collision chamber;

- 4) the interaction time between a parent ion and the target gas, which is determined by the collision energy (the potential difference between the ion source and the quad 2 offset);
- 5) the potential difference between the ion source and quad 1, which determines the energy distribution of ions entering the collision chamber and affects the resolution of quad 1;
- 6) the potential difference between the offsets of quads 2 and 3 (the quad 3 drawout), which determines which daughter ions enter quad 3 since daughter ions have a range of translational energies;
- 7) the RF voltage and field radius of quad 2; and
- 8) the type of detector used.

Martinez has recently identified several prerequisites for generating instrument-independent CAD spectra (26,27). He notes that unless the reaction dynamics of the CAD process are taken into account, CAD spectra will be highly instrument dependent even if reproducible instrumental operating conditions are maintained. The kinetics of CAD processes are an intrinsic property and thus can be used to determine the criteria for obtaining instrument-independent CAD spectra. He states that "the successful development of a generic, instrument-independent CAD spectral database for QQQ instruments necessitates that it be based on the measurement of absolute cross sections for the CAD of known ionic substructures" (28). In addition, single collision conditions must be used, since multiple collisions can lead to instrument-dependent CAD spectra (27). He also suggests that the

charge exchange reaction of argon be used to calibrate target gas thickness determinations and to verify that TQMS instruments are "kinetically well-behaved" (26,27). He proposes that a generic database of CAD spectra should consist of the target thickness and the identities of the fragmentation products for CAD of known ionic substructures.

Prior to my development of a CAD spectral database, little work had been published on the prerequisites for obtaining instrument-independent MS/MS spectra. Part of my work involved ascertaining the conditions necessary to obtain reproducible MS/MS spectra using an Extrel TQMS instrument in our laboratories. A diagram of this instrument is shown in Figure 2.1. A listing of the 25 instrumental parameters along with their logical device names, typical values, and scan limits is shown in Table 2.1. All of these parameters can be changed by the user and thus may have some effect on the data collected. The parameters which directly influence the ion path include the repeller, CI volume, EI volume, extractor lens, lens 1, lens 2, lens 3, quadrupole 1 offset, lens 4, quadrupole 2 offset, lens 5, quadrupole 3 offset, and the electron multiplier voltage. Other parameters which influence the quality of data include quadrupole operating parameters (DM1, DM3, RS1, RS3), and variables for the peak-finding algorithm, which include minimum and maximum peak widths (MWD and PWD), intensity threshold (THR) and rate of data acquisition. The effects of each of these parameters on daughter spectra were evaluated with the eventual goal of achieving reproducible MS/MS spectra on this instrument.

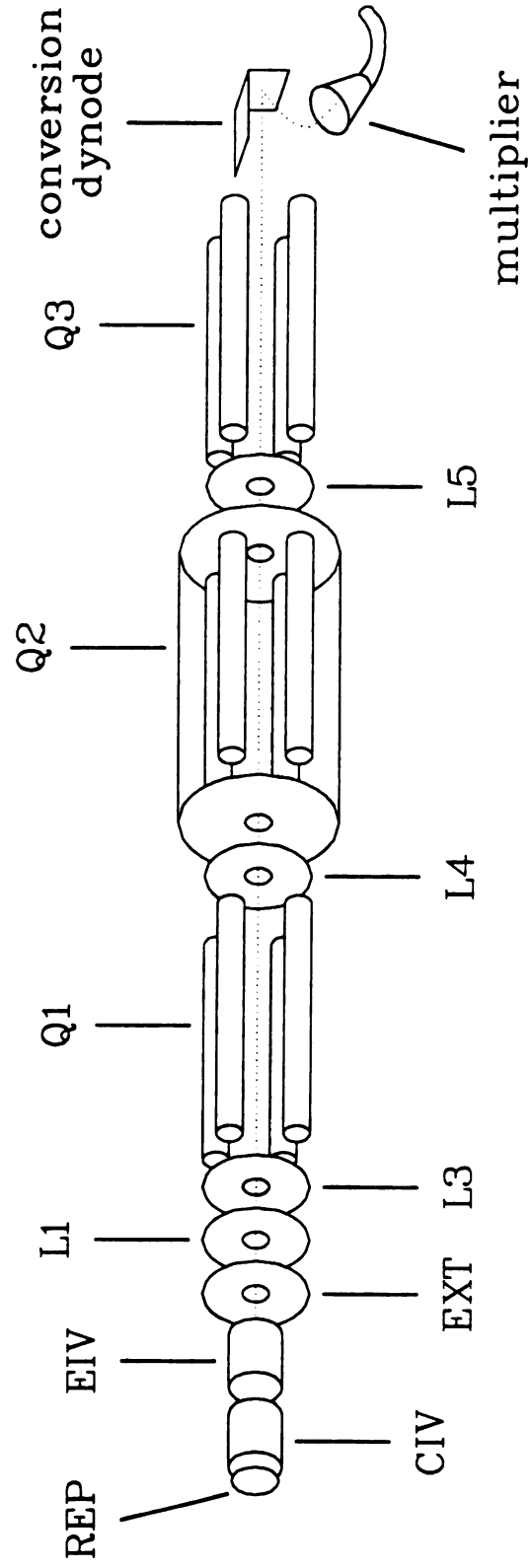


Figure 2.1 Schematic of the Extrel 400 series Triple Quadrupole Mass Spectrometer.

Table 2.1 Logical device names, typical values, and scan limits on the Extrel TQMS instrument.

Extranuclear, Inc. 12:20 9/3/85
 # 2 MS/MS TUNE MODE: EI

#	DEV	CURRENT	START	END
0	EV	70.0	0.0	100.0
1	REP	20.0	-100.0	35.0
2	CIV	18.8	-100.0	35.0
3	EIV	18.6	-100.0	35.0
4	EXT	-67.6	-200.0	200.0
5	L1	17.4	-200.0	200.0
6	L2	0.0	0.0	0.0
7	L3	-26.9	-200.0	200.0
8	Q1	4.9	-100.0	100.0
9	L4	-20.0	-100.0	100.0
10	Q2	3.0	-100.0	100.0
11	L5	-20.7	-100.0	100.0
12	Q3	0.0	-100.0	100.0
13	MHV	1900.0	0.0	3000.0
14	M1	77.0	10.0	280.0
15	M2	47.7	0.0	0.0
16	M3	94.0	10.0	110.0
17	DM1	-13.1	-100.0	100.0
18	DM3	-3.7	-100.0	100.0
19	RS1	6.3	-100.0	100.0
20	RS3	-2.4	-100.0	100.0
21	P2	452	0	0
22	THR	340	0	0
23	PWD	4	0	0
24	MWD	26	0	0
25	RTE	4	0	0

In my studies of the effects of the various instrumental parameters on daughter spectra, I found that *collision energy and collision gas pressure had the most significant effect on daughter spectra*. These parameters not only affected intensities, but the set of daughter ions produced. To establish standard collision energy and collision gas pressure ranges, their effects on the daughter spectra of the phthalate anhydride ion and several alkyl ions were studied. The results of these studies are presented below. In these studies, argon was used as the collision gas, collision gas pressure was measured using an ion gauge calibrated for argon connected to quad 2, and collision energy was calculated as the potential difference between the EI volume and quad 2.

Effects of Collision Gas Pressure and Collision Energy on Daughter Spectra of the Phthalate Anhydride Ion. The effects of collision gas pressure and collision energy on the daughter spectrum of the phthalate anhydride ion (m/z 149) from di-n-octyl phthalate, shown in Figure 2.2, were extensively studied on two different TQMS instruments: an Extrel model in our laboratories and a home-built TQMS instrument at Lawrence Livermore National Labs (LLNL) (29). In comparing results between these two instruments, one must keep in mind that they are constructed differently. For example, the LLNL instrument has an einzel lens set (a set of three lenses) between each quadrupole whereas the Extrel instrument has only one lens between each quadrupole. These instrumental variations certainly affect comparisons of data between these two instruments. In collecting data on these two instruments, it was not possible to rigorously standardize conditions between them.

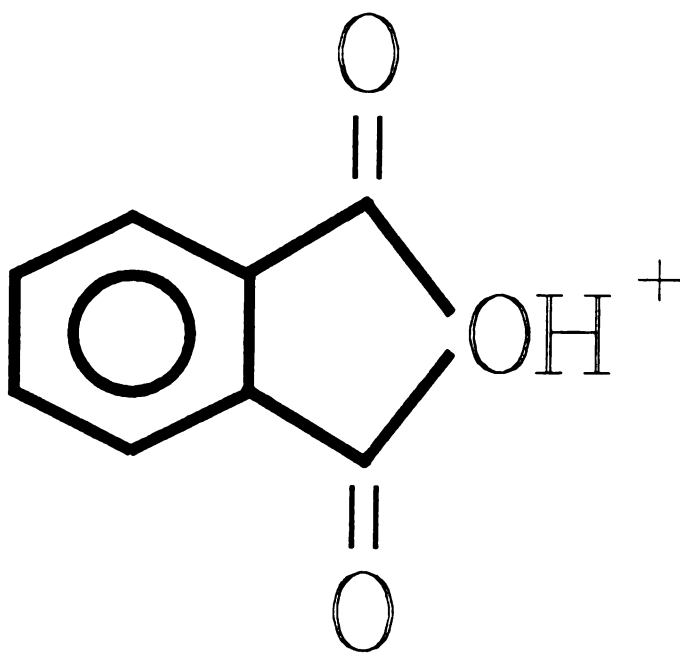


Figure 2.2 Structure of the phthalate anhydride ion.

Where possible, I have tried to hold independent variables constant when comparing results between these two instruments. Unless otherwise stated, the data which follow in this section are from the Extrel instrument.

Daughter spectra of the m/z 149 ion at three different collision gas pressures are shown in Figure 2.3, using a constant collision energy of 23.7 V. At a relatively low pressure of 0.052 mtorr, daughter ion intensities are all less than one percent of the base peak. At a pressure of 5.75 mtorr, the parent ion intensity is no longer the base peak and many different daughter ions are seen. Figure 2.4 shows daughter spectra of the same ion at three different collision energies using a collision gas pressure of 0.510 mtorr. From these spectra, it is evident that increasing the collision energy improves the efficiency of the fragmentation process. It is obvious that the standardization of collision gas pressure and energy is absolutely necessary to achieve consistent fragmentation.

Figures 2.5 and 2.6 are log-log plots of relative intensity versus collision gas pressure for the major daughter ions of the phthalate anhydride ion on the Extrel and LLNL instruments, respectively. The slope of such a log-log plot for a specific daughter ion is indicative of the reaction order, or the average number of collisions required to produce this ion. The reaction orders for formation of several daughter ions on these two instruments are shown in Table 2.2 below. These values were calculated using linear regression on the data points from Figures 2.5 and 2.6 at pressures between 0.1 and 1 mtorr.

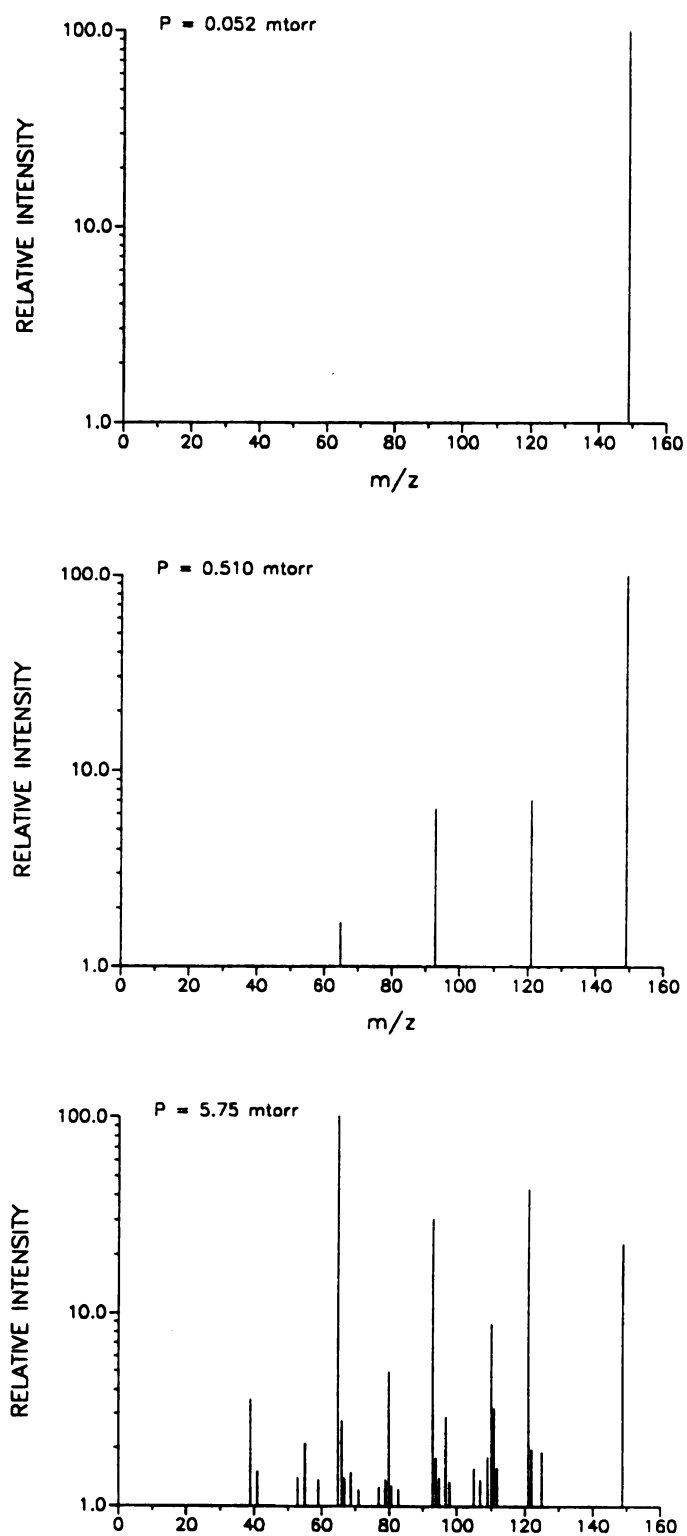


Figure 2.3 Daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate at three different collision gas pressures.

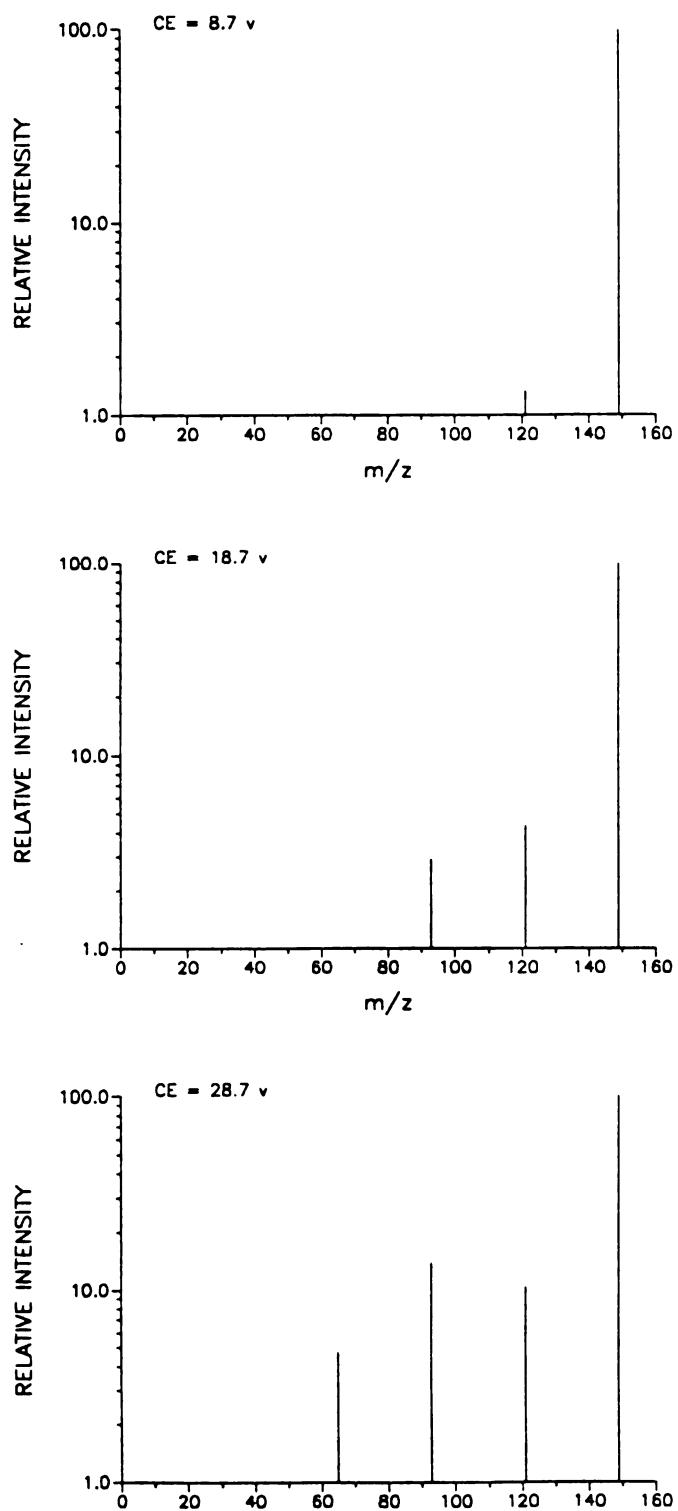


Figure 2.4 Daughter spectra of phthalate anhydride ion from di-n-octyl-phthalate the at three different collision energies.

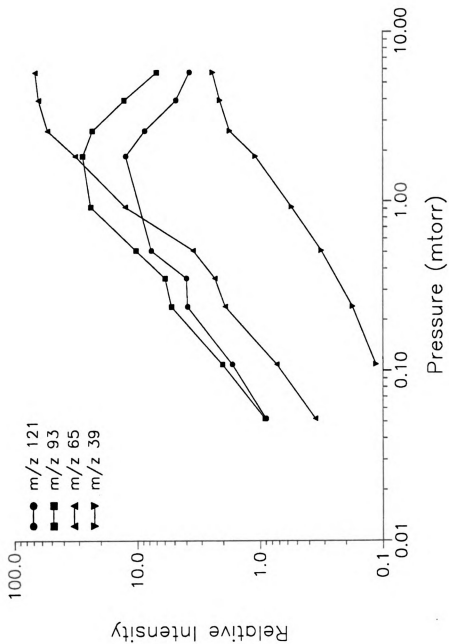


Figure 2.5 Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the phthalate anhydride ion from di-n-octylphthalate on the Extrel TQMS instrument.

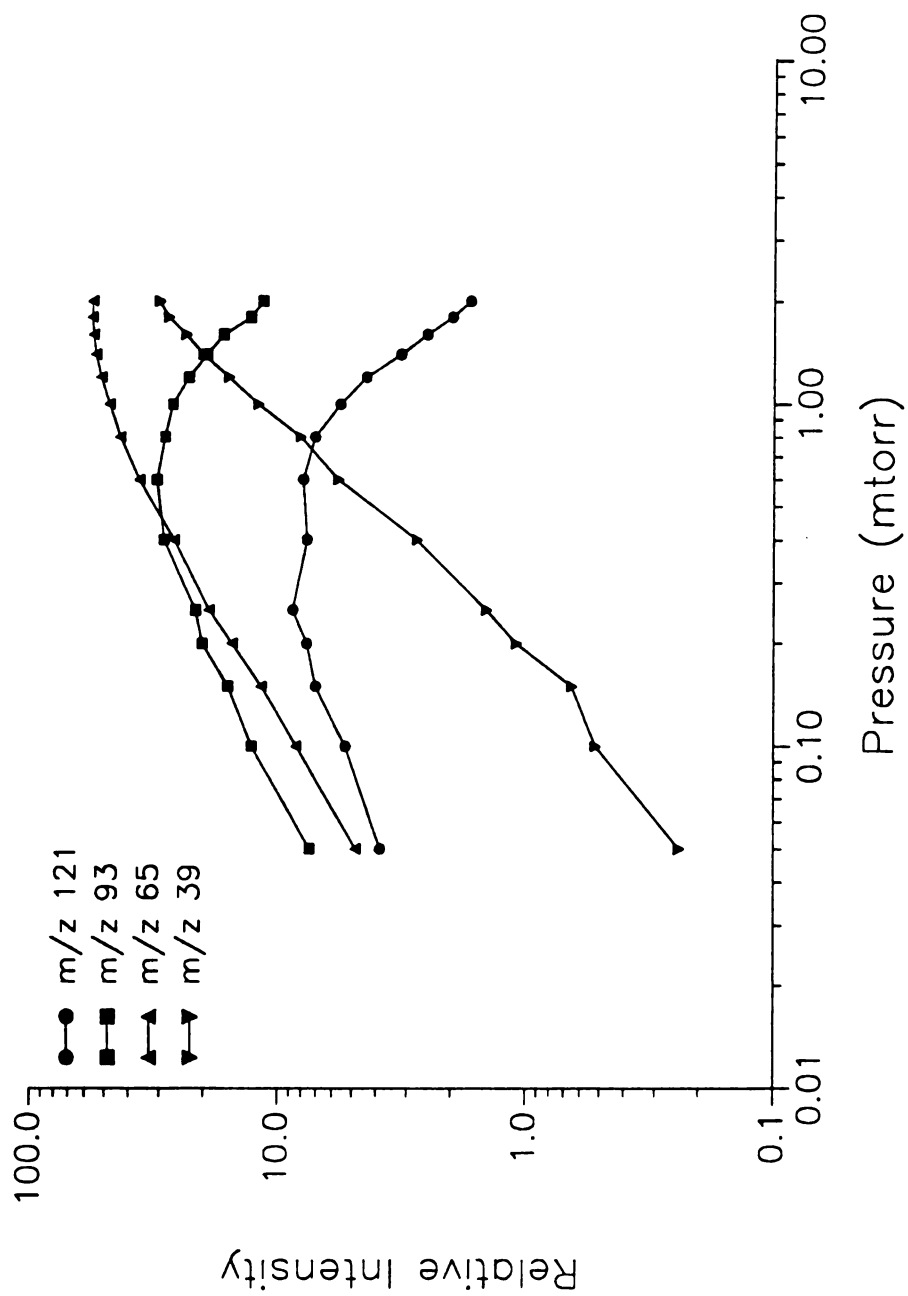


Figure 2.6 Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the phthalate anhydride ion from di-n-octylphthalate on the LLNL TQMS instrument.

Table 2.2 Reaction orders for the production of selected daughter ions from the phthalate anhydride ion on two different TQMS instruments.

Daughter Ion	Extrel Instrument	LLNL Instrument
121	0.90	0.51
93	1.0	0.67
65	1.0	0.84
39	0.80	1.1

Note that the slopes of these daughter ions are approximately one or less than one. This indicates that they are produced by either first order processes (slope = 1) or mixed metastable decompositions and first order processes (slope < 1). As the pressure increases, the slope of the m/z 121, 93 and 65 ions decreases, indicating that these ions are themselves undergoing further fragmentation. At high pressures, the slope of the m/z 39 ion increases, becoming equal to 1.4 at pressures above 1 mtorr on the LLNL instrument, indicating that it is being produced by both first and second order processes.

The general appearances of the plots for the individual daughter ions compare well between the two instruments. However, the relative intensities of the daughter ions is generally much greater at low pressures on the LLNL instrument as seen from Figures 2.5 and 2.6. This may be due to non-equivalent collision gas pressure readings between these instruments. It also may be attributed to the effects of different instrumental parameters between these two instruments, or to the different physical dimensions and designs of instruments themselves. It appears that the LLNL instrument has either better ion optics, higher

collection efficiencies at the exit of the collision chamber, higher fragmentation efficiencies, or some combination of these and other factors.

Figure 2.7 and 2.8 show the effects of collision energy on daughter spectra of the m/z 149 ion on the Extrel and LLNL instruments, respectively. From these spectra, it is apparent that higher collision energies provide higher daughter ion intensities. As collision energy is increased, more axial energy is imparted to the parent ion, increasing the efficiency of fragmentation and producing more daughter ions. This trend is seen on both instruments. Again, relative intensities of individual daughter ions do not compare favorably between the two instruments for the reasons mentioned above.

Collision gas pressure and energy affect the sensitivity of daughter spectra. Figure 2.9 plots the total ion count versus collision gas pressure for daughter spectra of m/z 149 at several collision energies. At higher pressures, scattering of ions in the collision chamber reduces the total ion count. Figure 2.10 is a plot of the daughter ion count versus the collision gas pressure for daughter spectra of m/z 149 at several collision energies. As pressure increases, more daughter ions are formed due to the increased fragmentation efficiency, and daughter ion count increases substantially. As the collision energy increases, ion abundances increase in both Figures 2.9 and 2.10. This is most likely due to the higher collection efficiency of ions at the entrance of quad 2 achieved at higher collision energies.

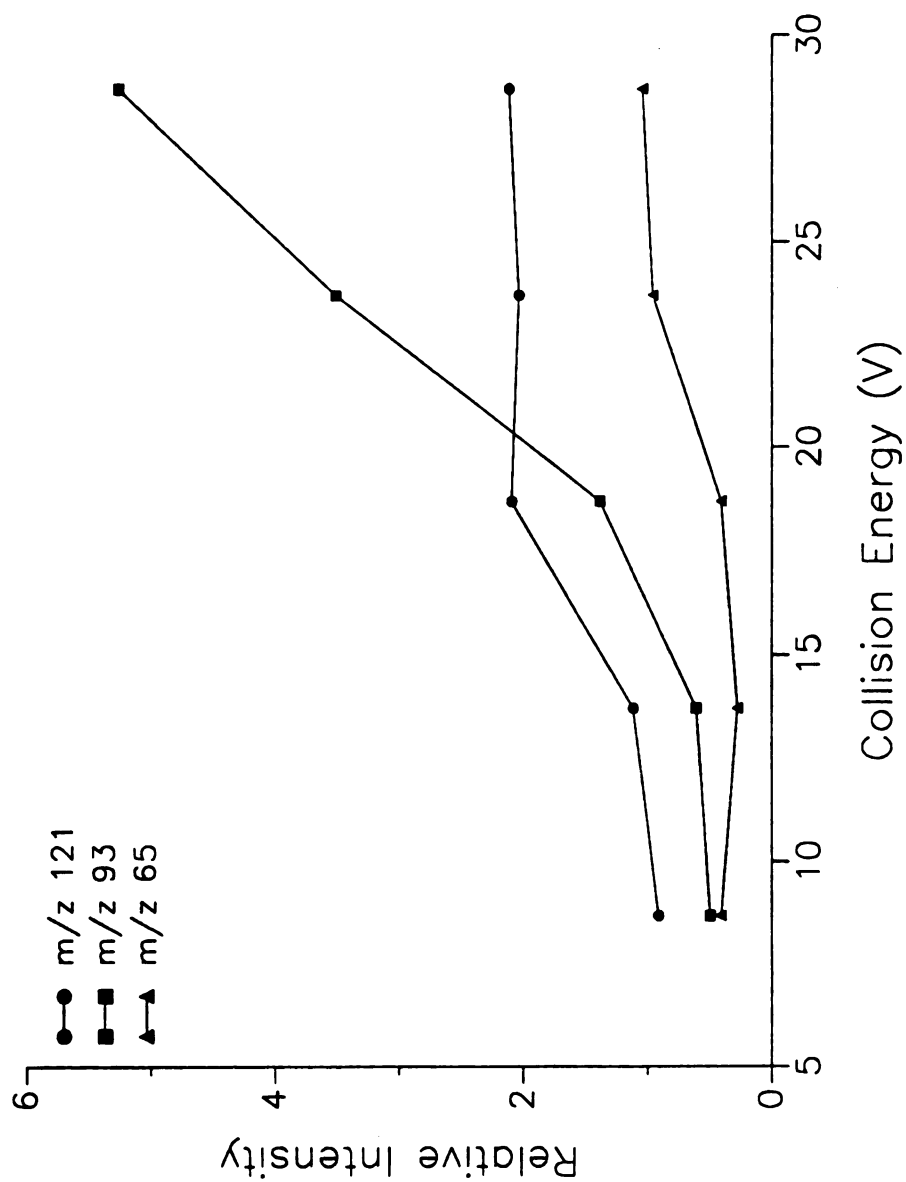


Figure 2.7 Plot of relative intensity versus collision energy for several daughter ions of the phthalate anhydride ion from di-n-octylphthalate on the Extrel TQMS instrument.

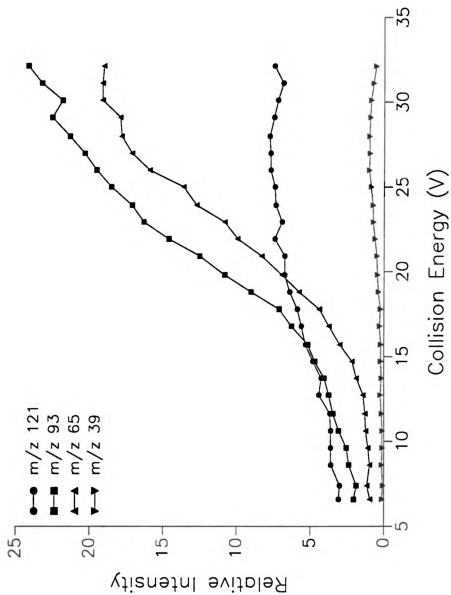


Figure 2.8 Plot of relative intensity versus collision energy for several daughter ions of the phthalate anhydride ion from di-n-octylphthalate on the LLNL TQMS instrument.

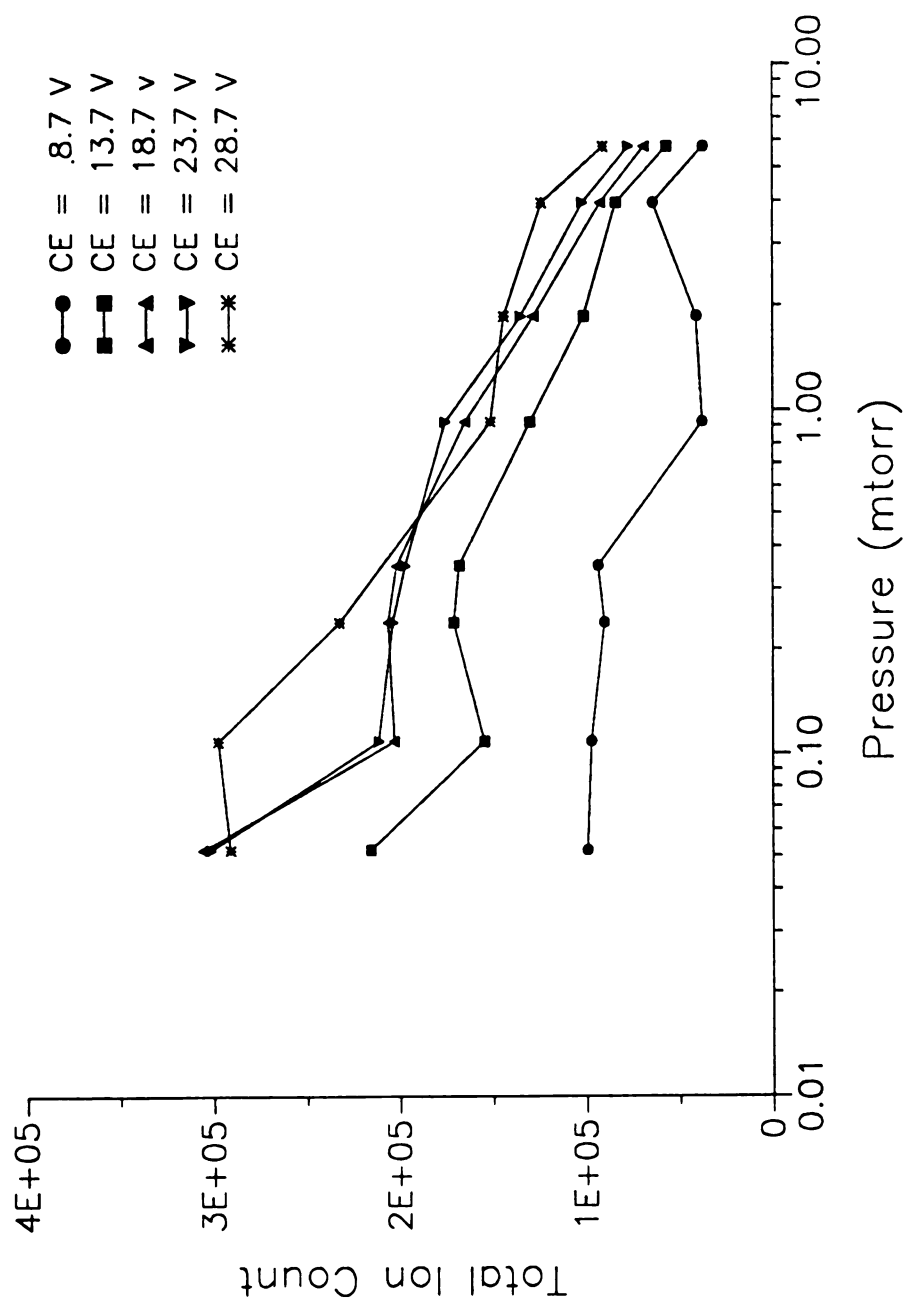


Figure 2.9 Plots of total ion count versus collision gas pressure at several collision energies for daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate.

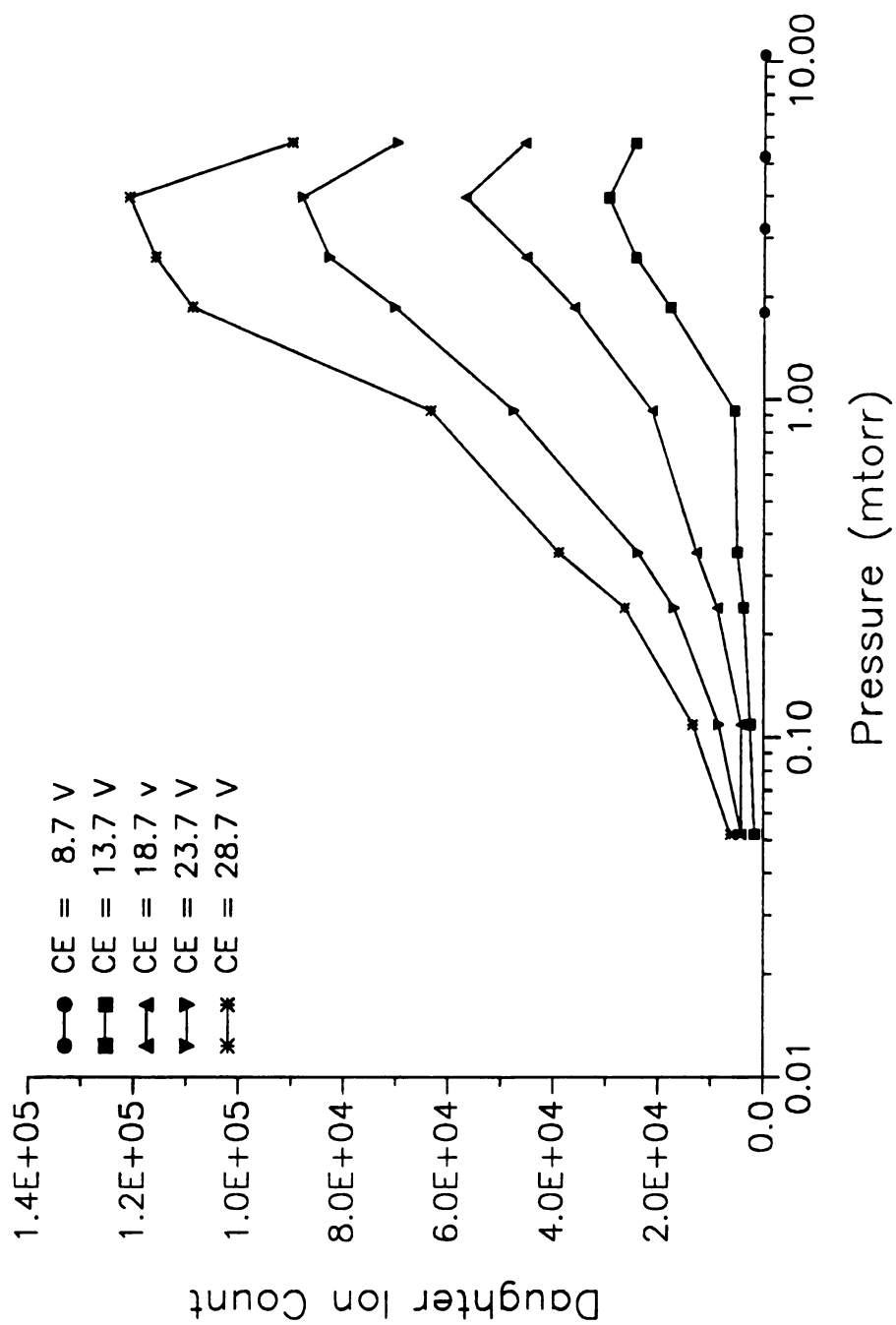


Figure 2.10 Plots of daughter ion count versus collision gas pressure at several collision energies for daughter spectra of the phthalate anhydride ion from di-n-octyl-phthalate.

Effects of Collision Gas Pressure and Collision Energy on Daughter Spectra of Alkyl Ions. Figures 2.11, 2.12, and 2.13 are log-log plots of relative intensity versus collision gas pressure for the major daughter ions of the m/z 71, 57, and 43 ions of n-heptane, respectively. At pressures of less than 1 mtorr, the slopes of these plots for individual daughter ions are between 0 and 1, indicating that these ions are produced by metastable and/or first order fragmentations. At pressures greater than 1 mtorr, second order fragmentations occur, as shown in Table 2.3.

Table 2.3 Reaction orders for the production of selected daughter ions from the m/z 71, 57, and 43 ions from n-heptane.

Parent Ion	Daughter Ion	Reaction Order

71	41	1.7
71	27	1.9
57	39	2.1
43	41	1.6
43	39	1.9
43	15	1.7

Figures 2.14, 2.15, and 2.16 show the effects of collision energy on the daughter spectra of the m/z 71, 57, and 43 ions from n-heptane, respectively. Unlike the corresponding collision energy versus intensity plots for daughter ions from the phthalate anhydride ion shown in Figures 2.7 and 2.8, the intensities of the daughter ions in these plots do not all increase as a function of collision energy. The intensities of many of the daughter ions shown in Figures 2.14, 2.15, and 2.16 decrease as a

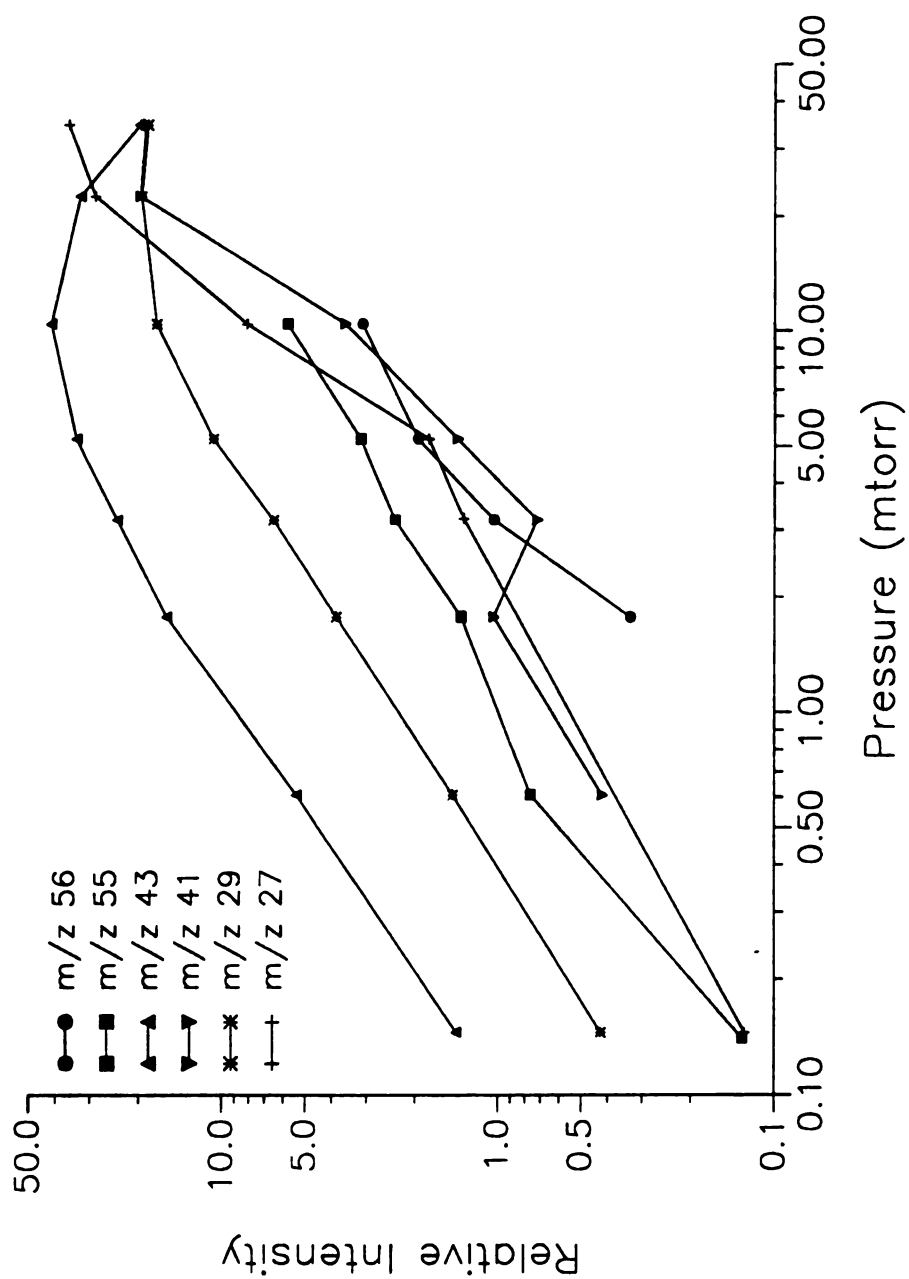


Figure 2.11 Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the m/z 71 ion from n-heptane.

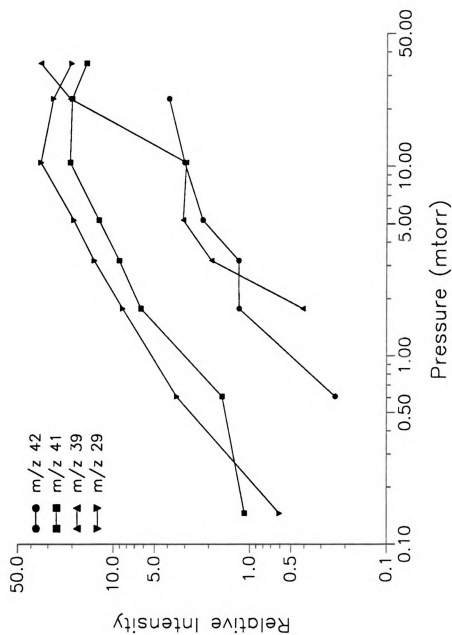


Figure 2.12 Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the m/z 57 ion from n-heptane.

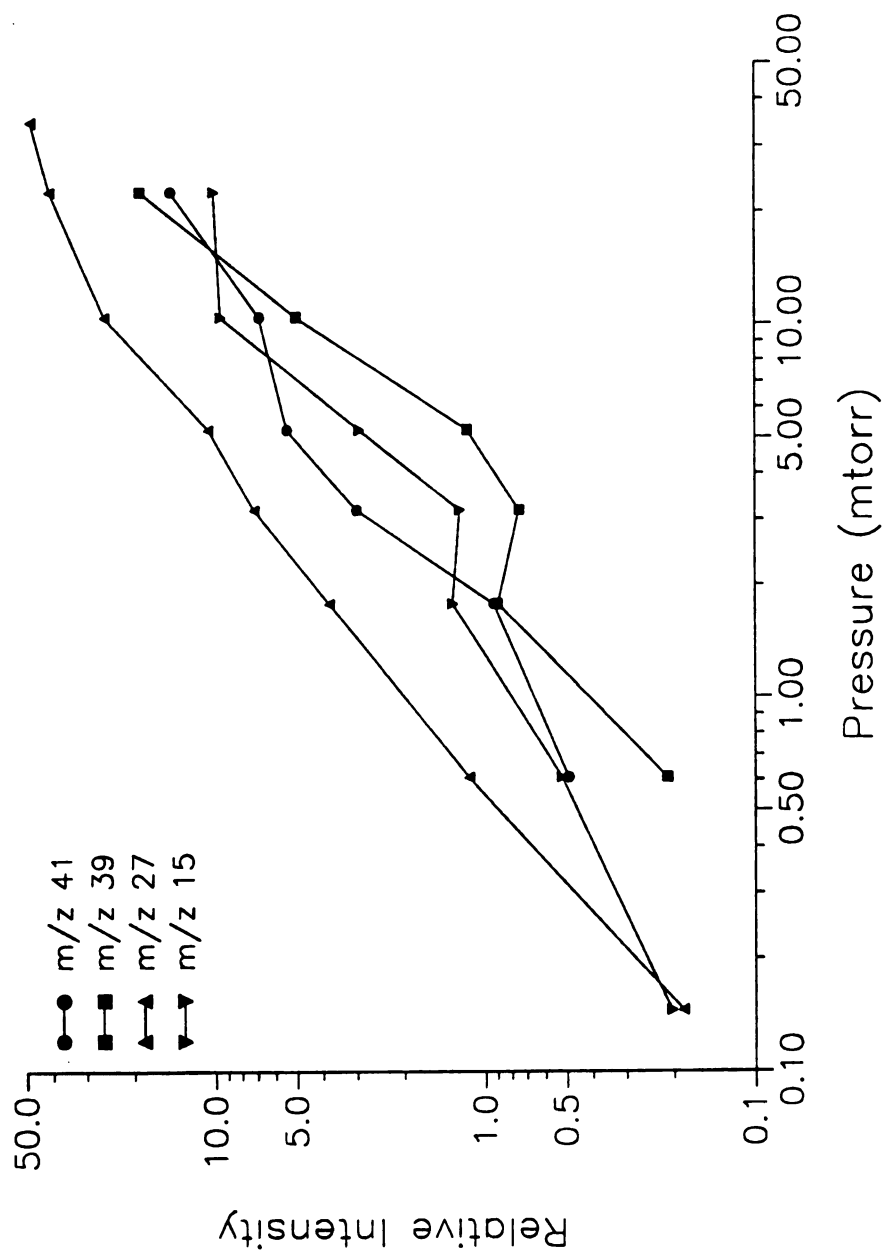


Figure 2.13 Log-log plot of relative intensity versus collision gas pressure for several daughter ions of the m/z 43 ion from n-heptane.

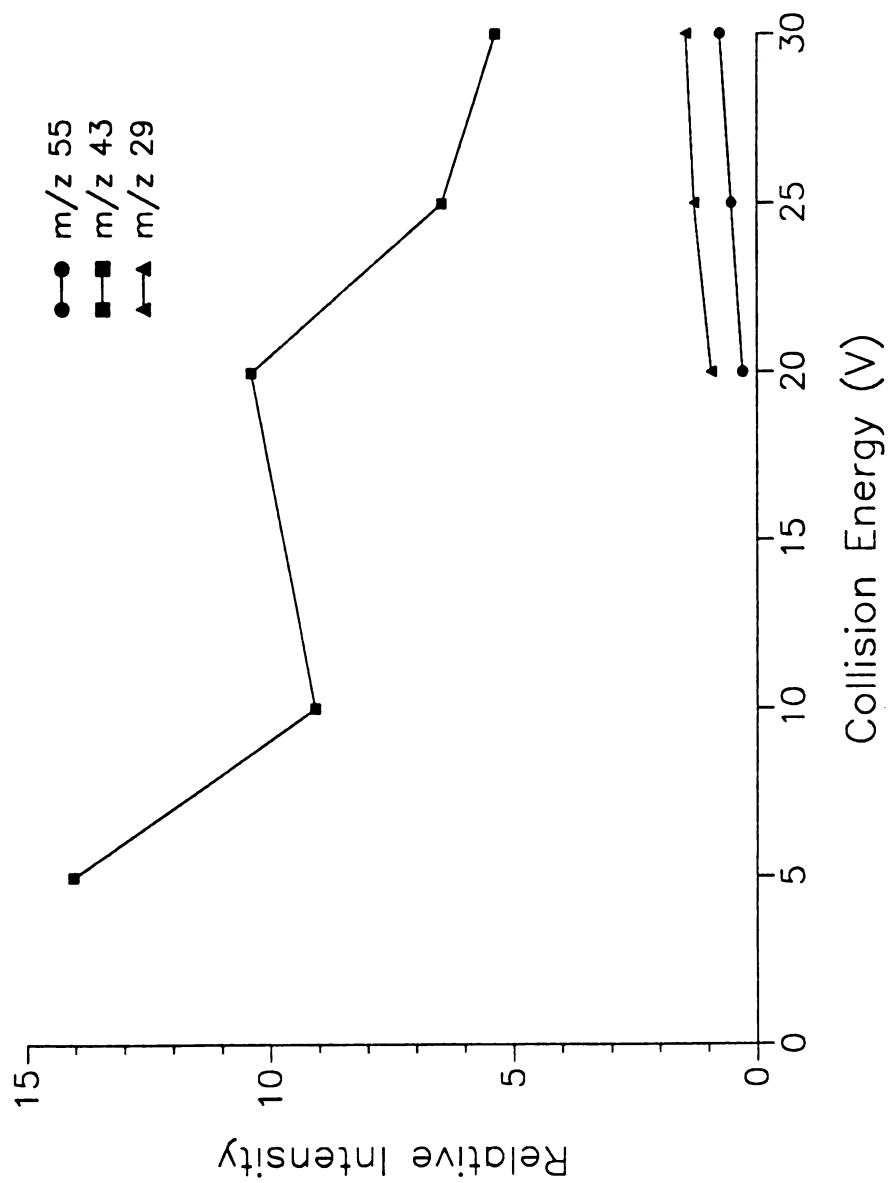


Figure 2.14 Plot of relative intensity versus collision energy for the m/z 71 ion from n-heptane.

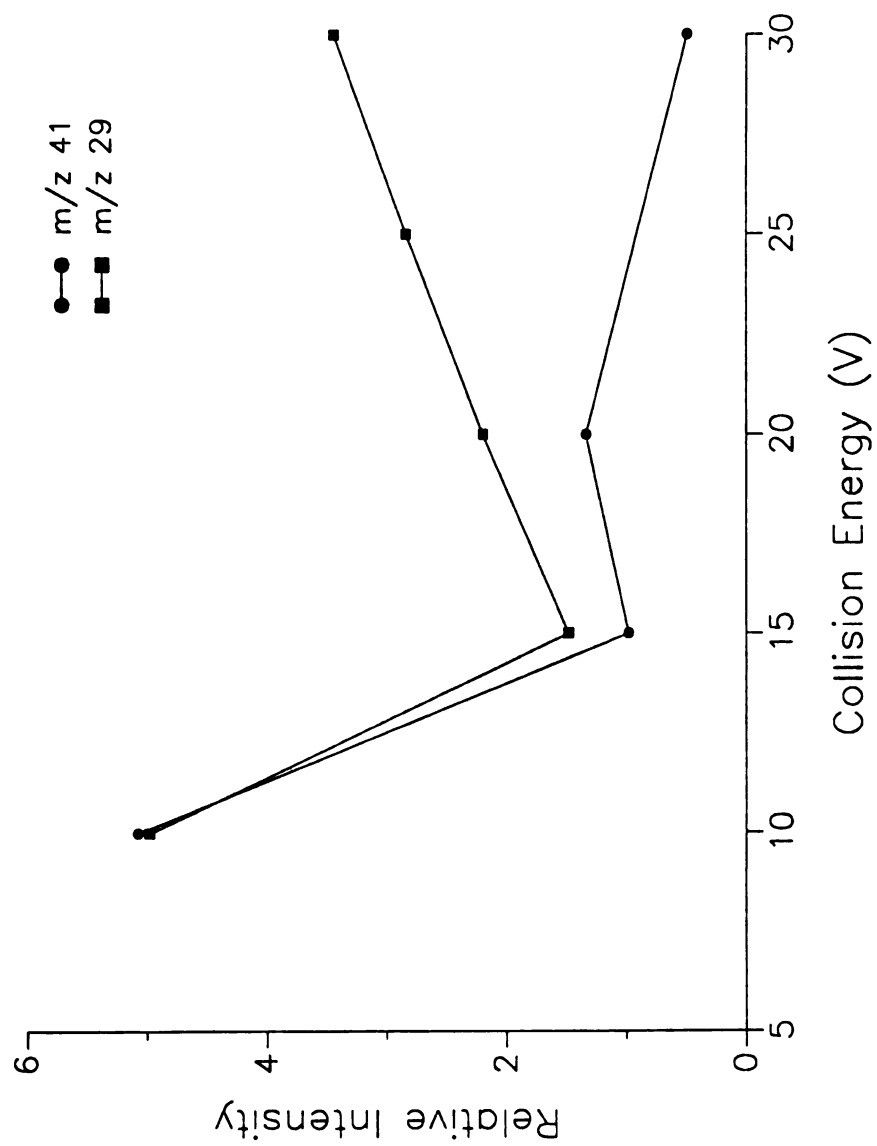


Figure 2.15 Plot of relative intensity versus collision energy for the m/z 57 ion from n-heptane.

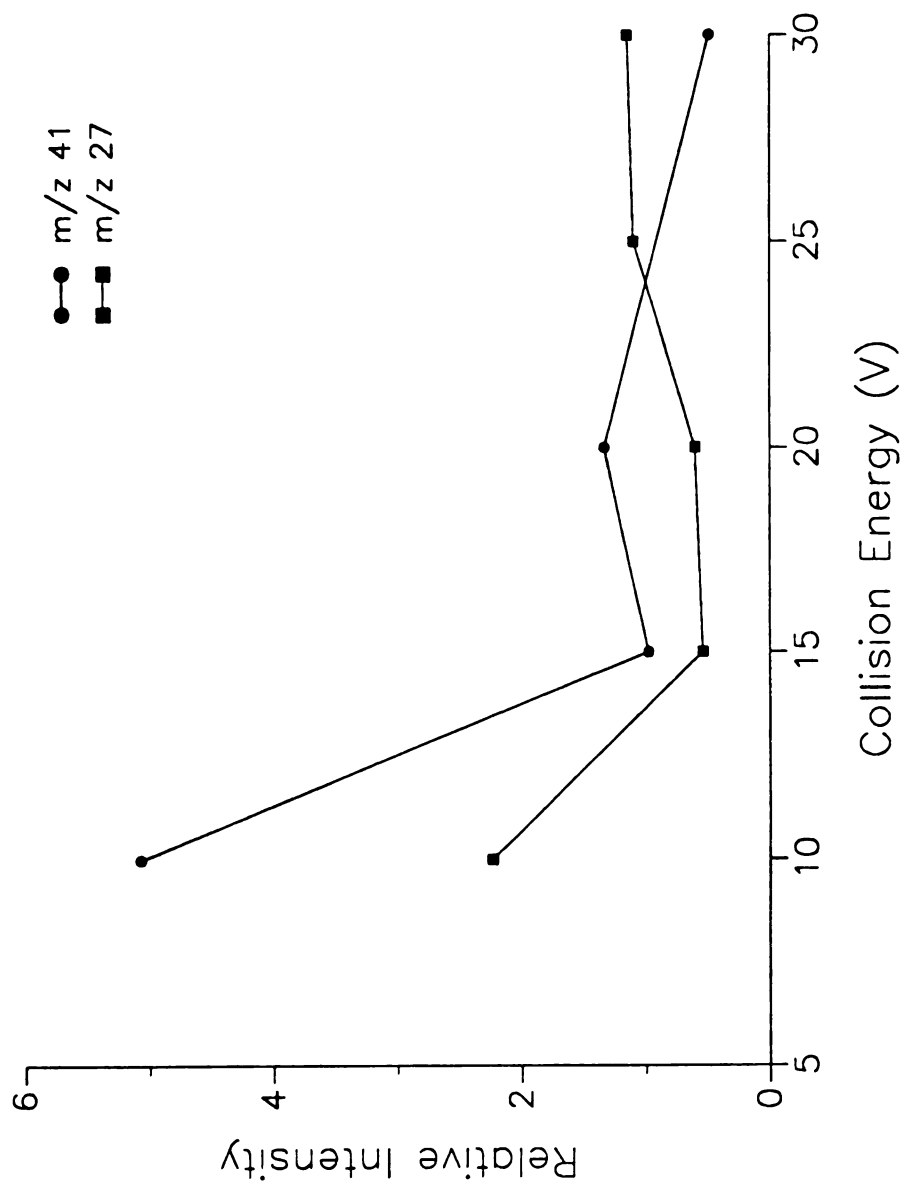


Figure 2.16 Plot of relative intensity versus collision energy for the m/z 43 ion from n-heptane.

function of collision energy. This indicates that fragmentation of the corresponding parent ions is a relatively facile processes. For the m/z 71, 57, and 43 parent ions, higher collision energies lead to lower mass daughters, some of which are not shown in these plots.

Criteria for Collecting a Database of MS/MS Spectra

Standard Conditions. As demonstrated in the previous section, collision gas pressure has a pronounced effect on daughter spectra. The optimum pressure depends on the type of information desired. At high collision gas pressures, multiple collision processes produce first, second, and higher order daughter ions. Although daughter spectra taken at high collision gas pressures provide more data, not all of these data are directly characteristic of the parent ion. Metastable decompositions and first order daughters predominate at lower collision gas pressures and the fragmentation efficiencies are lower.

The development of a database of CAD spectra requires consistent fragmentation. It is perhaps easiest to achieve this consistency in a relatively low collision gas pressure where all daughter ions produced result from single collision processes. This results in daughter ions and neutral losses which are directly characteristic of the parent ion and not some portion of it. This also increases the information content even though the data content is less. First order fragmentation products seem to provide the most useful information for substructure characterization. Higher order products may be misleading when single event neutral loss information is desired unless the reaction order of each daughter ion is

known. For the simple parent ions I have studied, first order collisions were maintained at collision gas pressures of 1 mtorr or less. This pressure produced consistent fragmentation for identical parent ions from different compounds. In this pressure regime, it was apparent from my results that the pressure could vary somewhat without affecting the set of daughters produced. The collision gas pressure required to ensure such first order collisions can be determined from brief kinetic studies as shown above.

Collision energy can be used as an extra dimension for daughter spectra. From collision energy versus intensity plots, appearance potentials of daughter ions may be deduced. However, each ion in a daughter spectrum may have a different appearance potential. Dawson mentioned that more than one collision energy may be required to adequately cover all types of parent ions. For highly conjugated molecules, such as biphenyls and polyaromatic compounds, collision energies of 50 V or higher are often required to produce fragmentation (30). For the simple, relatively small parent ions I have studied, collision energies of 20 to 25 V were sufficient for production of daughter ions of reasonable intensity.

Quad 3 drawout, which is the potential difference between quads 2 and 3, also affects the set of daughters produced. Because fragment ions have a distribution of translational energies, the quad 3 drawout potential determines whether or not daughter ions can enter quad 3. As this potential is increased, the daughter ion collection efficiency increases. However, peak shape suffers at large drawout potentials. Thus, a drawout potential of -10 volts was found to be optimal.

Lens voltages affect peak shape and intensities, but did not affect the set of daughters produced unless the voltage was changed to the point where all ions were effectively "stopped". These lens voltages and all other parameters were set to optimize sensitivity and not rigorously standardized, as they do not have as a profound effect on MS/MS spectra.

It must be emphasized that these standard conditions do not guarantee *instrument-independent CAD spectra*. However, I was able to achieve consistent fragmentation of specific parent ions from different compounds using these standard conditions. Shown in Figures 2.17, 2.18, and 2.19 are daughter spectra of the phthalate anhydride, phenylethyl, and benzoyl ions from several different compounds, respectively. It is evident from these figures that identical parent ions from different compounds can be fragmented to produce a consistent set of daughters through careful control of collision gas pressure and energy. More importantly, substructures of the same nominal mass (e.g., the phenylethyl and benzoyl substructures) may be unambiguously identified through the use of such data.

I have found that it is not possible to obtain consistent intensity data from parent ions without rigorous calibration and standardization of all the instrumental parameters. However, consistent intensity data were not required for the purposes of this research, and intensities can fluctuate by relatively large amounts without affecting the results obtained from the MAPS software, which is used to identify the presence and absence of substructures in unknowns. The reasons for this will be described more fully in Chapter 4 which discusses the use of intensity

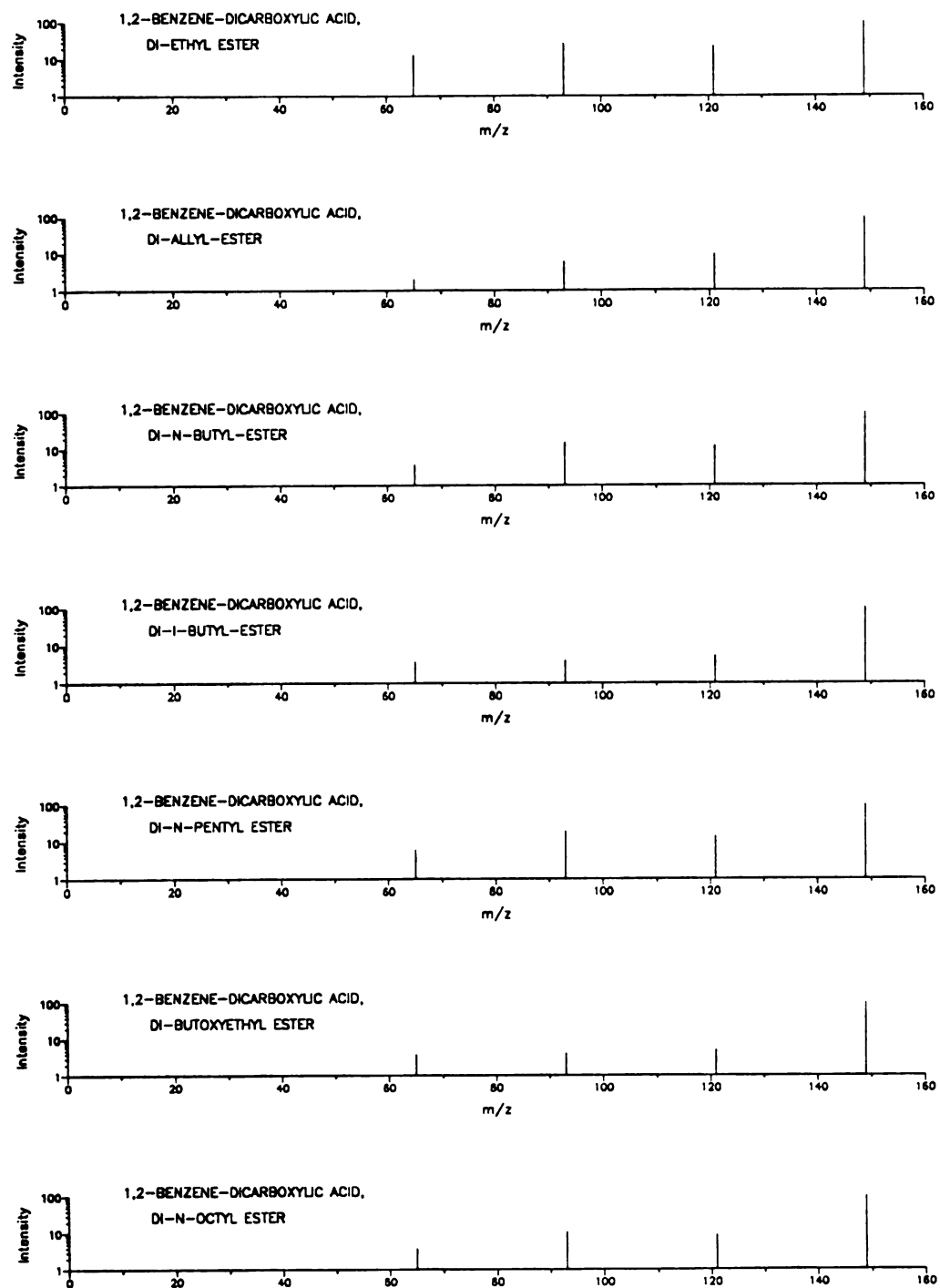


Figure 2.17 Daughter spectra of the phthalate anhydride ion (m/z 149) from several different compounds.

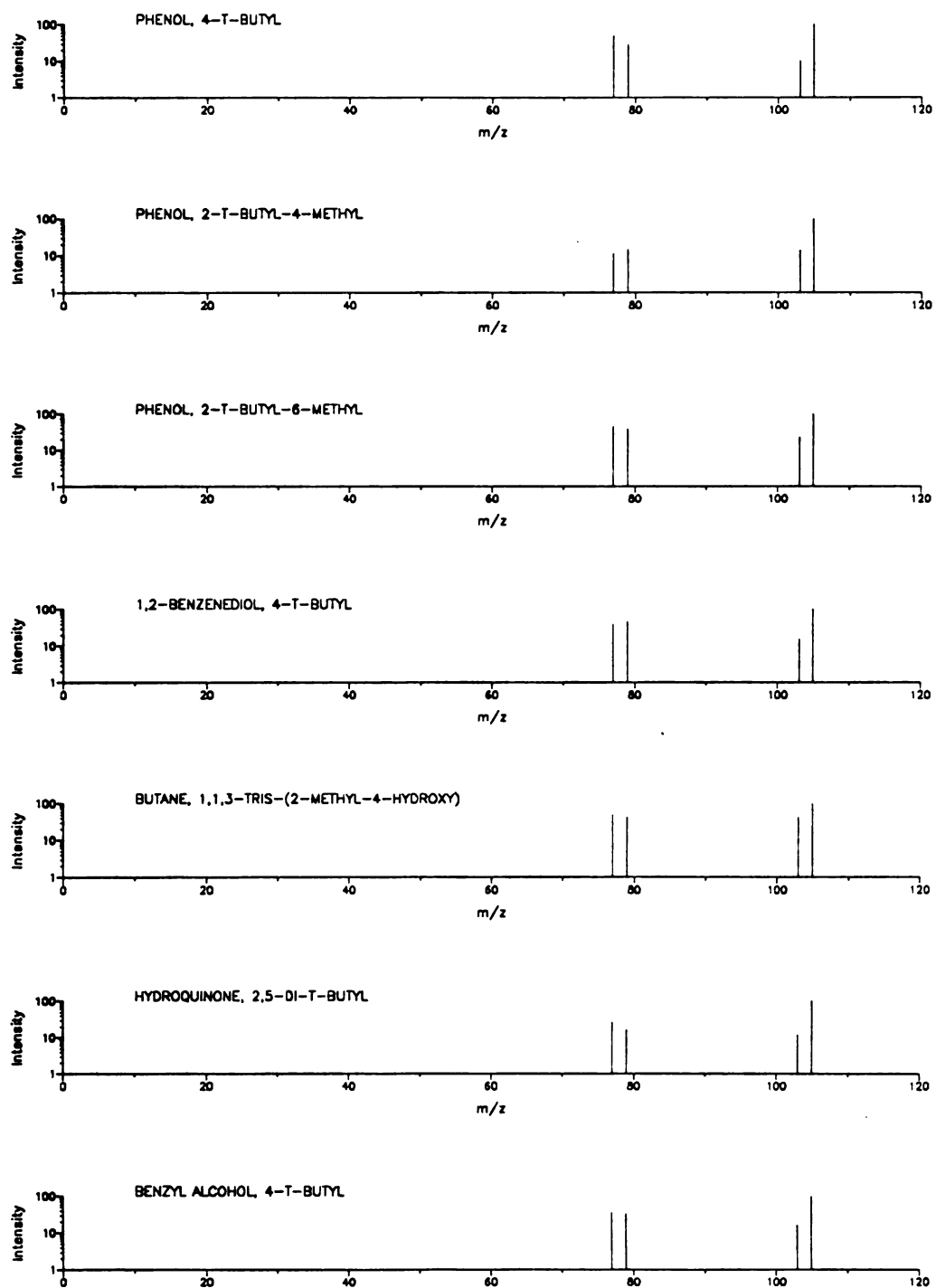


Figure 2.18 Daughter spectra of the phenylethyl ion (m/z 105) from several different compounds.

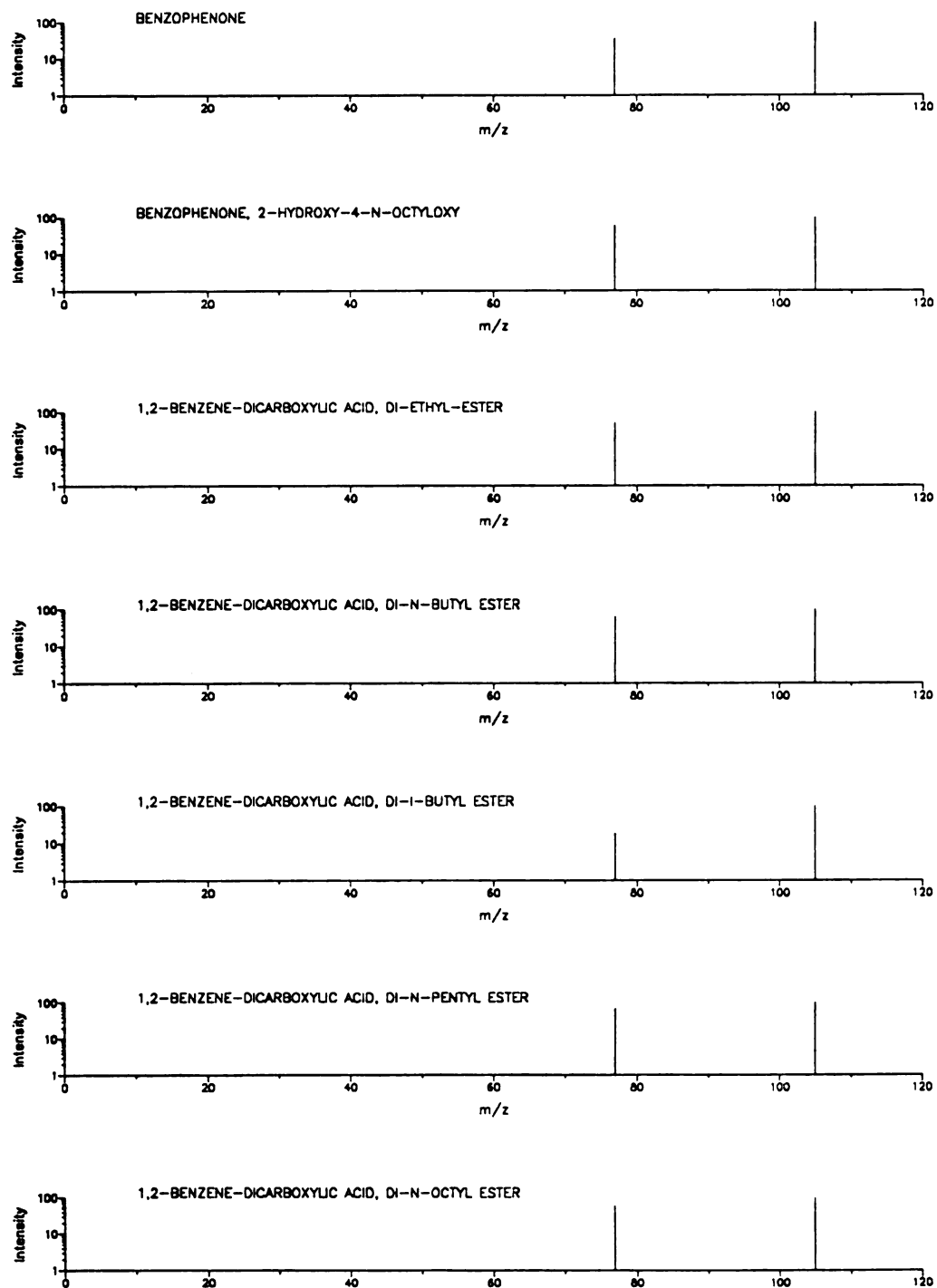


Figure 2.19 Daughter spectra of the benzoyl ion (m/z 105) from several different compounds.

data in MAPS. More research is required to identify the criteria for obtaining daughter spectra which are independent of instrument type or specific instrumental operating conditions.

Data Collection. It takes between two and five hours to collect MS/MS fragmentation maps for one compound using the Extrel TQMS instrument. This data collection process consists of several steps:

- 1) tuning the instrument with perfluorotributylamine to achieve intensities consistent with those in the reference mass spectrum of this compound,
- 2) inserting the sample into the instrument (via either a temperature-controlled solids probe or a liquid inlet),
- 3) taking a conventional mass spectrum of the compound,
- 4) checking the purity of the compound by comparing its mass spectrum to a reference spectrum if available,
- 5) tuning the instrument and optimizing its sensitivity for daughter spectra,
- 6) bringing the collision chamber up to the desired pressure,
- 7) setting all other parameters to their standard values, and
- 8) collecting three daughter spectra for each ion in the conventional mass spectrum of the compound, sensitivity permitting.

A complete MS/MS map, consisting of a daughter spectrum for every ion appearing in the conventional mass spectrum of the compound,

cannot be obtained due to the sensitivity limitations of this instrument. Daughter spectra can typically be obtained for ions whose relative intensity is greater than 1% of the base peak in a conventional mass spectrum. Three daughter spectra are collected for each parent ion. These daughter spectra may not be identical, due to the non-constant sample pressure in the source and the low sensitivity of the instrument. If one daughter spectrum is significantly different from the others, it is discarded. The remaining daughter spectra for the same parent ion are averaged before incorporation into the database. Manipulation of the data is accomplished by the MESSAGE program (31). This program allows the user to normalize spectra, average spectra, eliminate inaccuracies in the data, and put the data into proper format for inclusion in the database. This task is performed by the user using the MESSAGE program to ensure removal of inconsistencies from the data and allow manual inspection of the data before incorporation into the database.

Transfer of MS and MS/MS Data into the MAPS Database

The Multidimensional Database. Data created on the Extrel TQMS are uploaded from the FORTH-based instrument control computer (32,33) to a PDP-11 computer. This computer functions as a computing environment where mass spectral data can be manipulated and archived. Once MS and MS/MS data from an experiment are transferred to the PDP-11, they are immediately put into multidimensional database (MDDB) format. The MDDB format and associated software were

designed by Hugh Gregg and coworkers (34,35). Each MDDDB contains experimental data along with all of the associated experimental parameters.

A MDDDB, which can also be referred to as an experimenter's database, consists of three separate files: a header file, a pointer file, and a data file. The header file is used to store comments, the types and values of static instrumental parameters, and the types of variable parameters. The pointer data file stores starting record numbers for each scan in the experiment. The data file then contains X-Y data pairs and the values of the variable parameters for each scan. This database structure provides for efficient and rapid storage and retrieval of data. Once experimental data are in a MDDDB, it can then be inspected, massaged, plotted, and otherwise operated on within the PDP-11 environment by a variety of software tools developed in-house for these purposes.

The MAPS Database. The MAPS software runs on the Xerox 1108 AI workstation (36). The InterLISP-D environment on the Xerox 1108 is naturally built around list structures. Data for each compound are stored in a record data structure. This structure includes data fields for the registry name, compound name, molecular weight, MS and MS/MS data, empirical formula, and the substructures present in the compound. An abbreviated record for one such compound in the MAPS database is shown in Table 2.4. The compound name is stored in a list. This list allows for storage of more than one compound name, since several names may exist for the same compound (e.g., IUPAC name,

Table 2.4 Record for registry name CS520 in the MAPS database.

```

(CS520
(COMPOUND CS520
("1,2-BENZENEDICARBOXYLIC ACID, DI-N-PENTYL
  ESTER")
  306
  (( ( 306 "mol.ion" (      2.3293      10.9538))
    ( 237  69 100.0000)
    ( 220  86  24.8510)
    ( 219  87  61.5499)
    ( 176 130  11.6617)
    ( 174 132  26.0805)
    ( 149 157  86.2146))
  (( 238  68  2.4603))
    .      .      .
    .      .      .
    .      .      .
  (( 149 157 ( 100.0000 100.0000))
    ( 121  28  14.6733)
    (  93  56  19.1081)
    (  65  84  6.1040))
  (( 148 158  7.9196))
    .      .      .
    .      .      .
    .      .      .
  ((  44 262  4.4402))
  ((  43 263 31.0598))
  ((  41 265 14.9731)))
(C 18 H 26 O 4)
("PHTHALATE"
"PHTHALATE-ESTER"
"CARBOXYL"
"BENZOYL"
"ESTER"
"XPHENYL"
"METHYL"
"ETHYL"
"1-2-PHENYL"
"CARBONYL"
"PROPYL"
"BUTYL"
"PENTYL")) )

```

common name, trade name). Each data point is stored as a "node". For example, the node for m/z 41 in the conventional mass spectrum is ((41 265 14.9731)) and the node for the m/z 121 daughter from m/z 149 is (121 28 14.6733). Each node consists of the m/z value, the neutral loss which led to that ion's formation, and the intensity. The m/z values are rounded off to integers. Intensities are relative to the base peak in the spectrum and are real numbers. The list of substructures contained in the compound is obtained using the STRCHK routine in GENOA as described in Chapter 1.

Each compound in the database is identified by a unique alphanumeric registry name. These registry names have been defined somewhat arbitrarily. Most of the compounds in the database were obtained from a Chem Service compound kit (37). The registry names for these compounds consist of the initials CS followed by the Chem Service compound number. A set of phenolic compounds, obtained from Keith Olsen at General Motors Research Labs (38), were also included in the database. The registry names for these compounds consist of the initials GMR followed by the compound number. Other compounds in the training set are identified by unique registry names.

The records for all the compounds in the MAPS database are stored in a property list, allowing fast and facile access to the several fields of data in the records for individual compounds.

Requirements for Automatic Entry of Data into the MAPS Database. It would be extremely awkward to operate on the binary, unformatted, direct access files which make up a MDDB in the InterLISP-D

environment on the Xerox 1108. A MDDB stores all the experimental data for a compound, including instrumental parameters. The MAPS system does not require all of the information within a MDDB, but only the MS/MS fragmentation map. Also, the MAPS database requires additional information which is not included within a MDDB, such as the name, molecular weight, empirical formula, and the substructures present for each compound. This information must be obtained from the user.

The databases for MS and MS/MS data on the PDP-11 (MDDB) and the Xerox 1108 (the MAPS database) use different formats, due to their different contents and the requirements placed on efficient data storage and retrieval on these widely different computing environments. Therefore, to achieve the goal of automatic transfer of MS and MS/MS data from a MDDB on the PDP-11 to the MAPS database on the Xerox 1108, three requirements were:

- 1) a link between the two computers,
- 2) software on the PDP-11 to extract and verify the necessary data from an MDDB, obtain additional data from the user, and incorporate these data into a file that has a format compatible with the MAPS database, and
- 3) software on the Xerox 1108 to access the desired file on the PDP-11, transfer it over the link, verify the data, and store it in the MAPS database.

Linking the PDP-11 and the Xerox 1108. The InterLISP-D environment supports file transfers via the KERMIT protocol and includes software for this purpose. Since KERMIT software was already installed on our PDP-11, this protocol was the logical choice for transferring files between the two computers. A serial line was constructed to connect the two computers and the appropriate software was installed on the Xerox 1108.

Front End for Transfer of Data from the PDP-11 to the Xerox 1108. A program (DBXFER) was developed to extract MS/MS data from a MDDB, verify it, accept the additional data required by the MAPS database, and create an ASCII file containing these data. This file is then transferred to the Xerox 1108, where it is incorporated into the MAPS database.

Once the data from an experiment has been incorporated into a MDDB and massaged into the desired format, the DBXFER program is invoked. In addition to accomplishing the conversion of MS/MS data from MDDB to MAPS format, this program also verifies the data and checks for inconsistencies and inaccuracies. The following checks are performed.

- 1) The MDDB must include a conventional mass spectrum.
- 2) The remaining scans must be daughter spectra.
- 3) All spectra must be normalized to the base peak.
- 4) All daughter spectra must be unique (no duplicate daughter spectra of the same parent mass).
- 5) All m/z values in each spectra must be unique (no duplicate integral m/z values in a spectra).
- 6) There can be no peaks in the conventional mass spectrum at m/z values greater than the molecular weight plus 8 u.
- 7) There can be no peaks in daughter spectra at m/z values greater than the parent mass.

These last two tests screen against contamination peaks, adduct ions, ion-molecule reaction products, and noise peaks. Based on the standard conditions under which the data were collected, these situations should not occur and are undesirable.

DBXFER makes one pass through the data in a MDDB and performs the above tests. If any of these tests fail, the program prints an appropriate error message to the user and aborts without creating the MAPS compatible data file. At this point, the user must use the MESSAGE program to rectify any errors detected in the data, or collect new data if necessary. Once the data in a MDDB have been verified, DBXFER prompts the user for several types of information required by the MAPS database. These include the registry name of the compound, compound name, molecular weight, empirical formula, and the substructures present. After these data have been input, an ASCII file is

created which contains all the information required for this compound for the MAPS database.

Back End for Transfer of Data from the PDP-11 to the Xerox 1108. The ASCII file created by the DBXFER program is transferred over the serial line to the Xerox 1108 using the KERMIT software. Several functions were written to read in the file, verify the contents, notify the user of any errors if detected, create a new record for the compound, and incorporate it into the MAPS database.

Automatic Entry of Data into the MAPS Database. Prior to creation of the link between the PDP-11 and the Xerox 1108, entry of data into the MAPS database was performed manually. This proved to be tedious and prone to errors. The software and the link described above allow for automatic entry of data into the MAPS database. This scheme was tested and found to be an efficient and reliable. It has the additional advantage of incorporating checks and verifications of the data on both ends to eliminate errors which would otherwise appear in the database.

The Current MAPS Database

Shelley stated that "Most academic or industrial researchers will not be able to justify the expense of obtaining the spectra of pure, known compounds directly from instruments for the sole purpose of generating a reference database" (19). Given the amount of time I have invested into the development of a relatively small database of CAD spectra, I certainly

support this statement. The database I have developed represents MS and MS/MS data for 76 different compounds, comprising a variety of small, relatively common substructures. It includes 76 conventional mass spectra and 1114 daughter spectra, with an average of 15 daughter spectra per compound. This database characterizes several substructures quite adequately, most notably alkyl and aromatic substructures. However, it lacks structural diversity. The database does incorporate redundancy as it contains daughter spectra of the same parent ion from several different compounds. The quality of these spectra can certainly be analyzed using a modified form of McLafferty's quality index to identify dubious spectra or to choose the best spectrum from a set of duplicate spectra. This has not been attempted yet due to the small size of the database. One attractive feature about the use of the MDDB for storage of data is that it allows access to *all the instrumental parameters* for each daughter spectrum in the database. Thus, the instrumental parameters associated with each spectrum in the database may be evaluated through inspection of the appropriate MDDB. The connectivity tables for each structure in the database and defined substructures, created using GENOA structure entry routines, are stored in libraries on the MicroVAX. GENOA's substructure searching capabilities are used to obtain substructure information from the structures of each compound in the database. This information is used for making spectral feature/substructure correlations. While the database is still relatively small, it allowed reliable spectral feature/substructure correlations to be made, as will be seen in Chapter 3.

References

1. Stenhagen, E., Abrahamsson, S., McLafferty, F.W., "Registry of Mass Spectral Data", Wiley & Sons, New York, 1974.
2. Wiley & Sons, Electronic Publishing Division, 605 Third Ave, New York, 10158.
3. Heller, S.R., Milne, G.W.A., "EPA/NIH Mass Spectral Database", U.S. Government Printing Office, Washington, DC, 1978.
4. U.S. National Bureau of Standards, Office of Standard Reference Data, Physics Building, Room A-320, Gaithersburg, MD, 20899.
5. Milne, G.W.A., Potenzzone Jr., R., Heller, S.R., Science, **215**, 371 (1982).
6. Shelley, C.A., in Zupan, J. (Ed.), "Computer-Supported Spectroscopic Databases", John Wiley & Sons, New York, 1986, p. 6.
7. Henneberg, D., Adv. Mass Spectrom., **8**, 1511 (1980).
8. Speck, D.D., Venkataraghavan, R., McLafferty, F.W., Org. Mass Spectrom., **13**, 209 (1978).
9. Milne, G.W.A., Budde, W.L., Heller, S.R., Martinsen, D.P., Oldham, R.G., Org. Mass Spectrom., **17**, 547 (1982).
10. Eichelberger, J.W., Harris, L.E., Budde, W.L., Anal. Chem., **47**, 995 (1975).
11. Milne, G.W.A., Heller, S.R., J. Chem. Inf. Comput. Sci., **20**, 204 (1980).
12. Heller, S.R., in Zupan, J. (Ed.), "Computer-Supported Spectroscopic Databases", John Wiley & Sons, New York, 1986, p. 118.
13. Dillard, J.G., Heller, S.R., McLafferty, F.W., Milne, G.W.A., Org. Mass Spectrom., **16**, 48 (1981).
14. Gray, N.A.B., "Computer-Assisted Structure Elucidation", Wiley & Sons, New York, 1986.
15. Smith, E.G., "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill, New York, 1968.

16. Abe, H., Fujiwara, I., Nishimura, T., Okuyama, T., Kida, T. Sasaki, S., J. Chem. Inf. Comput. Sci., **24**, 212 (1984).
17. Wipke, W.T., Dyott, T.M., J. Am. Chem. Soc., **96**, 4825 (1974).
18. McLafferty, F.W., Stauffer, D.B., Int. J. Mass Spectrom. Ion Proc., **58**, 139 (1984).
19. Bozorgzadeh, M.H., Morgan, R.P., Beynon, J.H., Analyst, **103**, 1613 (1978).
20. Davidson, W.R., Fulford, J.E., 31st Annual Conference on Mass Spectrometry and Allied Topics, 1983, p. 559.
21. McLafferty, F.W., Hirota, A., Barbalas, M.P., Org. Mass Spectrom., **15**, 547 (1980).
22. Giordani, A.B., Gregg, H.R., Hoffman, P.A., Cross, K.P., Beckner, C.F., Enke, C.G., 32nd Annual Conference on Mass Spectrometry and Allied Topics, 1984, p. 648.
23. Dawson, P.H., French, J.B., Buckley, J.A., Douglas, D.J., Simmons, D., Org. Mass Spectrom., **17**, 205 (1982).
24. Martinez, R.I., Cooks, R.G., 35th Annual Conference on Mass Spectrometry and Allied Topics, 1987, p. 1175.
25. Dawson, P.H., Sun, W.F., Int. J. Mass Spectrom. Ion Proc., **55**, 155 (1983).
26. Martinez, R.I., Dheandhanoo, S., J. Res. Natl. Bur. Stand., **92**, 229 (1987).
27. Martinez, R.I., Rapid Comm. Mass Spectrom., **1**, 8 (1988).
28. Martinez, R.I., Rev. Sci. Instrum., **58**, 1702 (1987).
29. Wong, C.M., Crawford, R.W., Barton, V.C., Brand, H.R., Neufeld, K.W., Bowman, J.E., Rev. Sci. Instrum., **54**, 996 (1983).
30. Schoen, A.E., Syka, J.E.P., 34th Annual Conference on Mass Spectrometry and Allied Topics, 1986, p. 722.
31. Cross, K.C., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1986.
32. Myerholtz, C.A., Schubert, A.J., Kristo, M.J., Enke, C.G., Instruments and Computers, **6**, 11 (1985).
33. Myerholtz, C.A., Schubert, A.J., Kristo, M.J., Enke, C.G., Instruments and Computers, **6**, 13 (1985).

34. Gregg, H.R. Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1987.
35. Crawford, H.R., Brand, H.R., Wong, C.M., Gregg, H.R., Hoffman, P.A., Enke, C.G., Anal. Chem., 56, 1121 (1984).
36. Xerox Corporation, Artificial Intelligence Systems, 250 N. Halstead St., Pasadena, CA, 91109.
37. Chem Service Inc., P.O. Box 194, West Chester, PA, 19380.
38. General Motors Research Laboratories, Analytical Chemistry Department, 30500 Mound Rd., Warren, MI, 48090.

CHAPTER 3

THE MAPS SOFTWARE

Introduction

Mass spectrometry (MS) has long been recognized as a useful tool for structure elucidation. Some common fragment ions and assumed neutral losses (using the mass difference between two peaks) observed in conventional mass spectra have been recognized as fairly specific indicators for certain substructures. These have been tabulated and are widely used in spectral interpretation (1,2). MS/MS spectra of some compounds which have common substructures include *patterns* of features which are caused by those substructures. These patterns, which have been previously identified manually from daughter spectra (3), can be used to identify these substructures in unknowns. Although the literature contains many examples of the use of MS/MS to screen for specific compounds or compound classes (4-6), *until now there has been no attempt to exhaustively deduce and organize the correlations between MS/MS spectral features and substructures.* One problem is the dependence of MS/MS spectra on the experimental operating conditions and type of MS/MS instrument used; this is discussed in Chapter 2. In addition, the volume and dimensionality of data required to make correlations has been too large for many laboratory microcomputers to

handle efficiently. The time taken to acquire a library of MS/MS spectra on early instruments was prohibitive. However, greatly enhanced rates of data acquisition are now being achieved with newly developed instrumentation such as the Finnigan TSQ-70 (7) and a time-resolved ion momentum MS/MS system (8,9) using time array detection (10) currently under development in this laboratory.

MS/MS has several advantages over conventional MS for structure elucidation; these have been discussed in Chapter 1. Briefly, individual daughter spectra have fewer peaks than conventional mass spectra, and parent ions, daughter ions, and neutral losses may be separately and unequivocally observed. MS/MS not only yields more data per compound, but provides a second dimension of information which is more definitive than that provided by conventional MS. Also, MS/MS is potentially the better technique for identifying substructures, since sets of features due to particular substructures are less overlapped in the MS/MS data space than in conventional mass spectra.

Several types of features can be obtained from MS/MS data: lines in conventional mass spectra, lines in daughter spectra, neutral losses, and parent-to-daughter transitions. The most specific of these features are the parent-to-daughter transitions. Given a mass range of 1 to n u, the theoretical number of features which can be obtained from unit resolution MS/MS data can be calculated from the following equations.

$$\begin{aligned}
 \text{lines in conventional mass spectra} &= n \\
 \text{lines in daughter spectra} &= n \\
 \text{neutral losses} &= n \\
 \text{parent-to-daughter transitions} &= 0.5 \times (n^2 - n) \\
 \\
 \text{total number of potential features} &= 0.5 \times (n^2 + 5n)
 \end{aligned}$$

There are potentially 126,250 spectral features in a mass range of 0 to 500 u. For the same mass range there are only 500 directly observable MS features. As the mass range doubles, the number of potential MS/MS spectral features increases by a factor of four. With higher resolution MS/MS instruments, the number of potential features increases even further. Thus, the tremendous amount of information available from MS/MS data is apparent.

One of the goals of this research group has been to develop an effective tool for substructure identification as part of an overall system for more rapid and reliable identification of molecular structures using unit resolution MS and MS/MS data. Towards this end, we have developed an algorithm henceforth referred to as MAPS (**M**ethod for **A**nalyzing **P**atterns in **S**pectra). MAPS automatically deduces the relationships between the presence of substructures in molecules and the characteristic features they produce in MS and MS/MS spectra (11,12). The relationships found are expressed as IF-THEN rules that indicate which MS and MS/MS features have been associated with each substructure. MAPS uses chemical knowledge and elements of pattern recognition and artificial intelligence in the rule generation process. The

MAPS approach is an empirical scheme for discovering spectral feature/substructure relationships using both MS and MS/MS data. MAPS makes no assumptions about the fragmentation process or rearrangements, and correlates spectral features with molecular substructures rather than ionic structures. While MAPS currently uses MS and MS/MS data, the MAPS approach (and much of the software) is equally suited to multiple stage mass spectrometric data, from which further improved rules may be confidently expected. Advantages of searching for such spectral feature/substructure relationships by computer rather than manually include speed, completeness, lack of personal bias, and built-in automatic statistical checking.

MAPS is only one component of the ACES system described in Chapter 1. The rules developed by MAPS are of great utility for structure elucidation. These rules may be applied to MS and MS/MS data from unknowns to identify the presence of certain recognized substructures. This information may be used to develop constraints on the elemental composition of an unknown, and thus aid in the determination of its molecular formula. The substructures so identified substantially simplify the correct determination of an unknown's structure.

Choice of a Development Tool for MAPS

Finnigan MAT donated a Xerox 1108 artificial intelligence (AI) workstation (13) to our group in August 1985. This workstation came equipped with the InterLISP-D programming environment and KEE (Knowledge Engineering Environment). The Xerox 1108 and its

associated software proved to be invaluable for the development of the MAPS software. The utility of both KEE and LISP as development vehicles for the MAPS software are described in this section.

KEE. KEE is a set of software tools developed by Intellicorp for assisting in the development of knowledge-based systems (14). It includes several well-known artificial intelligence methodologies, such as object-oriented programming, frame-based knowledge representation, inheritance, forward and backward-chaining rule-based reasoning, and LISP programming. It is an excellent tool for rapid prototyping, design, debugging, and testing of expert system applications. KEE was initially used to develop a system for identifying substructures from their corresponding daughter spectra using a methodology similar to the original system for structure elucidation from MS/MS data described in Chapter 1. This involved using object-oriented paradigms provided by KEE to develop a knowledge base of substructures and their associated daughter spectra. A backward-chaining rule-based approach was then used to identify these substructures in unknowns. This approach involved developing procedures in the LISP language for matching daughter spectra. However, this system still did not use all of the information available from MS/MS data for substructure identification and had many of the drawbacks associated with the original system described in Chapter 1. In the course of developing this prototype system, KEE was found to have several drawbacks which made its use for this project unnecessarily complicated. It is memory-intensive, slow, and forces the user into a specific syntax. KEE also insulates the user

from the LISP environment and thus can complicate programming, since the user must implement data structures in KEE format and algorithms in LISP code.

LISP. The LISP programming language was developed in 1958 by John McCarthy and has become perhaps the most popular AI programming language (15-17). Several excellent reasons for this are provided by McCarthy himself : "LISP is now the second oldest programming language in present widespread use (after FORTRAN)... Its core occupies some kind of local optimum in the space of programming languages given that static friction discourages purely notational changes. Recursive use of conditional expressions, representation of symbolic information externally by lists and internally by list structure, and representation of program in the same way will probably have a very long life" (15). LISP is a symbolic language and differs substantially from the more conventional programming languages. It uses a single primitive data type: the list. Hence the name of the language, which is an acronym for "**LISt Processing**", is well-deserved. LISP also differs in that its programs are described not as a sequence of steps, but rather as functions which are defined in a somewhat mathematical format. Each function call is in list format: the first element represents the function name and the remaining elements denote the arguments. The arguments may themselves be function calls; thus the language is extensible. The LISP language also implements recursion (a recursive function is one which can call itself), which is a very powerful programming technique. LISP code can be interpreted or compiled.

The InterLISP-D system on the Xerox 1108 is one of the most advanced LISP programming environments (18,19). The name InterLISP stands for "**Interactive LISP**". It has several powerful features which include:

- 1) a structure editor;
- 2) the Programmer's Assistant, which remembers user commands and allows the user to select old commands for correction, re-execution, or undoing;
- 3) Masterscope, which cross references function calls and variable usage, and is very useful for managing large applications;
- 4) the File Package system, which allows the user to identify and save functions and variables which have been modified or created during the course of a programming session;
- 5) a graphics system which allows for multiple windows;
- 6) icons and mouse-driven menus; and
- 7) an excellent error handling system which includes spelling correction and a flexible and very powerful debugger.

LISP does have several disadvantages. All of the available storage space on the system may eventually get used up during the course of a programming session or the execution of a program. LISP systems implement "garbage collection" to reclaim used storage space. During garbage collection, execution is suspended. Since LISP uses dynamic allocation of storage space, garbage collection is a necessity and

represents part of the price paid for the flexibility provided by the language. LISP systems are usually memory intensive. The InterLISP-D system consumes *several megabytes of memory*; all of the utility functions and useful programming tools have their price. LISP is also conceived as slow, especially for mathematical calculations. Better compilers and faster microprocessors specifically built for LISP have substantially improved its execution speed. A serious problem facing LISP is the lack of a language standard. Unfortunately, almost all of the applications of LISP have been in the AI community. With the current popularity of artificial intelligence techniques and expert systems, there has been tremendous interest in LISP over recent years. In response to this, the industry appears to be moving towards Common LISP as a standard (20). This will hopefully encourage a wider acceptance of LISP language.

We decided to use the LISP language for development of the MAPS software. LISP is well suited to this task since the natural data structures of this problem are lists of alphanumeric data. The InterLISP-D system on the Xerox 1108 was found to be an excellent programming environment for this work. Development of MAPS in LISP allowed more compact code, faster execution, and a simpler, more direct implementation than in the KEE system. The remainder of this chapter describes the initial version of the MAPS code as developed by Dr. Adrian Wade. The substructure identification rules and the results from rule application given in this chapter were obtained using a training set developed from the MS and MS/MS spectra I have collected.

Generation of Substructure Identification Rules Via MAPS

The rule generation process takes place in four stages as illustrated in Figure 3.1. First, the training set is constructed from substructural, MS and MS/MS data from known compounds. Next, the "feature bucket" and "substructure bucket" lists are created to facilitate the third step, which is correlation of features with substructures. Finally, filters are applied to remove false correlations. This process results in rules which express the correlation of MS and MS/MS spectral features with particular fragments of molecules. Each of these steps are described in detail below.

Construction of a Training Set. A training set comprises substructural information and MS and MS/MS data for several known compounds. Data for each compound are stored in record data structures which contain separate fields for storing the alphanumeric identification code, full name, molecular weight, MS and MS/MS spectra, empirical formula, and a list of recognized substructures present. This list is obtained using the GENOA substructure searching routine, which involves searching a given molecular structure for recognized structures. To accomplish this, GENOA requires a library of known substructures (i.e., benzyl, nonyl, carboxyl) in which the names and connectivity tables for each substructure are specified. MAPS maintains its own abbreviated version of the GENOA substructure library. This is in the form of a list which contains the name, lowest and highest possible fragment mass or neutral loss and the empirical formula of each recognized substructure.

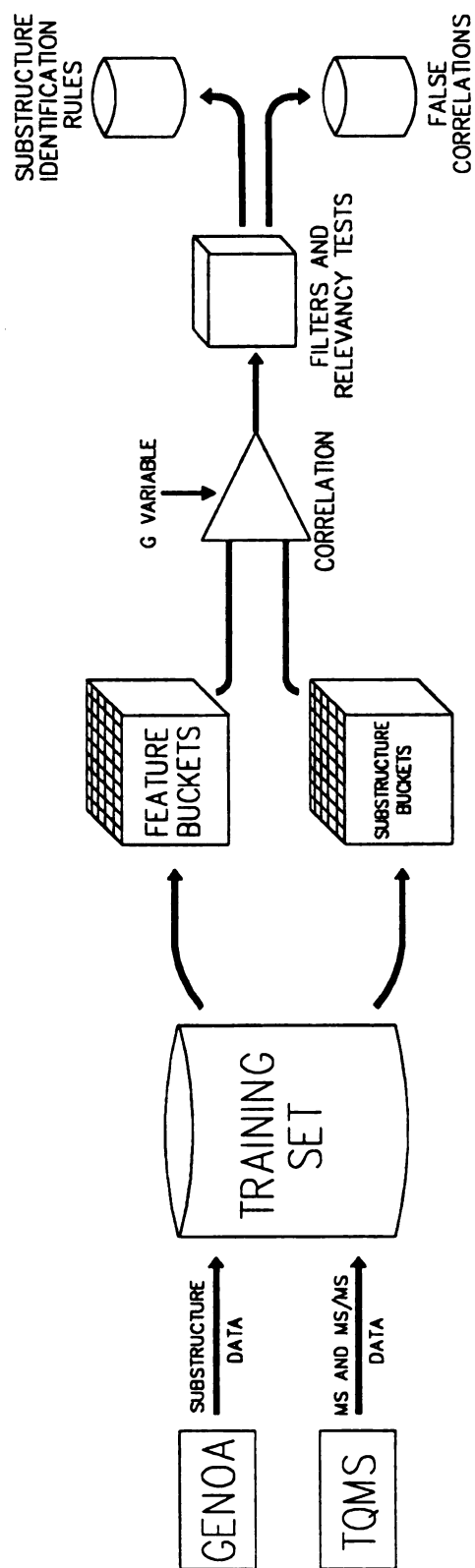


Figure 3.1 Schematic of rule generation process.

As examples, entries for the benzyl and carbonyl substructures appear as the lists (BENZYL 1 77 (C 7 H 7)) and (CARBONYL 12 28 (C 1 O 1)).

The MS and MS/MS spectra from a compound may be displayed in a graphical format as shown in Figure 3.2 using an additional MAPS function. This fragmentation tree concisely indicates the parentage of each ion. Several training sets obtained under different sets of experimental conditions (e.g., EI, CI, different collision gas pressures) may be stored separately on the Xerox 1108.

Construction of Feature and Substructure Bucket Lists. To facilitate the correlation of spectral features with recognized substructures, two list data structures, referred to as the "feature bucket list" and the "substructure bucket list" must be (re)generated. The term "bucket list" arises because each association can be thought of as a bucket: the head of the list (the first item) is like a label on a bucket; the tail of the list (that is, everything except the head) is the contents of that bucket.

Creation of the feature bucket list requires the extraction of recognized features from the MS and MS/MS spectra. Currently the features recognized are the m/z values seen in conventional and daughter scans, neutral losses, and parent-to-daughter transitions. New types of features (e.g., multiplicity of daughters from parents and daughter spectra intensity values) may readily be added. The format of the feature bucket list is

Figure 3.2 Abbreviated fragmentation tree for 1,2-benzene-dicarboxylic acid, diethyl ester (CS525).

```
((<feature-a> <compd-a1> <compd-a2> ... <compd-an>)
 (<feature-b> <compd-b1> <compd-b2> ... <compd-bn>)
 (    ...           ...           ...           ...           ...   ))
```

where *<feature-a>* might be "daughter ion at m/z 29" and *<compd-a₁>...<compd-a_n>* are the identifiers for the compounds which had this feature in their MS/MS data space.

The substructure bucket list associates the individual recognized substructures with the compounds in which they are known to occur. New substructures may readily be added. The list has the form

```
((<substr-i> <compd-i1> <compd-i2> ... <compd-in>)
 (<substr-j> <compd-j1> <compd-j2> ... <compd-jn>)
 (    ...           ...           ...           ...           ...   )).
```

Correlation of Features with Substructures. The first step in this stage of the rule generation is to filter the substructure and feature bucket lists to remove any bucket containing less than three compounds, since these represent insufficient instances of a substructure or feature to make reliable correlations.

The resulting feature buckets and substructure buckets are correlated against each other to identify features which correlate with the remaining substructures. This results in a third list structure which forms the basis for the interpretation rules and is the first step towards the deduction of reliable spectral feature/substructure correlations. This list has the form

```

(((<substr i> s-occurrences-i (<feature i1> f-occurrences-i1)
      (      ...                ...                )
      (<feature in>                ...                ))
(<substr j> s-occurrences-j (<feature j1>                ...                )
      (      ...                ...                ))
(<substr m> s-occurrences-m (      ...                ...                )))

```

where s-occurrences is the number of occurrences of compounds in the training set which contain that substructure and f-occurrences is the number of compounds which had that substructure present AND had that feature in their MS/MS spectra. One measure of the level of correlation between $\langle \text{feature } i_n \rangle$ and $\langle \text{substr } i \rangle$ is the fraction given by $(f\text{-occurrences-}i_n / s\text{-occurrences-}i)$. This value must ordinarily be not lower than some minimum level of correlation required for a feature to be considered relevant. This level is set by the C variable (correlation factor for rule generation) and is currently 75%. This implies that for to be included in a rule, it must occur in at least three out of four compounds that have that substructure.

Some special features correlate *specifically* with a particular substructure and few others, but may not occur *whenever* that substructure is present in a compound. Their absence may be due to the variety of fragmentation pathways that the molecule can follow. A simple example of this type of feature is a loss of 15 u which is often (but not always) seen when a methyl substructure is present in a compound. Such features can be of value in identifying the presence of a

substructure even though they may have too low a level of correlation to *ordinarily* be included in the rule. In this version of MAPS, such unique features are included in a rule if their level of correlation with the substructure is at least one half of the C value and their uniqueness (defined as the percentage of compounds with the *feature and the substructure* out of the total number of compounds in the training set with that feature) is greater than one half of the C value.

Filtering the Rules to Retain only Relevant Features. Features obtained through the process described above are then filtered to remove false correlations. These filters currently include low and high mass limits for each substructure, and constraints to define legal fragment masses based on their elemental composition. The low and high mass limits correspond to the minimum and maximum fragment masses or neutral loss that can be attributed to each substructure. These may be defined when the substructure is added to the library. One or two hydrogen shifts are allowed onto the fragment, so long as sufficient free valences are available and enough hydrogen atoms are accessible from other portions of the molecule.

The features in each rule are scrutinized to ascertain whether the fragment mass(es) or neutral loss(es) can be attributed to that substructure, based on combinations of the atoms present in the substructure and the rules of valence; any that cannot are removed. For example, a daughter ion of 18 u cannot be attributed to the chloromethyl substructure, as no feasible combination of carbon, chlorine and hydrogen atoms satisfy this mass. While a *daughter ion* of 18 u cannot

occur from a t-butyl substructure, a double neutral loss of CH_4 and H_2 from this substructure would satisfy this mass loss and has been observed. Determination of what is and is not a legal fragment can be complex. For example, the 60 u fragment C_5 may be possible based on valence criteria and elemental composition, but is highly unlikely to be formed under the CAD conditions used here. It is more likely that this 60 u fragment is due to $\text{C}_3\text{H}_8\text{O}$ or $\text{C}_2\text{H}_6\text{NO}$. When a rule is listed, the formula(e) consistent with each fragment mass or neutral loss is listed along with each feature. Where several formulae may be attributed to a particular fragment or neutral loss in a rule, MAPS makes no attempt to determine which is most likely.

False correlations due to artifacts in the training set can also occur. For example, features that correlate highly with the t-butyl substructure might be due to the break-up of the phenyl moiety if most of the t-butyl containing compounds in the training set are phenyl compounds with a t-butyl substituent. Such artifacts are expected to decrease as the variety of compounds in the database grows.

The surviving sets of features may then be printed as rules and have the general form:

```
IF    [x1/y] (feature a)    (possible fragment ion formula(e))
AND   [x2/y] (feature b)    ( .. .. .. )
AND   [xn/y] ..            ( .. .. .. )
THEN (likely substructure)
```


Each spectral feature in each rule has some level of correlation with the substructure. This is expressed as an [x/y] fraction, which indicates that x of the y training set compounds that had that substructure also had that feature in their MS/MS spectra. MAPS postulates candidate formulae for the fragment masses in each rule based on the elemental composition of the substructure and valence rules. Intensity values were not used in this version of MAPS and were incorporated later. Generating the rules and evaluating their reliability using the current training set of 76 compounds currently takes about four hours on the Xerox 1108.

Table 3.1 shows the ethyl rule at three different G values. Features with levels of correlation below the G value are those which are infrequent but highly specific to the ethyl substructure. As the G value decreases, more features correlate with the substructure, and the length of the rule increases. Rules may be generated at several different G values and selectively used to achieve the desired performance level for identifying substructures.

Evaluation of Unknowns via MAPS

"Expert" investigation of the substructures present in unknowns may be obtained by matching the substructure rules against their MS and MS/MS spectra. The minimum percentage of features from a rule that must be present in an unknown's MS and MS spectra for that substructure to be identified is called the "match value". The AND's shown in the rules are applied rigorously only at match values of 100%,

Table 3.1 Rule composition for the ethyl substructure at three different C values.

C = 100%	IF [27/27] neutral loss of 28 amu	C ₂ H ₄
	THEN functionality indicated is ETHYL	
C = 75%	IF [21/27] neutral loss of 16 amu	CH ₄
	AND [22/27] neutral loss of 26 amu	C ₂ H ₂
	AND [12/27] neutral loss of 27 amu	C ₂ H ₃
	AND [27/27] neutral loss of 28 amu	C ₂ H ₄
	AND [23/27] daughter ion at m/z 29	C ₂ H ₅
	THEN functionality indicated is ETHYL	
C = 50%	IF [20/27] neutral loss of 2 amu	H ₂
	AND [7/27] neutral loss of 4 amu	H ₂ + H ₂
	AND [9/27] neutral loss of 14 amu	CH ₂
	AND [19/27] neutral loss of 15 amu	CH ₃
	AND [21/27] neutral loss of 16 amu	CH ₄
	AND [18/27] neutral loss of 18 amu	CH ₄ + H ₂
	AND [22/27] neutral loss of 26 amu	C ₂ H ₂
	AND [12/27] neutral loss of 27 amu	C ₂ H ₃
	AND [27/27] neutral loss of 28 amu	C ₂ H ₄
	AND [12/27] neutral loss of 29 amu	C ₂ H ₅
	AND [15/27] neutral loss of 30 amu	C ₂ H ₆
	AND [14/27] daughter ion at m/z 15	CH ₃
	AND [16/27] daughter ion at m/z 27	C ₂ H ₃
	AND [23/27] daughter ion at m/z 29	C ₂ H ₅
	THEN functionality indicated is ETHYL	

which implies that *all* of the features in the rule must be present in the spectra of an unknown for an assignment to be made. This is not practical as not all of the spectral features associated with the presence of a particular substructure are seen in the spectra of every compound which contains that substructure. However, in many cases a sufficient proportion will be observed for the association to be made. The match value is ideally set to maximize the reliability of the predictions.

MS and MS/MS spectra of unknowns are entered into an "unknowns database". The MAPS software then looks for high correlations between the empirically derived substructure rules and features in the spectra of the unknowns. This process takes only a few minutes per compound. When such correlations are found, it concludes that the corresponding substructures are present. Spurious m/z values in the unknown's spectra due to contaminants or other components of a mixture may mislead conventional spectral matching algorithms, but will not affect MAPS since, in general, they will not correlate with particular substructures. At this point, GENOA is invoked to perform structure generation. Given an empirical formula and constraints (substructures identified as present in a compound), GENOA exhaustively generates all possible structures.

Evaluation of Rule Performance

The current training set for MAPS includes MS and MS/MS data for 76 compounds. These data comprise 2526 spectral features and 67 different substructures. From this, MAPS was able to generate 45

substructure identification rules containing a total of 2085 features, of which 285 are unique. The number of spectra necessary in the database for meaningful spectra/substructure relationships to be discovered was relatively small. The frequency of erroneous identifications is expected to decrease as the database grows. The range of training set of compounds selected must be broad enough to accurately characterize each substructure.

Substructures are identified in a compound through a simple match of the rules against the set of MS and MS/MS features from the compound. In the absence of any weighting function, the match value specifies the minimum percentage of features in a rule that must be present in MS and MS/MS data for that substructure to be identified.

The predictive capabilities of the rules have been assessed by applying them to all of the compounds in the training set. This function looks for high correlations between the rules and the compounds, and predicts which substructures are present in each compound. These results are then tabulated into two categories: correct and incorrect predictions of the presence of substructures (inclusions). This process is diagrammed in Figure 3.3. The number of correct assignments for each substructure rule are then determined at several match values.

A rule represents a composite of the features characteristic of that substructure. As the number and variety of training set compounds containing that substructure increases, the rule becomes more reliable. Individual training set compounds containing a given substructure may exhibit very few or all of that rule's features. Thus, a valid test of the

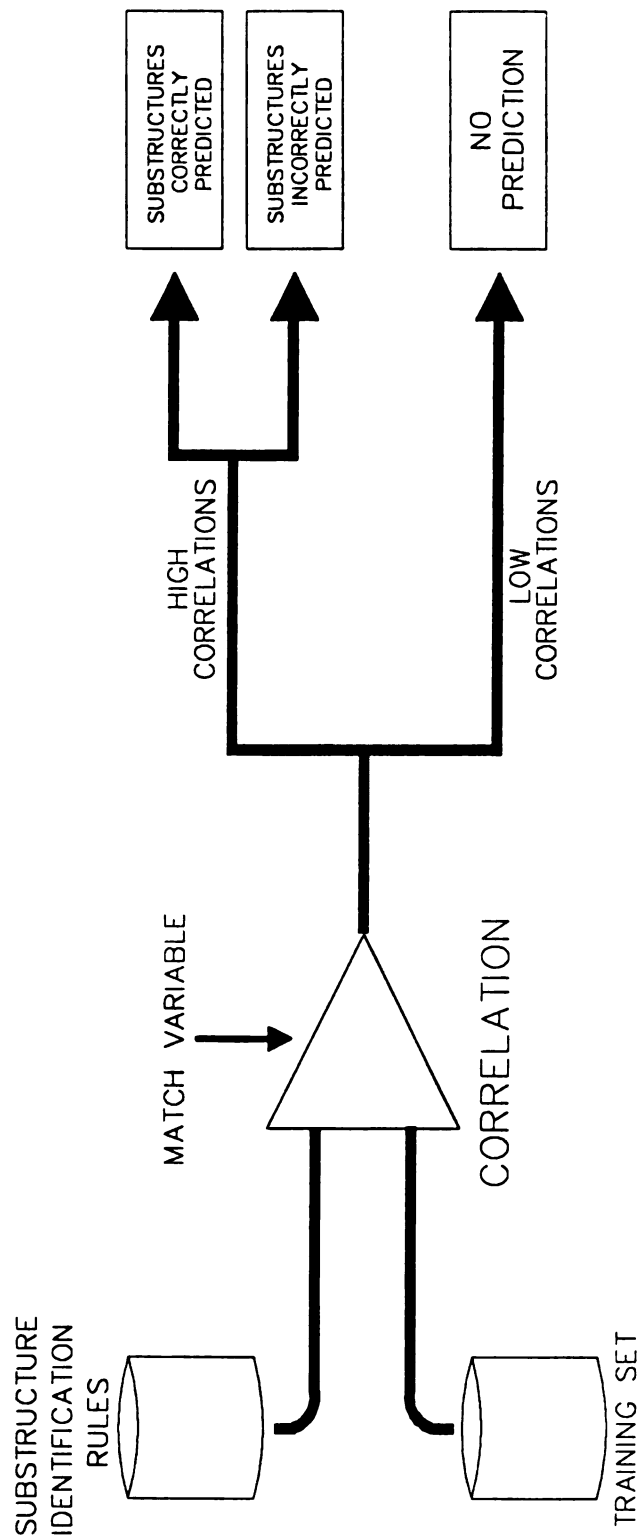


Figure 3.3 Schematic of rule validation process.

predictive capabilities of the rules can be obtained by applying them to the training set.

The predictive capabilities of the rules can be described by three quantities: recall, false positives, and a reliability factor. Recall and false positives are the percentage of correct and incorrect predictions out of the total number possible, respectively. The reliability factor is a measure of the percentage of predictions which are correct. Each of these terms can be calculated for a given match value as shown below.

$$\text{Recall} = \frac{\text{number of correct predictions}}{\text{total number possible}} \times 100\%$$

$$\text{False Positives} = \frac{\text{number of incorrect predictions}}{\text{total number possible}} \times 100\%$$

$$\text{Reliability factor} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \times 100\%$$

Table 3.2 shows recall, false positives, and the reliability factor for selected substructures at different match values. As the match value decreases, recall and false positives both increase. Ideally, the match value should be chosen such that recall is maximized while keeping false positives at the minimum practical level. The optimal match value differs for each substructure rule. For example, incorrect predictions of the presence of substructures occur at match values of *less than* 60% for the xphenyl substructure (which is a phenyl substructure with one substituent) and 90% for the phthalate-ester substructure.

There are certain substructures whose identification is unavoidably ambiguous from MS and MS/MS data alone. These substructures produce rules with poor reliability factors, and unless

Table 3.2 Results of application of selected substructure rules to the training set at several match values.

XPHENYL		34 clauses in rule	
match value	Recall	False Positives	Reliability Factor
100%	0	0	undefined
90%	12	0	100
80%	32	0	100
70%	67	0	100
60%	86	0	100
50%	95	16	95

PHTHALATE-ESTER		61 clauses in rule	
match value	Recall	False Positives	Reliability Factor
100%	0	0	undefined
90%	38	0	100
80%	62	3	71
70%	88	19	35
60%	88	32	24
50%	88	51	17

CARBOXYL		2 clauses in rule	
match value	Recall	False Positives	Reliability Factor
100%	43	18	35
50%	100	95	19

identified and treated specially, they will often be erroneously identified as present in unknowns. Most of these substructures are small and therefore provide few mass spectral features by which they can be characterized. When these features are quite common, which is the case for the carboxyl and carbonyl rules shown below, such rules suffer a relatively large percentage of false positives even at a match value of 100% (as seen for the carboxyl rule in Table 3.2) and, as such, are poor predictors.

```
IF  (12/14) line in primary scan at m/z 45          CO2H
AND (6/14) neutral loss of 28 u                      CO
THEN Functionality indicated is CARBOXYL
```

```
IF (22/23) neutral loss of 28 u                      CO
THEN Functionality indicated is CARBONYL
```

It can be seen that the carbonyl rule contains a single feature, indicating a neutral loss of 28 u. There is no doubt that the rule is correct and complete. However, in many molecules this neutral loss could equally well be attributed to C₂H₄ or N₂. Thus, this rule, used in isolation, would cause some molecules to be erroneously flagged as having carbonyl present. Fortunately the majority of rules contain more than two features. When the rules are generated at a C value of 75%, the average number of features per rule is currently 46, and the largest rule contains 120 features.

Rules with a large number of features can suffer from a different problem, namely that a substructure can be erroneously identified as present when in fact a similar substructure is actually present in a compound. For example, the rule for "pentadecyl" is principally a subset of the rule for "hexadecyl". The presence of the pentadecyl substructure may cause the hexadecyl substructure to be erroneously identified at sufficiently low match values. This type of interference is a result of the somewhat rudimentary matching algorithm used in the initial form of this software. This interference can at present be predicted from a cross correlation of the rules, which indicates their degree of commonality. Part of such a cross correlation is shown in Table 3.3. Correlation factors of 1.0 indicate that all of the features in the first rule are contained within the second, whereas factors of 0.0 indicate that the two rules have no common features. Note in particular the correlations between alkyl substructures (high), carbonyl-containing substructures (high), and widely different substructures such as bromo and benzoyl (low).

MAPS recognizes unreliable rules on the basis of their performance when applied to the training set. A rule is designated as "unreliable" when its reliability factor becomes less than 50%. MAPS maintains and automatically updates a list of the substructures whose rules are unreliable and notes the match values at which they become unreliable (Table 3.4). This list grows as the match value decreases, since the reliability of a rule decreases as a function of the match value. At match values of 100%, the list mainly includes the small, poorly characterized substructures mentioned previously; the identification of these

Table 3.3 Partial results from cross correlation of the rules.

1.0	between ALDEHYDE and CARBONYL
1.0	between ALDEHYDE and CARBOXYL
0.0	between ALDEHYDE and CHLORO
1.0	between ALDEHYDE and ESTER
1.0	between ALDEHYDE and ETHYL
1.0	between ALDEHYDE and PHTHALATE
0.026	between BENZOYL and BROMO
0.842	between BENZOYL and PHTHALATE
0.842	between BENZOYL and PHTHALATE-ESTER
0.0	between BUTYL and BROMO
0.903	between BUTYL and DECYL
0.967	between BUTYL and HEPTYL
0.967	between BUTYL and HEXYL
0.967	between BUTYL and PENTYL
0.806	between BUTYL and TRIDECYL
0.0	between CHLORO and BROMO
0.0	between CHLORO and ESTER
0.0	between CHLORO and NONYL
0.0	between CHLORO and PHTHALATE-ESTER
0.008	between PENTADECYL and BROMO
0.818	between PENTADECYL and DODECYL
1.0	between PENTADECYL and HEXADECYL
1.0	between PENTADECYL and TETRADECYL
0.956	between PENTADECYL and TRIDECYL

substructures is unreliable at *any* match value using MS and MS/MS data alone. At a match value of 50%, the list includes nearly all of the substructures for which there are rules. However, the optimal match value is greater than this for nearly all the rules. A majority of the features in the rules are derived from MS data, which are not as specific as MS/MS data. Some rules become unreliable at low match values because they contain a large percentage of MS features. Later versions of MAPS replaced the match value with a criterion which considers the predictive value of each rule clause.

Table 3.4 Unreliable rule lists at three different match values.

match value	SUBSTRUCTURES WHOSE RULES ARE UNRELIABLE
100%	ester, aldehyde, carboxyl, carbonyl, methoxy, ethoxy, primary-alcohol, xphenyl
75%	the above substructures plus dimethylamino, 1-2-4-5-phenyl, and tolyl
50%	39 out of the 45 recognized substructures for which there are rules

MAPS can exclude unreliable rules from consideration during analysis of unknowns. This substantially decreases the number of incorrect assignments. As shown in Figure 3.4, the number of correct predictions is also decreased through removal of unreliable rules. However, the quality of the remaining predictions (as described by the reliability factor), which is of prime importance, is markedly improved as shown in Figure 3.4. The overall reliability factor for *all* the rules is 80%

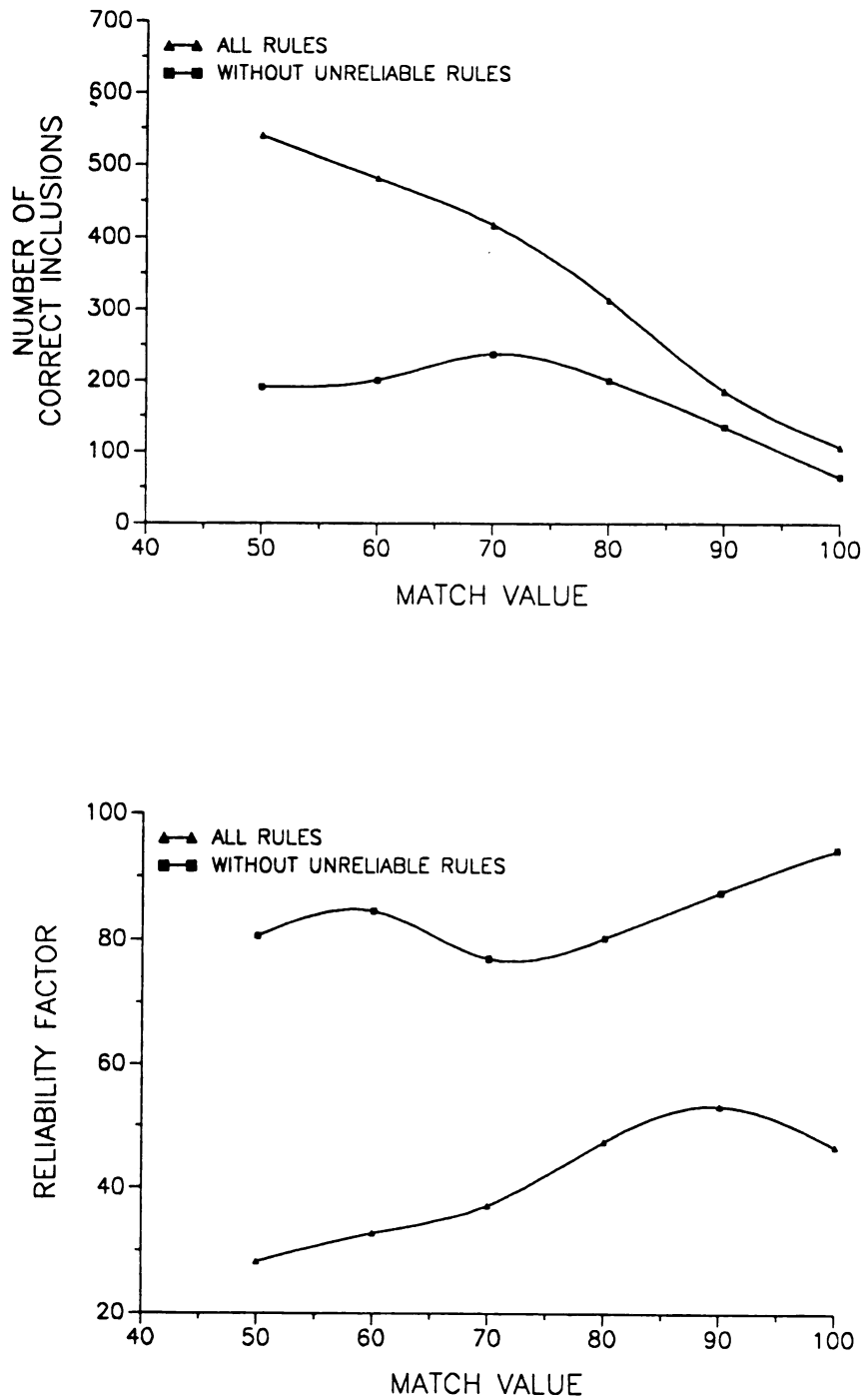


Figure 3.4 Performance graphs for rules generated by MAPS at a C value of 75% showing the effect of exclusion of unreliable rules.

when they are applied to the training set at a match value of 75%. When the unreliable rules are removed from consideration, the reliability factor approaches 98%. These results can be improved even further by optimizing the C and match values for each substructure.

The MAPS rules showed much promise for spectral interpretation, screening for particular classes of compounds, and structure elucidation. Several different methodologies were envisioned to improve the quality and number of predictions made by the rules generated by this initial version of the MAPS code. These are discussed in detail in the next two chapters.

References

1. McLafferty, F.W., "Interpretation of Mass Spectra", University Science Books, Mill Valley, CA, 1980, p. 282.
2. McLafferty, F.W., Venkataraghavan, R., "Mass Spectral Correlations", American Chemical Society, Washington, DC, 1982.
3. Cross, K.C., Palmer, P.T., Giordani, A.B., Beckner, C.F., Hoffman, P.A., Gregg, H.R., Enke, C.G., in Pierce, T.H., Hohne, B.H. (Eds.), "Artificial Intelligence Applications in Chemistry", ACS Symposium Series No. 306, American Chemical Society, Washington, DC, 1986, p. 321.
4. Hunt, D.F., Shabinowitz, J., Harvey, T.M., Coates, M., Anal. Chem., 57, 525 (1985).
5. Bauer, M.R., Ph.D. Thesis, Michigan State University, East Lansing, MI, 1987.
6. Wood, K.V., Cooks, R.G., Laugal, J.A., Benkeser, R.A., Anal. Chem., 57, 692 (1985).
7. Finnigan MAT, 355 River Oaks Parkway, San Jose, CA, 95134.
8. Stults, J.T., Enke, C.G., Holland, J.F., Anal. Chem., 55, 1323 (1983).
9. Eckenrode, B., Newcome, B.H., Holland, J.F., Enke, C.G., 35th Annual Conf. on Mass Spectrometry and Allied Topics, 1987, p. 241.
10. Holland, J.F., Erickson, E.D., Eckenrode, B., Watson, J.T., 35th Annual Conf. on Mass Spectrometry and Allied Topics, 1987, p. 277.
11. Wade, A.P., Palmer, P.T., Hart, K.J., Enke, C.G., Anal. Chim. Acta, in press.
12. Wade, A.P., Palmer, P.T., Hart, K.J., Enke, C.G., 34th Annual Conf. on Mass Spectrometry and Allied Topics, 1986, p. 426.
13. Xerox Corporation, Artificial Intelligence Systems, 250 N. Halstead St., Pasadena, CA, 91109.
14. KEE Manufacturer's Literature, Intellicorp, 1975 El Camino Real W., Mountain View, CA, 94040.
15. McCarthy, J., "History of LISP", SIGPLAN Notices, p. 217.

16. Winston, P.H., Horn, B.K.P., "LISP", Addison-Wesley, Menlo Park, CA, 1984.
17. Steele, G.L., "Common Lisp", Digital Press, Maynard, MA, 1984.
18. Sheil, B., Masinter, L.M. (Eds.), "Papers on InterLISP-D", Xerox PARC, Palo Alto, CA, 1983.
19. Teitelman, W., Masinter, L.W., IEEE Transactions on Computers, 4, 25 (1981).
20. Allen, J.R., AI Expert, 2, 48 (1987).

CHAPTER 4

MODIFICATIONS TO THE MAPS SOFTWARE

Introduction

The MAPS program described in Chapter 3 showed great promise for structure elucidation via substructure identification. However, further modifications to the code were required to improve the quality and predictive capabilities of the rules. This chapter describes the modifications which have been made to the rule generation and application processes for these purposes. In the previous version of MAPS, MS and MS/MS spectral feature/substructure relationships were used to generate rules for predicting the *presence* of substructures. These relationships can be exploited for predicting the *absence* of substructures as well. Such negative information (substructures known to be absent) can be used by GENOA to constrain the generation of candidate structures and thus is useful information. Since different criteria exist for predicting the presence and absence of substructures from MS and MS/MS data, it became necessary to differentiate between the two different types of rules required to predict the presence and absence of substructures. Henceforth, these two rule types are referred to as inclusion and exclusion rules, which are used to predict the presence and absence of substructures, respectively. Since exclusion

rules use the same spectral feature/substructure relationships as inclusion rules, modification of the rule generation process to obtain exclusion rules was relatively straightforward.

The correspondence between a given feature f_i and a given substructure ss_j can be described by correlation and uniqueness factors as shown below.

$$\text{Correlation factor } (f_i, ss_j) = \frac{\# \text{ occurrences of } f_i \text{ and } ss_j}{\# \text{ occurrences of } ss_j}$$

$$\text{Uniqueness factor } (f_i, ss_j) = \frac{\# \text{ occurrences of } f_i \text{ and } ss_j}{\# \text{ occurrences of } f_i}$$

These factors are obtained from training set statistics. Certain features appear in MS and MS/MS spectra whenever a specific substructure is present in compounds analyzed under similar instrumental conditions. The absence of such features suggests the absence of that substructure. Thus, common sense indicates that each exclusion rule should contain features with high correlation factors. However, these same features may not be useful for predicting the presence of that substructure if they have moderate to high correlation factors with other substructures as well, since they may lead to false positives. Thus, inclusion rules should emphasize features which have high uniqueness factors. The MAPS code has been modified to exploit these criteria for the generation of both rule types.

Intensity data have been incorporated into the rule generation process through the use of three intensity classes (strong, medium, and weak). In addition, a weighting scheme which exploits the uniqueness factor of rule clauses has been developed for the rule application process.

This chapter describes the modifications to the rule generation and application processes and discusses the effects of the various parameters associated with rule generation and application, intensity data, and uniqueness weighting on rule performance (1,2). The results have been used to optimize the performance of both rule types. Lastly, a substructure hierarchy which describes the inheritance relationships between substructures has been developed to eliminate redundant information produced by a MAPS analysis of the substructures present in an unknown.

Modifications to the Rule Generation Process

The rule generation process has been described in detail in Chapter 3 and elsewhere (3). Several modifications to this process were required to improve the predictive capabilities of both rule types. Briefly, the rule generation process works as follows. MAPS extracts several types of features from the MS and MS/MS data fields of known compounds. These include the m/z values seen in conventional and daughter spectra, neutral losses, and parent-to-daughter transitions. This spectral feature data combined with substructure data for each known compound comprises the training set. Correlation and uniqueness factors are calculated for each feature-substructure combination and are used to determine whether or not that feature should be included in the rule for that substructure. For each substructure, low and high fragment mass limits and constraints to define legal fragment masses based on the substructure's elemental

composition are used to remove features which cannot be attributed to the substructure from its rule. By this process, a complete rule set is generated. Since different criteria exist for predicting the presence and absence of substructures, optimization of inclusion and exclusion rules required the development of two different filters for rule generation.

Use of Correlation and Uniqueness Filters in Rule Generation. One of two filters may be applied in the rule generation: a correlation filter or a uniqueness filter. If the correlation filter is used, the user must specify the value of the *C variable*, which represents the *minimum correlation factor* necessary for a feature to be included in a rule. Likewise, if the uniqueness filter is used, the user must specify the value of the *U variable*, which represents the *minimum uniqueness factor* necessary for a feature to be included in a rule. Table 4.1 shows the exclusion rule for the phthalate-ester substructure generated at three different C values. Table 4.2 shows the inclusion rule for the phthalate-ester substructure generated at three different U values. Correlation or uniqueness factors are printed along with each clause in the rules and are designated by the letters CF or UF, respectively. Fragment formulae for each clause in the rules, which are normally postulated by MAPS based on the elemental composition of the substructure and valence rules, are not shown for the rules in this table. Rule length increases as the C or U value decreases, as more features have the necessary levels of correlation or uniqueness. Rules may be generated at different C and U values, and their content and performance thus depend on these variables.

Table 4.1 Exclusion rules for the phthalate-ester substructure generated at three different C values: a) C = 70%, b) C = 85%, and c) C = 100%.

```

a) IF NO [CF= 88%] daughter ion at m/z 39
   OR NO [CF=100%] daughter ion at m/z 65
   OR NO [CF= 75%] daughter ion at m/z 76
   OR NO [CF= 88%] daughter ion at m/z 77
   OR NO [CF=100%] daughter ion at m/z 93
   OR NO [CF= 88%] daughter ion at m/z 121
   OR NO [CF= 88%] daughter ion at m/z 149
   OR NO [CF= 75%] neutral loss of 26 amu
   OR NO [CF=100%] neutral loss of 28 amu
   OR NO [CF=100%] neutral loss of 56 amu
   OR NO [CF=100%] neutral loss of 84 amu
   OR NO [CF= 75%] neutral loss of 148 amu
   OR NO [CF= 75%] line in primary scan at m/z 41
   OR NO [CF= 88%] line in primary scan at m/z 43
   OR NO [CF= 88%] line in primary scan at m/z 45
   OR NO [CF=100%] line in primary scan at m/z 50
   OR NO [CF= 88%] line in primary scan at m/z 51
   OR NO [CF= 75%] line in primary scan at m/z 53
   OR NO [CF= 75%] line in primary scan at m/z 55
   OR NO [CF= 75%] line in primary scan at m/z 57
   OR NO [CF=100%] line in primary scan at m/z 63
   OR NO [CF=100%] line in primary scan at m/z 65
   OR NO [CF= 88%] line in primary scan at m/z 66
   OR NO [CF= 75%] line in primary scan at m/z 67
   OR NO [CF= 75%] line in primary scan at m/z 71
   OR NO [CF=100%] line in primary scan at m/z 76
   OR NO [CF=100%] line in primary scan at m/z 77
   OR NO [CF= 75%] line in primary scan at m/z 78
   OR NO [CF= 75%] line in primary scan at m/z 83
   OR NO [CF=100%] line in primary scan at m/z 93
   OR NO [CF= 75%] line in primary scan at m/z 94
   OR NO [CF=100%] line in primary scan at m/z 104
   OR NO [CF=100%] line in primary scan at m/z 105
   OR NO [CF= 88%] line in primary scan at m/z 106
   OR NO [CF= 88%] line in primary scan at m/z 121
   OR NO [CF= 88%] line in primary scan at m/z 123
   OR NO [CF=100%] line in primary scan at m/z 132
   OR NO [CF= 88%] line in primary scan at m/z 133
   OR NO [CF= 75%] line in primary scan at m/z 135
   OR NO [CF=100%] line in primary scan at m/z 149
   OR NO [CF= 88%] line in primary scan at m/z 150
   OR NO [CF= 88%] line in primary scan at m/z 151
   OR NO [CF= 75%] daughter of m/z 65 from m/z 93 (28 amu)

```

Table 4.1 (continued)

OR NO [CF= 75%] daughter of m/z 65 from m/z 121 (56 amu)
 OR NO [CF= 88%] daughter of m/z 65 from m/z 149 (84 amu)
 OR NO [CF= 75%] daughter of m/z 77 from m/z 105 (28 amu)
 OR NO [CF= 75%] daughter of m/z 93 from m/z 121 (28 amu)
 OR NO [CF= 88%] daughter of m/z 93 from m/z 149 (56 amu)
 OR NO [CF= 88%] daughter of m/z 121 from m/z 149 (28 amu)
 THEN the PHTHALATE-ESTER substructure is ABSENT

b) IF NO [CF= 88%] daughter ion at m/z 39
 OR NO [CF=100%] daughter ion at m/z 65
 OR NO [CF= 88%] daughter ion at m/z 77
 OR NO [CF=100%] daughter ion at m/z 93
 OR NO [CF= 88%] daughter ion at m/z 121
 OR NO [CF= 88%] daughter ion at m/z 149
 OR NO [CF=100%] neutral loss of 28 amu
 OR NO [CF=100%] neutral loss of 56 amu
 OR NO [CF=100%] neutral loss of 84 amu
 OR NO [CF= 88%] line in primary scan at m/z 43
 OR NO [CF= 88%] line in primary scan at m/z 45
 OR NO [CF=100%] line in primary scan at m/z 50
 OR NO [CF= 88%] line in primary scan at m/z 51
 OR NO [CF=100%] line in primary scan at m/z 63
 OR NO [CF=100%] line in primary scan at m/z 65
 OR NO [CF= 88%] line in primary scan at m/z 66
 OR NO [CF=100%] line in primary scan at m/z 76
 OR NO [CF=100%] line in primary scan at m/z 77
 OR NO [CF=100%] line in primary scan at m/z 93
 OR NO [CF=100%] line in primary scan at m/z 104
 OR NO [CF=100%] line in primary scan at m/z 105
 OR NO [CF= 88%] line in primary scan at m/z 106
 OR NO [CF= 88%] line in primary scan at m/z 121
 OR NO [CF= 88%] line in primary scan at m/z 123
 OR NO [CF=100%] line in primary scan at m/z 132
 OR NO [CF= 88%] line in primary scan at m/z 133
 OR NO [CF=100%] line in primary scan at m/z 149
 OR NO [CF= 88%] line in primary scan at m/z 150
 OR NO [CF= 88%] line in primary scan at m/z 151
 OR NO [CF= 88%] daughter of m/z 65 from m/z 149 (84 amu)
 OR NO [CF= 88%] daughter of m/z 93 from m/z 149 (56 amu)
 OR NO [CF= 88%] daughter of m/z 121 from m/z 149 (28 amu)
 THEN the PHTHALATE-ESTER substructure is ABSENT

Table 4.1 (continued)

c) IF NO [CF=100%] daughter ion at m/z 65
OR NO [CF=100%] daughter ion at m/z 93
OR NO [CF=100%] neutral loss of 28 amu
OR NO [CF=100%] neutral loss of 56 amu
OR NO [CF=100%] neutral loss of 84 amu
OR NO [CF=100%] line in primary scan at m/z 50
OR NO [CF=100%] line in primary scan at m/z 63
OR NO [CF=100%] line in primary scan at m/z 65
OR NO [CF=100%] line in primary scan at m/z 76
OR NO [CF=100%] line in primary scan at m/z 77
OR NO [CF=100%] line in primary scan at m/z 93
OR NO [CF=100%] line in primary scan at m/z 104
OR NO [CF=100%] line in primary scan at m/z 105
OR NO [CF=100%] line in primary scan at m/z 132
OR NO [CF=100%] line in primary scan at m/z 149
THEN the PHTHALATE-ESTER substructure is ABSENT

Table 4.2 Inclusion rules for the phthalate-ester substructure generated at three different U values: a) U = 33%, b) U = 50%, and c) U = 100%.

- a) IF [UF= 80%] daughter ion at m/z 76s
 OR [UF= 33%] neutral loss of 84m amu
 OR [UF= 42%] neutral loss of 148s amu
 OR [UF=100%] neutral loss of 166m amu
 OR [UF= 50%] line in primary scan at m/z 76s
 OR [UF= 75%] line in primary scan at m/z 104s
 OR [UF= 33%] line in primary scan at m/z 149s
 OR [UF= 42%] daughter of m/z 39s from m/z 93 (54 amu)
 OR [UF= 35%] daughter of m/z 65s from m/z 93 (28 amu)
 OR [UF= 75%] daughter of m/z 65s from m/z 121 (56 amu)
 OR [UF= 80%] daughter of m/z 65m from m/z 149 (84 amu)
 OR [UF=100%] daughter of m/z 76s from m/z 104 (28 amu)
 OR [UF= 71%] daughter of m/z 93s from m/z 149 (56 amu)
 OR [UF= 38%] daughter of m/z 121m from m/z 149 (28 amu)
 THEN the PHTHALATE-ESTER substructure is PRESENT
- b) IF [UF= 80%] daughter ion at m/z 76s
 OR [UF=100%] neutral loss of 166m amu
 OR [UF= 50%] line in primary scan at m/z 76s
 OR [UF= 75%] line in primary scan at m/z 104s
 OR [UF= 75%] daughter of m/z 65s from m/z 121 (56 amu)
 OR [UF= 80%] daughter of m/z 65m from m/z 149 (84 amu)
 OR [UF=100%] daughter of m/z 76s from m/z 104 (28 amu)
 OR [UF= 71%] daughter of m/z 93s from m/z 149 (56 amu)
 THEN the PHTHALATE-ESTER substructure is PRESENT
- c) IF [UF=100%] neutral loss of 166m amu
 OR [UF=100%] daughter of m/z 76s from m/z 104 (28 amu)
 THEN the PHTHALATE-ESTER substructure is PRESENT

Use of Intensity Data in Rule Generation. The previous version of MAPS did not use intensity data from MS and MS/MS spectra. These intensities are dependent on a large number of instrumental parameters as shown in Chapter 2. In addition, they are not as important as the presence or absence of a particular feature for identifying substructures. Hence, they play a decreased role in identifying the presence and absence of substructures. For these reasons, intensity classes rather than numerical intensities were used for implementing intensity data into the rules. A given intensity I_X which has been normalized to the base peak is categorized into one of the three defined intensity classes using the equations shown below.

$$\begin{array}{ll} \text{strong:} & I_X \geq 10 \\ \text{medium:} & 1 \leq I_X < 10 \\ \text{weak:} & I_X < 1 \end{array}$$

When intensity data are used in the rule generation, the intensity of each feature is converted to an intensity class and associated with that feature.

When a rule is generated with intensity data, it may contain multiple clauses for the same feature with *different intensity classes*. This only occurs for neutral losses and daughter ions, since these features may appear in several daughter spectra with different intensities. For example, the ethyl rule with intensities in Table 4.3 contains two clauses for daughter ion at m/z 29 and neutral losses of 26 and 28 u, with associated intensity classes of medium and strong.

Table 4.3 Exclusion rules for the ethyl substructure generated at a C value of 50%. Parts a and b represent the ethyl exclusion rule generated with and without use of intensity data, respectively.

a)	IF NO [CF=52%] daughter ion at m/z 29m	C ₂ H ₅
	OR NO [CF=56%] daughter ion at m/z 29s	C ₂ H ₅
	OR NO [CF=59%] neutral loss of 2m amu	H ₂
	OR NO [CF=56%] neutral loss of 15s amu	CH ₃
	OR NO [CF=56%] neutral loss of 16m amu	CH ₄
	OR NO [CF=52%] neutral loss of 26m amu	C ₂ H ₂
	OR NO [CF=63%] neutral loss of 26s amu	C ₂ H ₂
	OR NO [CF=74%] neutral loss of 28m amu	C ₂ H ₄
	OR NO [CF=93%] Neutral loss of 28s amu	C ₂ H ₄
	THEN the ETHYL substructure is ABSENT	
b)	IF NO [CF= 52%] daughter ion at m/z 15	CH ₃
	OR NO [CF= 59%] daughter ion at m/z 27	C ₂ H ₃
	OR NO [CF= 85%] daughter ion at m/z 29	C ₂ H ₅
	OR NO [CF= 74%] neutral loss of 2 amu	H ₂
	OR NO [CF= 74%] neutral loss of 15 amu	CH ₃
	OR NO [CF= 78%] neutral loss of 16 amu	CH ₄
	OR NO [CF= 67%] neutral loss of 18 amu	CH ₄ + H ₂
	OR NO [CF= 81%] neutral loss of 26 amu	C ₂ H ₂
	OR NO [CF=100%] neutral loss of 28 amu	C ₂ H ₄
	OR NO [CF= 55%] neutral loss of 29 amu	C ₂ H ₅
	OR NO [CF= 55%] neutral loss of 30 amu	C ₂ H ₆
	THEN the ETHYL substructure is ABSENT	

When intensity data are used in the rules, the letter s, m, or w is associated with each feature mass. These letters correspond to intensity classes of strong, medium, and weak, respectively. Fragment formulae are not shown the rules in this table as well.

Since the use of intensity data decreases both the number of occurrences of a given feature f_i and substructure ss_j and the number of occurrences of f_i , correlation and uniqueness factors for features with intensity data are not the same as those for features without intensity data. Thus, intensity data affects rule content. This is now shown for rules generated at given C and U values.

When an intensity class is associated with a feature, the correlation factor of that feature without intensity data is divided among three equivalent features with intensity classes of strong, medium, and weak. Thus, features with intensity data have lower correlation factors. Because of this, rules with intensity data generated at a given C value usually contain fewer features. This is demonstrated in Table 4.3, which shows exclusion rules for the ethyl substructure generated at a C value of 100%, with and without the use of intensity data. Note that several clauses in the rule without intensity data (daughter ions at m/z 15 and 27, and neutral losses of 18, 29, and 30 u) do not appear in the rule with intensities, since they do not possess the required correlation factors. Also note that the correlation factors for common clauses between these two rules are smaller when intensity data is used. Using a C value of 100%, 25 exclusion rules were generated with an average of 7 clauses per rule with intensity data, while 41 exclusion rules were generated with an average of 14 clauses per rule without intensity data. Thus, intensity

data should not be used in generating exclusion rules at any given C value in order to increase the number of rules generated and the number of clauses per rule.

Since the compounds which contain a given feature are divided among three equivalent features with intensity classes of strong, medium, and weak when an intensity class is associated with that feature, the uniqueness factors for features with intensities may be greater than, less than, or equal to the uniqueness factor of the same feature without intensity data. Usually, at least one of the features with an associated intensity class has a higher uniqueness factor. Thus, at a given U value, inclusion rules with intensity data will have more features than those without intensity data. This is demonstrated in Table 4.4, which shows inclusion rules for the benzyl substructure generated at a U value of 100% with and without the use of intensity data. Note that several clauses in the rule with intensity data (neutral losses of 75 and 78, line in primary scan at m/z 90, daughter of m/z 29 from m/z 69, and daughter of m/z 51 from 91) do not appear in the rule without intensities, since they do not possess the required uniqueness factors. Using a U value of 100%, 18 inclusion rules were generated with an average of 6 clauses per rule without intensity data, while 22 inclusion rules were generated with an average of 9 clauses per rule with intensity data. Thus, intensity data should be used in generating inclusion rules at any given U value in order to increase the number of rules generated and the number of clauses per rule.

Table 4.4 Inclusion rules for the benzyl substructure generated at a U value of 100%. Parts a and b represent the benzyl inclusion rule generated with and without use of intensity data, respectively.

a)	IF neutral loss of 60m amu	C ₅
	OR neutral loss of 60s amu	C ₅
	OR neutral loss of 75m amu	C ₆ H ₃
	OR neutral loss of 78w amu	C ₆ H ₆
	OR neutral loss of 92m amu	C ₇ H ₈
	OR line in primary scan at 90s amu	C ₇ H ₆
	OR daughter of m/z 29s from m/z 69 (40 amu)	C ₃ H ₄
	OR daughter of m/z 51m from m/z 65 (14 amu)	C ₂ H ₂
	OR daughter of m/z 51w from m/z 91 (40 amu)	C ₃ H ₄
	OR daughter of m/z 91s from m/z 93 (2 amu)	H ₂
	THEN the BENZYL substructure is PRESENT	
b)	IF neutral loss of 60 amu	C ₅
	OR neutral loss of 62 amu	C ₅ H ₂
	OR neutral loss of 92 amu	C ₇ H ₈
	OR daughter of m/z 49 from m/z 77 (28 amu)	C ₂ H ₄
	OR daughter of m/z 51 from m/z 65 (14 amu)	C ₂ H ₂
	OR daughter of m/z 91 from m/z 93 (2 amu)	H ₂
	THEN the BENZYL substructure is PRESENT	

Modifications to the Rule Application Process

The rule application process has been described in Chapter 3 and elsewhere (3). Several modifications to this process were required to predict the absence as well as the presence of substructures. Briefly, the rule application process works as follows. MS and MS/MS spectra of an unknown are entered into an "unknowns database". MAPS then extracts the standard features from these data; that is, lines present in conventional mass spectra and daughter spectra, neutral losses from daughter spectra, and parent-to-daughter transitions. Each rule is then applied to the unknown to identify the substructures which are present and absent. At this point, GENOA is invoked to generate all possible structures consistent with a given molecular formula and the substructural constraints identified by MAPS.

Identifying the Presence and Absence of Substructures. The performance of a rule can be evaluated as a function of a "match value", which specifies the minimum or maximum percentage of features from a rule that must appear in an unknown's MS and MS/MS spectra for a prediction to be made. In applying a rule to an unknown, MAPS first identifies the degree of match between each rule and the unknown. In the absence of any weighting, the degree of match is simply the number of common clauses between the rule and the unknown divided by the number of clauses in the rule. For exclusion, a substructure is predicted to be absent when the degree of match is *less than or equal* to the match value. For inclusion, a substructure is predicted to be present if the

degree of overlap is *greater than or equal* to the match value. Thus, the logical OR's in the exclusion rules shown in Table 4.1 are enforced only at match values of 99.9999%, which implies that at least one of the features in the rule must be absent from the unknown spectra for a prediction to be made. Likewise, the logical OR's in the inclusion rules shown in Table 4.2 are enforced only at match values of 0.0001%, which implies that at least one of the features in the rule must be present in the unknown spectra for a prediction to be made. Both of these previous two statements hold true for rules which have two million clauses or less. Spurious m/z values in the unknown's spectra due to contaminants or other components of a mixture may mislead conventional spectral matching algorithms, but will not affect MAPS since, in general, they will not correlate with particular substructures.

The predictive capabilities of the rules have been assessed by applying them to all of the compounds in the training set. This function looks for high or low correlations between the rules and the compounds, and predicts which substructures are present in or absent from each compound. These results are then tabulated into four categories: correct and incorrect predictions of the presence and absence of substructures. This process is diagrammed in Figure 4.1. The predictive capabilities of the rules are described by three quantities which have been defined in Chapter 3: recall, false positives, and a reliability factor. Recall and false positives are the percentage of correct and incorrect predictions out of the total number possible, respectively. The reliability factor is a measure of the percentage of predictions which are correct. These terms are calculated for rules generated at a given C or U value and applied to

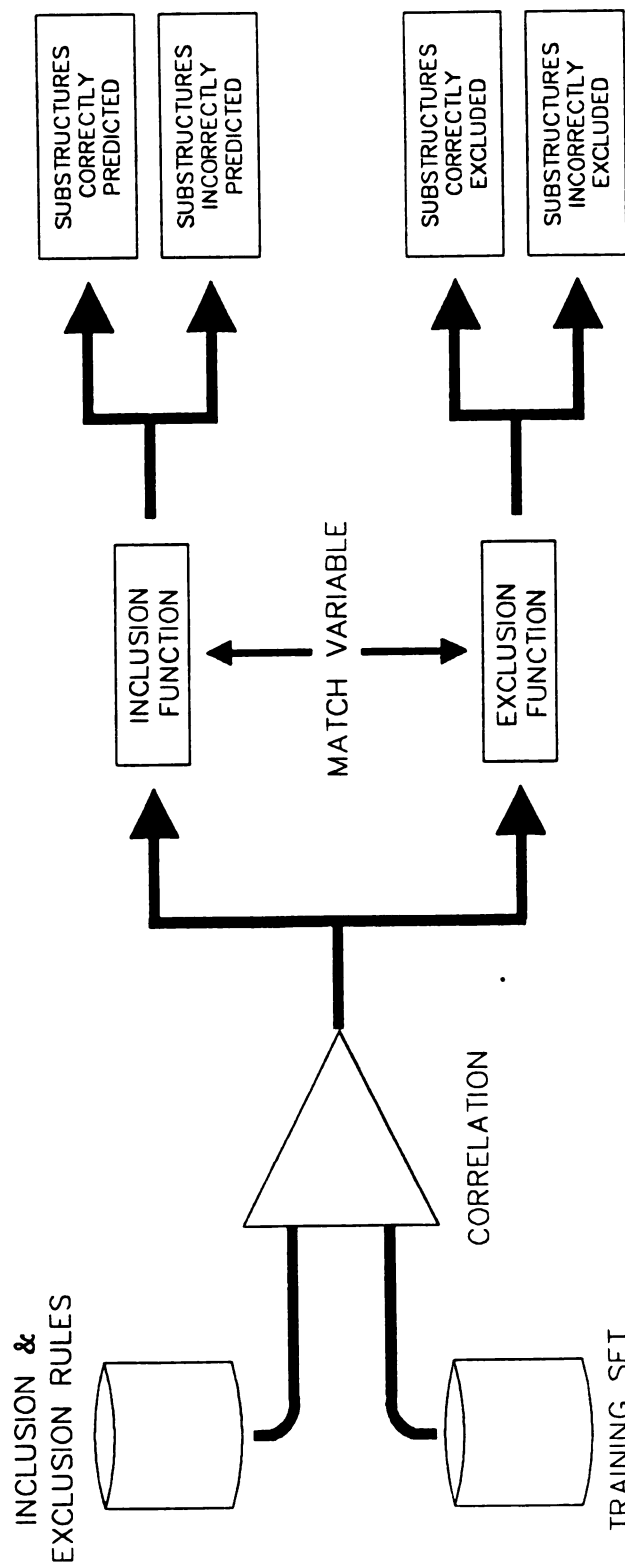


Figure 4.1 Schematic of rule validation process.

the training set at a given match value. A false positive guarantees that the set of candidate structures will *not* contain the correct structure. Therefore, only predictions of the highest quality should be used by GENOA for structure generation. *Thus, in optimizing rule performance, the main criteria are minimizing false positives without reducing recall to the level of uselessness.*

Development of a Weighting Function which uses Uniqueness Factors. A weighting scheme was developed in an attempt to improve rule performance. This scheme takes into the account the uniqueness factor for each clause in a rule. In this respect, it is similar to McLafferty's probability-based matching (PBM) system, which also weights features by their uniqueness with respect to the database (4). The following equation is used to describe the degree of match using uniqueness weighting.

$$\text{degree of match} = \frac{\sum_{i=1}^n (U_i * X)}{\sum_{i=1}^n U_i}$$

where n = number of clauses in rule

U_i = uniqueness factor for feature f_i and substructure ss_j

$X = 1$ if unknown possesses rule clause

$X = 0$ if unknown does not possess rule clause

A fuzzy-logic-based scheme was developed for matching the real intensities in unknowns to the intensity classes in the rules, but was found to degrade rule performance.

Exclusion Rule Optimization

Intensity information is usually not considered for predicting the absence of substructures, since it is the absence of certain characteristic features which indicate the absence of the corresponding substructure. In an attempt to verify this, the performance of the exclusion rules was evaluated with and without the use of intensity data. Figure 4.2 plots recall versus the match value for all exclusion rules (generated at a C value of 100%) applied to the training set, with and without the use of intensity data in the rules. False positives and the reliability factor were not plotted as they were 0% and 100% respectively at all match values for rules generated at this C value. More exclusion rules with a greater number of features per rule are produced without the use of intensity data, since more features have the required minimum correlation factor. Thus, when intensity data is not used, higher recall values are achieved as shown in Figure 4.2. This is because it is the absence of a particular feature rather than the absence of a specific feature-intensity combination which is more important in predicting the absence of a substructure.

Figure 4.3 shows plots of recall, false positives, and the reliability factor as a function of the match value when exclusion rules (without the use of intensity data or weighting) generated at three different C values are applied to the training set. This information is used to determine the optimal C and match values for exclusion rules. As the match value increases, the number of rule clauses required to be absent from an unknown for an exclusion prediction to be made decreases. Hence,

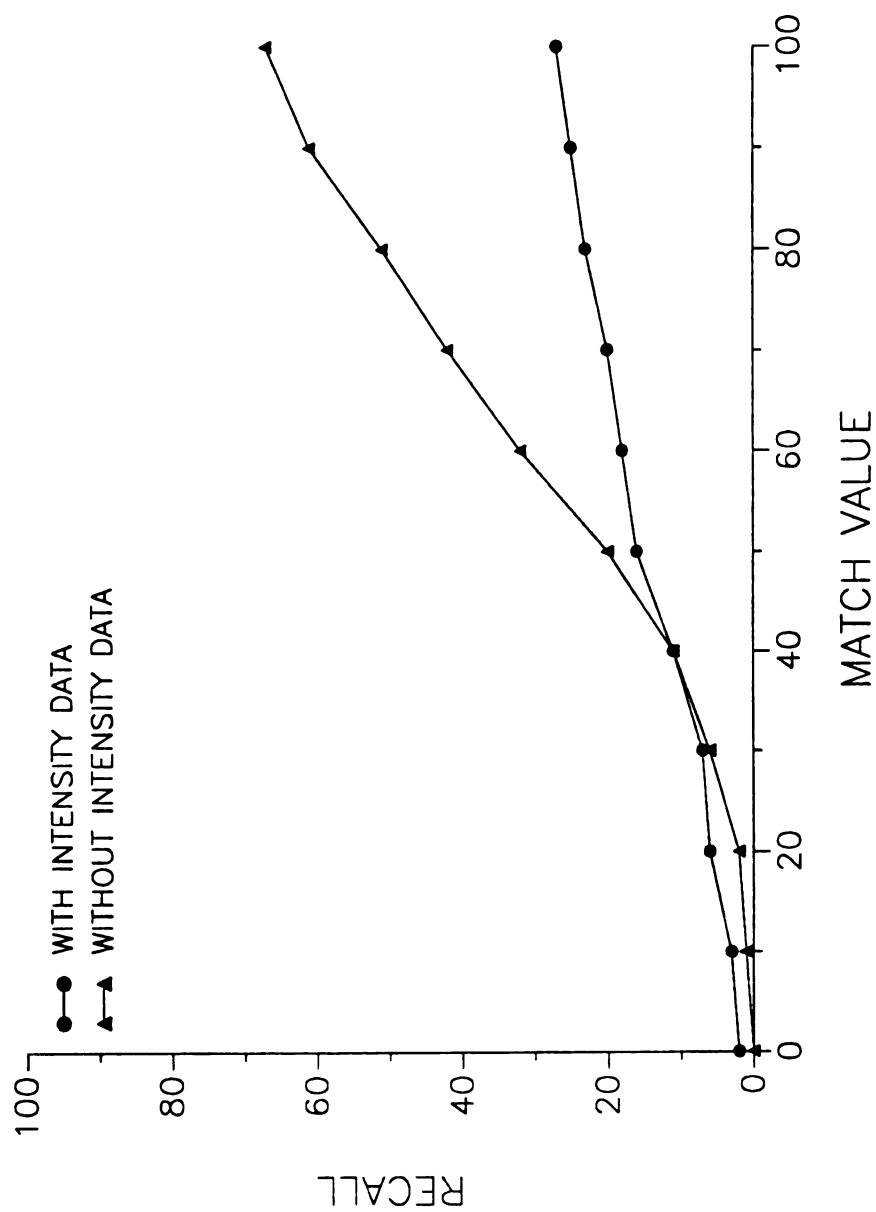


Figure 4.2 Recall versus match value for exclusion rules (generated at a C value of 100%) applied against the training set, with and without use of intensity data in the exclusion rules.

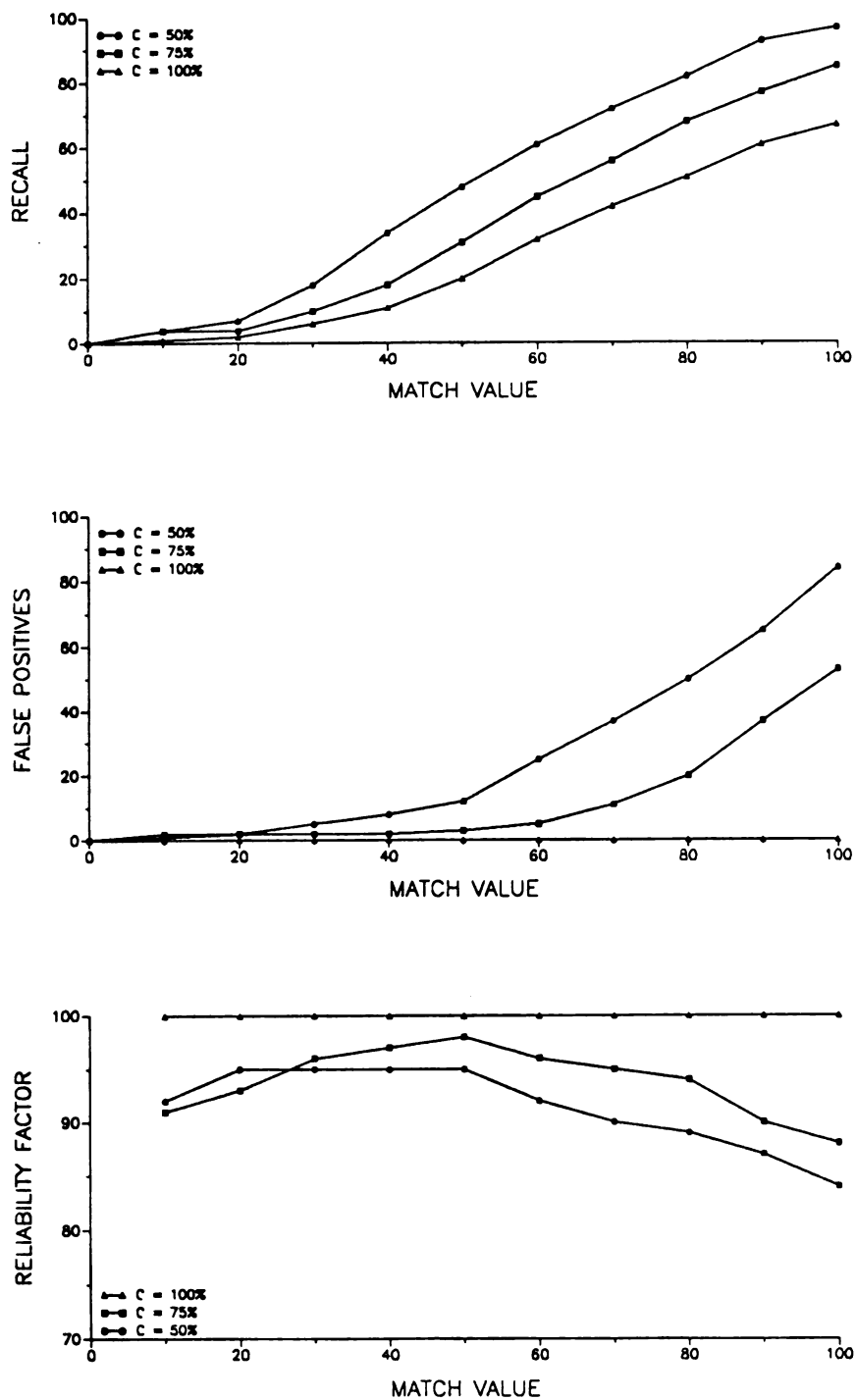


Figure 4.3 Recall, false positives, and the reliability factor versus match value for exclusion rules generated at three different C values applied to the training set.

higher match values essentially loosen the conditions necessary for an exclusion prediction to be made. This is demonstrated in Figure 4.3, where both recall and false positives increase as the match value increases. The match value specifies the *minimum* degree of match necessary between an exclusion rule and an unknown for a prediction to be made. Thus, given an exclusion rule which contains 10 clauses, a match value of 90% would cause a prediction to occur if 9 rule clauses or less were present in the unknown. Conversely, if the unknown was missing at least one of the 10 rule clauses, a prediction would be made at this match value. Now consider both ends of the extremes for the match value: 0% and 100%. At a match value of 0%, all of the exclusion rule clauses must be absent in an unknown for a prediction to be made. This will not lead to optimal recall as some of the clauses in an exclusion rule may be present in an unknown if these features are not unique to this substructure. A substructure will be predicted to be absent at a match value of 100% if an unknown possesses all of the exclusion rule clauses. This match value, however, is useless since it is the *absence* of features which is indicative of the absence of substructures. At a match value of 99.9999% (used in Figure 4.3), *only one exclusion rule clause* must be missing in an unknown for a prediction to be made. Thus, the use of this match value for exclusion rule application maximizes recall while minimizing false positives.

Given a spectral feature which has a high level of correlation with a substructure, its absence from an unknown strongly suggests the absence of that substructure. Assume that another spectral feature has a lower level of correlation with the same substructure and thus does not

always appear whenever that substructure is present in a compound. The absence of such a feature in an unknown may not always indicate the absence of that substructure and thus may lead to false positives (which in this case is defined as an incorrect prediction of the absence of a substructure) since that feature does not have a correlation factor of 100% with respect to that substructure. As the C value decreases, more features are allowed into each exclusion rule, effectively loosening the conditions necessary for an exclusion prediction to be made. Thus recall and false positives both increase as shown in Figure 4.3. Only by using a C value of 100% for exclusion rule generation are false positives minimized.

When the exclusion rules are generated at a C value of 100% and applied to the training set at match values of 99.9999%, recall is maximized while minimizing false positives, and the reliability factor is optimized, as shown in Figure 4.3. Since rules generated at this C value contain only those features which correlate perfectly with each substructure, the degree of match between a given rule and each compound in the training set that contains the substructure will always be 100%. Since only those cases where the degree of match is less than 100% will lead to an exclusion prediction, false positives will be 0% when the exclusion rules are applied against the training set. When a C value of 100% and a match value of 99.9999% are used, exclusion rules logically operate in the following manner:

IF NO (spectral feature a)

OR NO (spectral feature b)

...

...

...

THEN the X substructure is ABSENT

Since all of the features in an exclusion rule generated at a C value of 100% have correlation factors of 100%, they have equal value in predicting the absence of that substructure. As expected, weighting of clauses in exclusion rules did not improve their performance.

False positives may occur when the exclusion rules are applied to true unknowns. For example, the exclusion rule for the phthalate-ester substructure contains "line in daughter scan at m/z 149" as a clause. However, 1,2-benzene-dicarboxylic acid, di-phenyl ester does not produce this feature, most likely due to the steric hindrance involved in fragmenting this molecule to form this daughter ion. When the phthalate-ester exclusion rule is applied to this compound at match values greater than 94% (given that the phthalate-ester rule contains 15 clauses), a false positive results. This problem underlines the importance of accurately characterizing each substructure in the training set. As the number of training set compounds containing a given substructure grows, fewer features will have correlation factors of 100%, and thus the number of features in each exclusion rule will decrease. This results in lower recall, which is similar to the trend seen in Figure

4.3 in which higher C values lead to lower recall. However, only by using a C value of 100% will false positives be minimized.

Inclusion Rule Optimization

Adding intensity data to the inclusion rules increases their information content and results in features with higher uniqueness factors. Thus, intensity data was expected to improve inclusion rule performance. Figure 4.4 plots recall as a function of the match value when the inclusion rules (generated at a U value of 100%) are applied to the training set with and without the use of intensity data. False positives and the reliability factor were not plotted as these were 0% and 100% respectively at all match values for rules generated at this U value. More inclusion rules with a greater number of features per rule are produced using intensity data, since more features have the required minimum uniqueness factor. Thus, when intensity data is used, recall is substantially improved as shown in Figure 4.4.

Figure 4.5 shows plots of recall, false positives, and the reliability factor versus the match value when the inclusion rules (generated at a U value of 50%) are applied to the training set with and without the use of uniqueness weighting. At match values of less than 100%, recall is slightly reduced by using uniqueness weighting. The real value of uniqueness weighting is in reducing false positives. The use of this weighting function results in a further improvement in inclusion rule reliabilities.

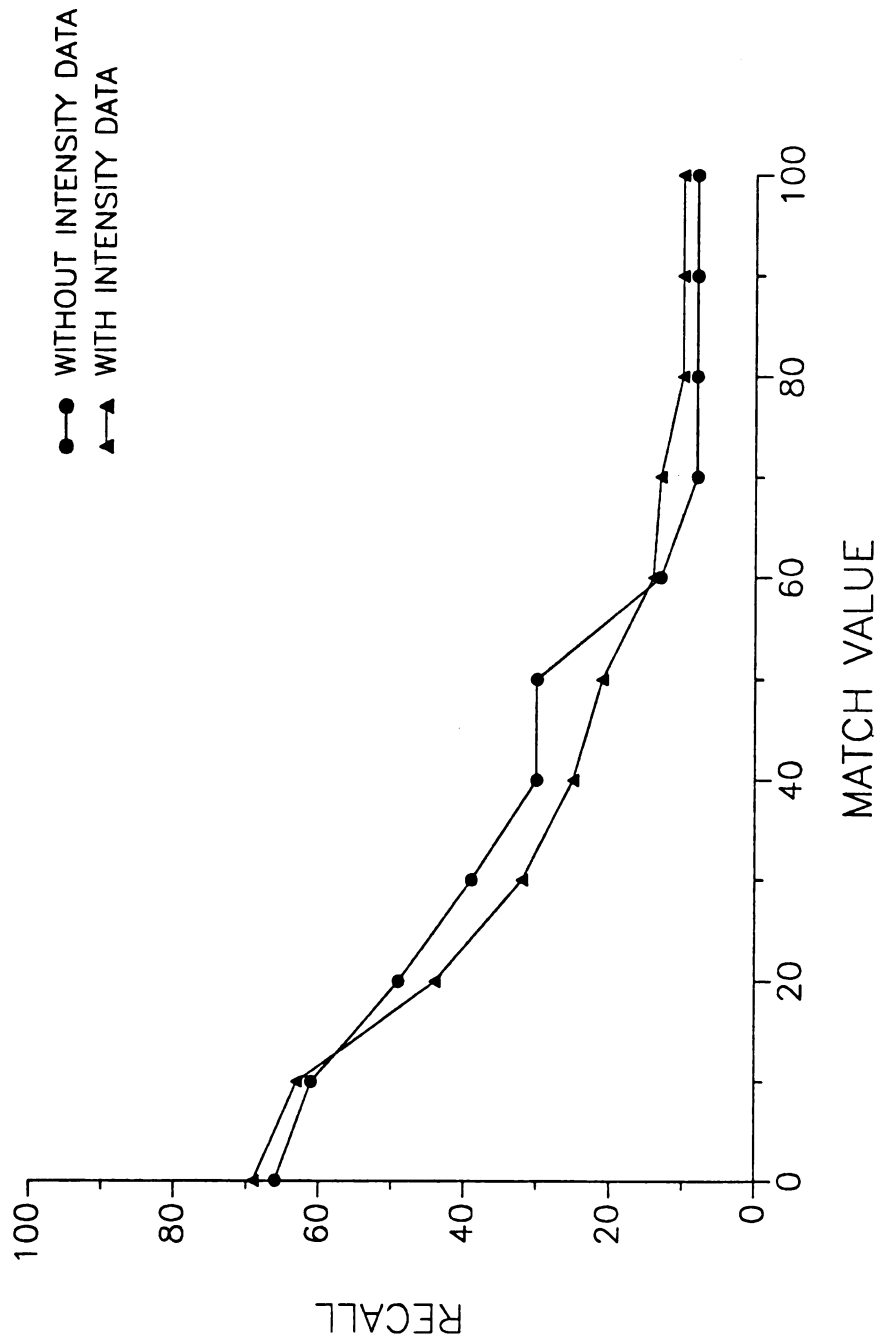


Figure 4.4 Recall versus match value for inclusion rules (generated at a U value of 100%) applied to the training set, with and without use of intensities.

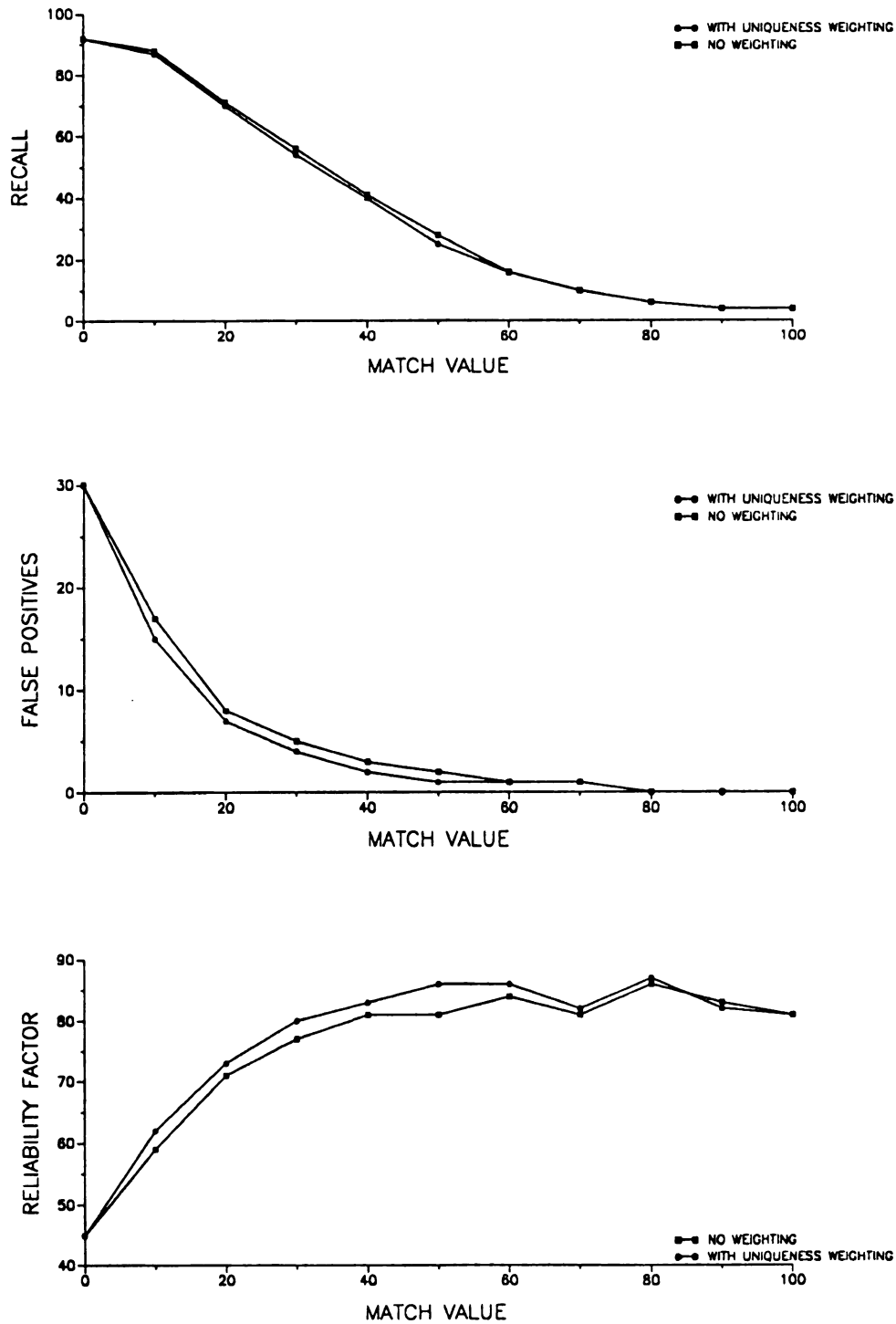


Figure 4.5 Recall, false positives, and the reliability factor versus match value for inclusion rules (generated at a U value of 50%) applied to the training set, with no weighting and uniqueness weighting.

Figure 4.6 shows plots of recall, false positives, and the reliability factor versus the match value when inclusion rules (using intensity data and uniqueness weighting) generated at three different U values are applied to the training set. This information is used to determine the optimal U and match values for inclusion rules. Since all of the features in an inclusion rule generated at a U value of 100% have uniqueness factors of 100%, they have equal value in predicting the absence of that substructure. Thus, uniqueness weighting has no effect on rule performance for rules generated at this U value. As the match value decreases, the number of rule clauses required to be present in an unknown for an inclusion prediction to be made decreases. Hence, lower match values essentially loosen the conditions necessary for an inclusion prediction to be made. This is demonstrated in Figure 4.6, where both recall and false positives increase as the match value decreases. The match value specifies the minimum degree of match necessary between an inclusion rule and an unknown for a prediction to be made. Thus, given an inclusion rule which contains 10 clauses, a match value of 90% would cause a prediction to occur if at least 9 rule clauses were present in the unknown. Again consider both ends of the extremes for the match value: 0% and 100%. At a match value of 100%, a prediction will be made if an unknown possesses all of an inclusion rule's features. Since many of these features are unique, they may not always be exhibited whenever that substructure is present in an unknown. At a match value of 0%, a prediction will be made if an unknown possesses none of the features in an inclusion rule. This match value is useless for predicting the presence of substructures. At a match value of 0.0001% (used in

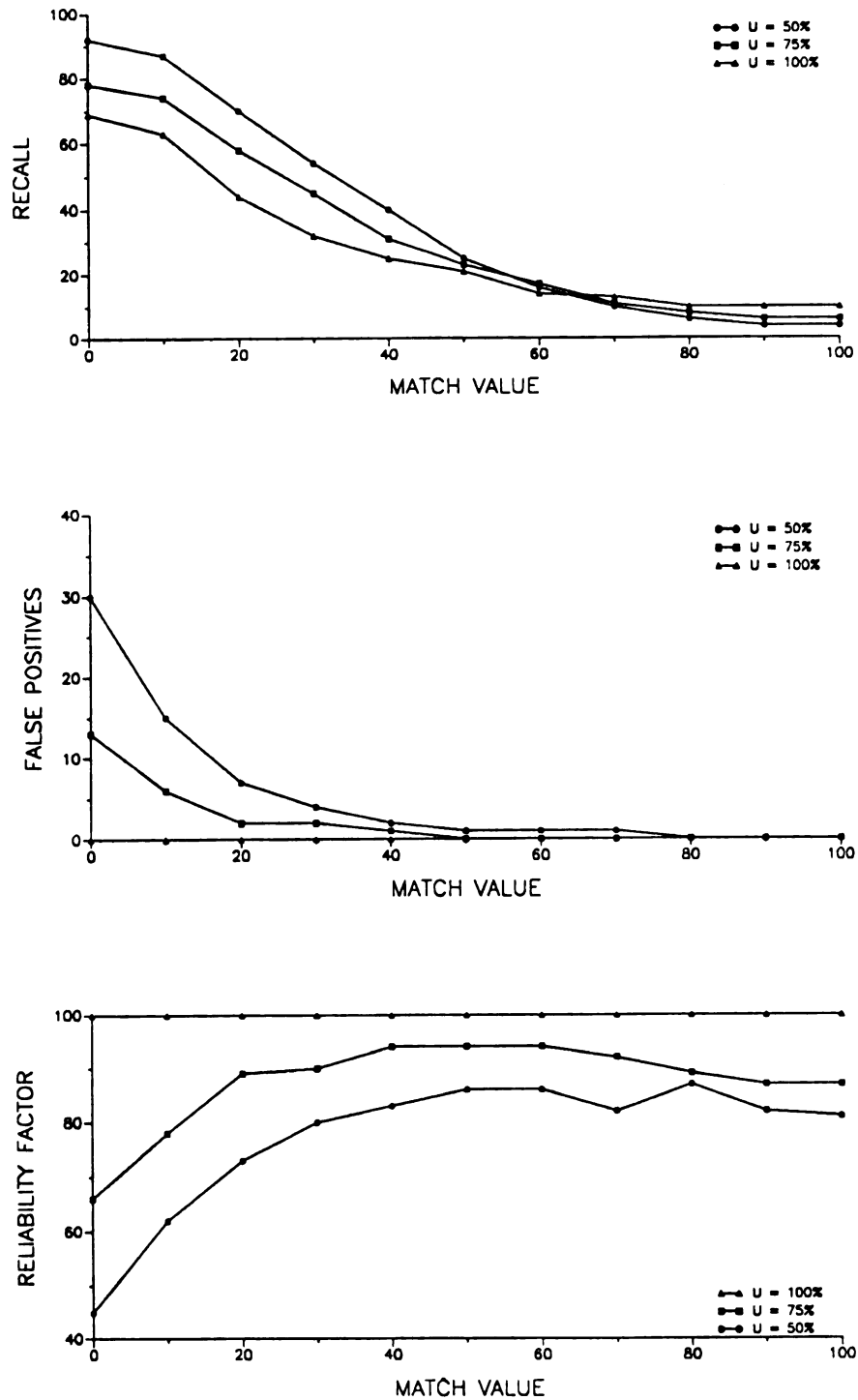


Figure 4.6 Recall, false positives, and the reliability factor versus match value for inclusion rules generated at three different U values applied to the training set.

Figure 4.6), only *one inclusion rule clause must be present in an unknown for a prediction to be made*. Thus, the use of this match value for inclusion rule application maximizes recall while minimizing false positives.

Given a spectral feature which is unique to a substructure, its presence in an unknown strongly suggests the presence of that substructure. Assume that another spectral feature is not unique to that same substructure, and thus may be indicative of other substructures. The presence of such a feature in an unknown may not always indicate the presence of that substructure and thus may lead to false positives. As the U value decreases, more features are allowed into each inclusion rule, effectively loosening the conditions necessary for an inclusion prediction to be made. Thus, recall and false positives both increase as shown in Figure 4.6. Only by using a U value of 100% for inclusion rule generation are false positives minimized.

When the inclusion rules are generated at a U value of 100% and applied to the training set at match values of 0.0001%, recall is maximized while minimizing false positives, and the reliability factor is optimized, as shown in Figure 4.6. Since rules generated at this U value contain only those features that are unique to each substructure, the degree of match between a given rule and any compound in the training set which does not contain the substructure will always be 0%. Since only those cases where the degree of match is greater than 0.0001% will lead to the prediction, false positives will be 0% when the inclusion rules are applied against the training set. When a U value of 100% and a

match value of 0.0001% are used, inclusion rules logically operate in the following manner:

IF (spectral feature a)

OR (spectral feature b)

...

...

...

THEN the X substructure is PRESENT

False positives may occur when the inclusion rules are applied to true unknowns. Features which had formerly been unique to each substructure based on the training set may be produced by the presence of other substructures in unknowns. For example, the inclusion rule for the bromo substructure contains a clause which represents a medium intensity neutral loss of 82 u, which is attributed to a loss of H^{81}Br . However, 1,3-benzene-dicarboxylic acid, di-allyl ester produces this feature, which can most likely be attributed to a loss of a $\text{C}_4\text{H}_2\text{O}_2$. When the bromo inclusion rule is applied to this compound at match values less than 33% (given that there are three clauses in this rule), a false positive results. This problem again underlines the importance of accurately characterizing each substructure in the training set. As the number of training set compounds containing a given substructure grows, fewer features will have uniqueness factors of 100%, and thus the number of features in each inclusion rule will decrease. This results in lower recall, which is similar to the trend seen in Figure 4.6, in which

higher U values lead to lower recall. However, only by using a U value of 100% will false positives be minimized.

The Substructure Hierarchy

A substructure hierarchy data structure has been implemented into MAPS. It is used for elimination of redundant substructures for more efficient generation of candidate structures, and can also be used to control the order in which rules are applied. The hierarchy defines the familial relationships between substructures and is best visualized in a tree-type format. Portions of this hierarchy are shown in Figures 4.7 and 4.8. Note that this hierarchy contains entries which are not *true* substructures. Examples of these include all-ss, org-ss, and o-ss, which represent all substructures, organic (carbon-containing) substructures, and oxygen-containing substructures. These are referred to as virtual substructures and are used to organize the hierarchy for more completeness. Substructures higher up in the hierarchy or near the "top" of the tree, such as methyl and all-ss, are usually either small or virtual substructures, whereas substructures near the "bottom" of the tree represent larger or more specific substructures. Inheritance relationships in this hierarchy are passed down the tree. For example, the carboxyl substructure contains the hydroxyl substructure as well as the virtual substructures o-ss and all-ss. Parts of this hierarchy which have been left out of Figures 4.7 and 4.8 due to lack of space are denoted by node links on the right edges of these figures. The substructure hierarchy data structure currently implemented in MAPS defines the

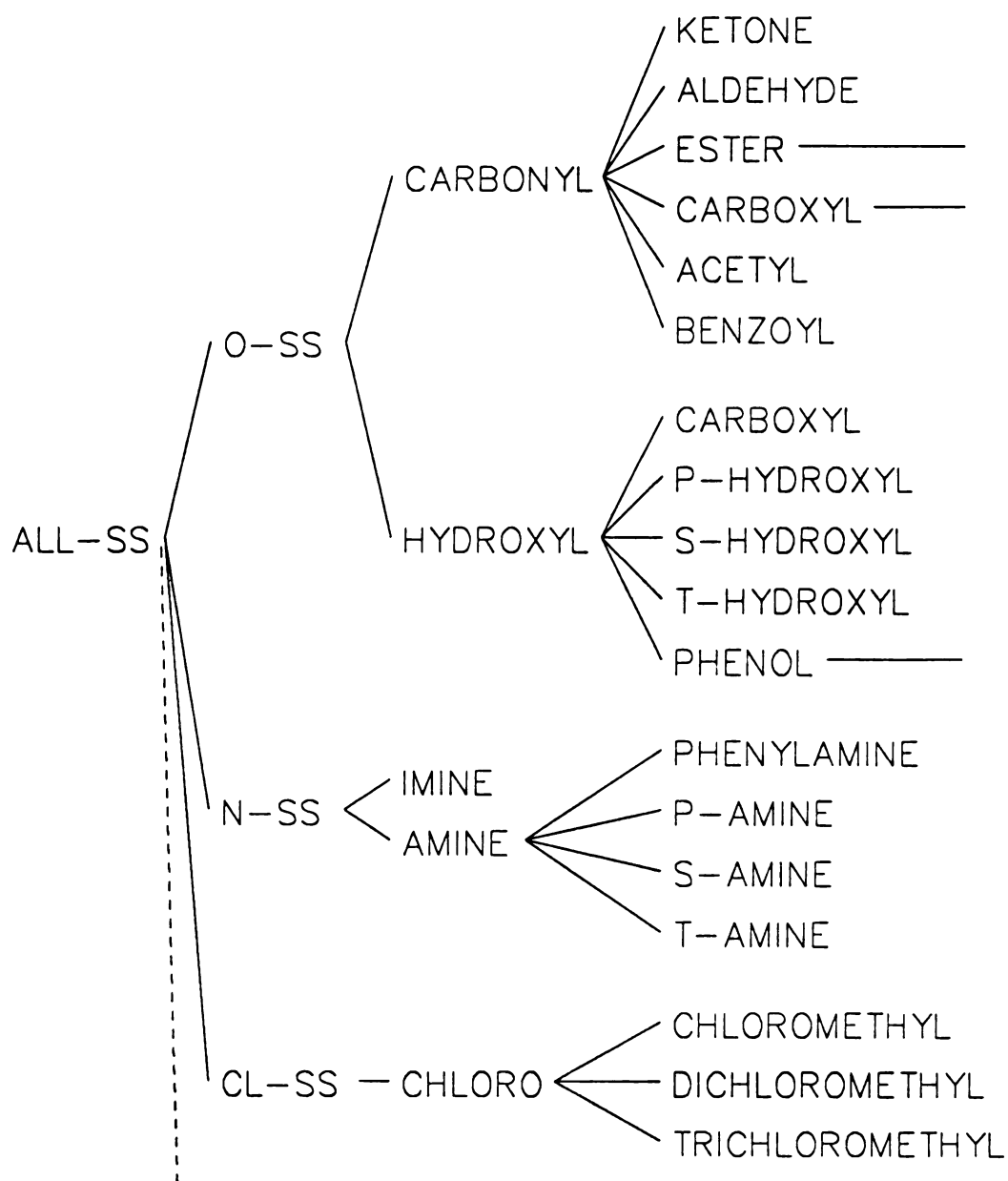


Figure 4.7 Portion of the substructure hierarchy defining inheritance relationships between substructures.

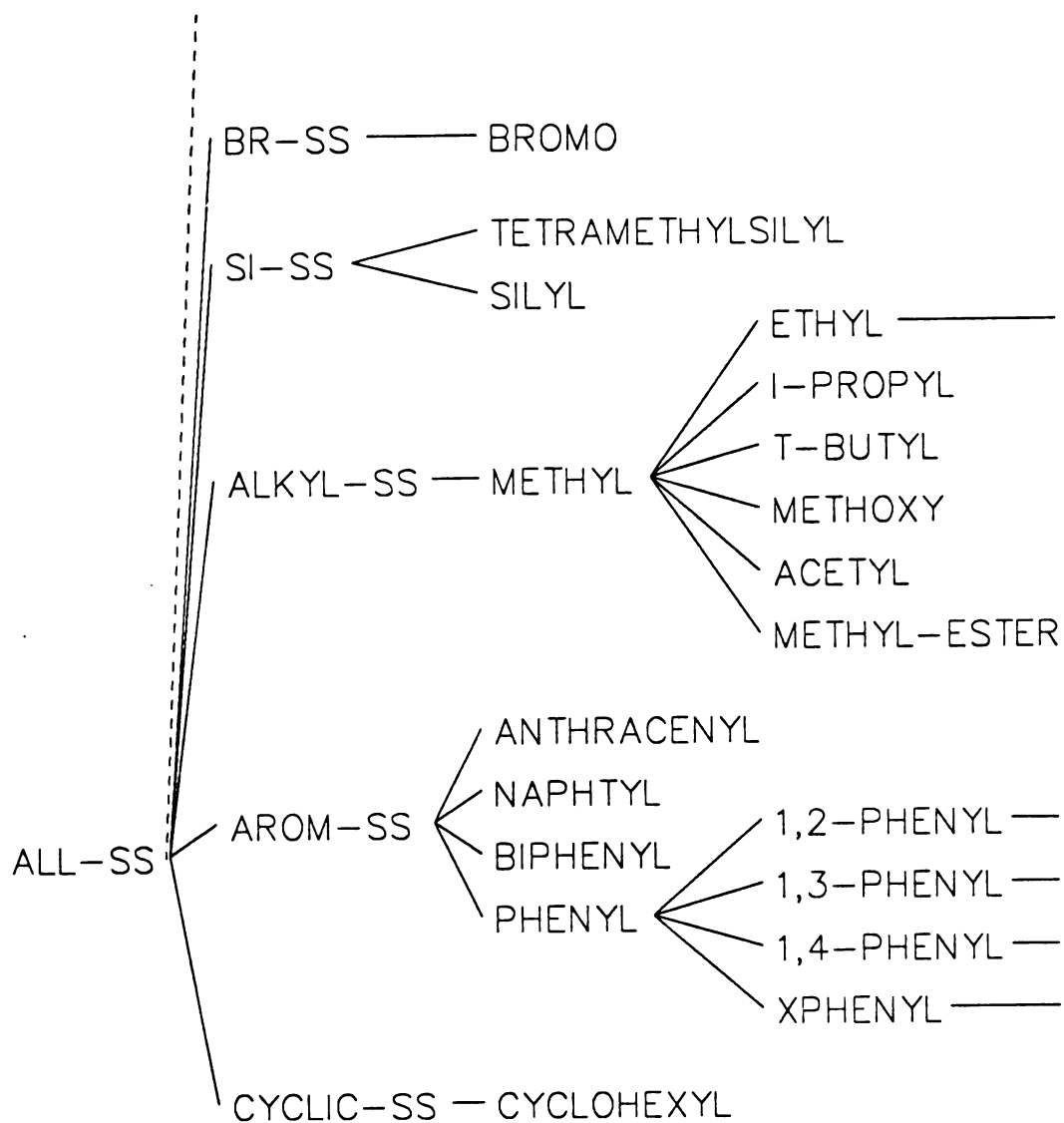


Figure 4.8 Portion of the substructure hierarchy defining inheritance relationships between substructures.

inheritance relationships between all known substructures currently used for rule development. Additional software has been developed to display the hierarchy on a terminal in a graphical, tree-type format to assist the user in extending or correcting the relationships.

This hierarchy finds use in the application of the rules to an unknown. MAPS applies all the rules against the set of MS and MS/MS data from an unknown to identify the presence or absence of as many substructures as possible. These data are used as constraints for GENOA. Although GENOA accepts redundant substructures as constraints, these increase the execution time for generation of candidate structures. Given the following list of substructures identified as present (methyl, ethyl, propyl, butyl, pentyl, carbonyl, carboxyl, ester, phthalate, phthalate-ester, xphenyl, 1-2-phenyl), the nonredundant substructures would be simply pentyl and phthalate-ester, as all other substructures are included within these. In the rule application process, the substructure hierarchy is used to eliminate such redundant substructures from the list of substructures identified as present or absent by MAPS for more efficient generation of candidate structures.

Conclusions

The modifications made to the MAPS code have identified several criteria for optimal rule performance and have substantially improved rule performance (1,2). When only those features which correlate perfectly with substructures are used to form exclusion rules, the absence of any one of these features indicates the absence of the

substructure. Using such criteria, 25 exclusion rules were generated which had reliability factors of 100% and recall of 67% when applied to the training set. Exclusion predictions were not possible with the previous version of MAPS. When only those features which are unique to each substructure are used to form inclusion rules, the presence of any one of these features indicates the presence of the substructure. Using these criteria and intensity data, 22 inclusion rules were generated which had an overall recall of 69% and no false positives when applied to the training set. Using the previous version of MAPS, only 13 reliable inclusion rules were obtained with an overall recall of only 12% and 0.2% false positives when applied to the training set. *Thus, the use of intensity data and optimal U and match values has greatly improved the predictive capabilities of the inclusion rules.* In addition, an unreliable rule list, which is obtained by monitoring rule performance against the training set in the previous version of MAPS, is no longer required. Since inclusion rules are now generated using a uniqueness filter rather than a correlation filter, rules are not obtained for substructures whose rules were formerly unreliable, since these substructures are not characterized by features with high uniqueness factors.

This approach, however, is not without its drawbacks. Optimization of the rules has decreased false positives at the expense of recall. As the training set grows to characterize more substructures in a variety of structural environments, fewer features will have correlation and uniqueness factors of 100% with respect to any given substructure. Thus, the rules produced at C and U values of 100% (for optimal performance) will contain fewer features. In addition, some

substructures may not be characterized by such features and thus no rules will be generated for them. In addition, the use of the optimal match values for inclusion and exclusion rules leads to predictions that are based on the presence or absence of at least one feature, which may lead to false positives when the rules are applied to unknowns. These problems have been addressed by a new approach for substructure identification described in the next chapter.

References

1. Palmer, P.T., Wade, A.P., Hart, K.J., Enke, C.G., presented at 13th Annual FACSS Conference, Detroit, MI, October 1987.
2. Palmer, P.T., Wade, A.P., Hart, K.J., Enke, C.G., in preparation for submission to Talanta.
3. Wade, A.P., Palmer, P.T., Hart, K.J., Enke, C.G., Anal. Chim. Acta, in press.
3. McLafferty, F.W., Stauffer, D.B., J. Chem. Inf. Comput. Sci. **25**, 245 (1985).

CHAPTER 5

AUTOMATED GENERATION OF MS AND MS/MS SPECTRAL FEATURE COMBINATIONS FOR RELIABLE SUBSTRUCTURE IDENTIFICATION

Introduction

In Chapter 4, the criteria for optimization of inclusion and exclusion rules are identified. These criteria, however, do have their limitations for substructure identification. These limitations are now described.

The use of a U value of 100% for inclusion rule generation minimizes false positives. However, many substructures cannot be characterized by MS and MS/MS features with uniqueness factors of 100% and such features become more unlikely as the training set grows. For example, the carbonyl substructure produces only one characteristic feature: a neutral loss of 28 u (loss of CO). However, a variety of other substructures can also produce this feature (due to a loss of N₂ or C₂H₄). Since the carbonyl substructure does not produce MS or MS/MS spectral features which have uniqueness factors of 100%, no rule can be generated for this substructure at a U value of 100%. The use of a lower U value allows more features into a rule and results in higher recall. However, only by using a U value of 100% are false positives minimized.

Use of a match value of 0.0001% for inclusion rule application maximizes recall. This match value causes a prediction to be made if *at least one feature* from the rule is present in MS and MS/MS data from an unknown. However, basing the identification of a substructure on the presence of a single feature can lead to false positives when the inclusion rules are applied to unknowns. Using a match value of 50% would cause a prediction to occur if an unknown produced at least one half of the features in an inclusion rule. Such higher match values would reduce false positives but lead to lower recall.

It is obvious that there is a tradeoff between recall and false positives for inclusion rules generated from a given training set. Recall can be improved by lowering the U value, but this increases false positives. False positives can be reduced by increasing the match value, but this decreases recall. Enhancing recall without causing a concomitant increase in false positives requires a completely new approach - one which is not limited to the use of features with uniqueness factors of 100% and does not require the use of a match variable.

As demonstrated in this chapter, the presence of a certain substructure in a compound is better indicated by specific *combinations of individual MS and MS/MS spectral features*. Such feature combinations have a uniqueness factor which is greater than or equal to the highest uniqueness factor of any of its component features. Thus, the use of such feature combinations for substructure identification will produce predictions which are inherently more reliable than those based on the presence of single features. Several different combinations of features

may be required to improve recall, since a given substructure may produce a variety of fragmentation patterns in different compounds due to the directing effects of surrounding substructures. These feature combinations can be used to formulate a new type of rule whose format is shown below.

```

IF (fa AND fb)
OR (fb AND fc)
OR (fc AND fd AND fe)
..          ..
..          ..
..          ..
THEN the X substructure is present

```

Such a rule can be comprised of feature combinations which have 0% false positives. Again, the rule should be composed of several feature combinations to improve recall.

This chapter addresses some of the deficiencies in the previous version of MAPS described in Chapter 4. The major goals of this work are improving the quality and predictive capabilities of the rules. Towards this end, two new feature types, multiple daughter ions and neutral losses from the same parent ion, have been implemented into the rule generation process to improve inclusion rule performance. Also, algorithms developed for automatic generation of feature combinations for substructure identification are described (1). Although much of the

work described in this chapter is applicable to exclusion rules as well as inclusion rules, exclusion rules are not discussed further.

Addition of Multiple Daughter Ion and Neutral Loss Feature Types to the Rule Generation Process

A symbolism for representation of the available scan modes for a BEQQ instrument developed by Cooks and coworkers (2) has been extended by Wade et. al. to identify the scan modes for multiple stage mass spectrometry (MS^n) (3). This symbolism uses a filled circle to represent a fixed mass and an open circle to represent a variable mass. Likewise, a heavy arrow is used to represent a fixed mass transition (i.e., a neutral loss or gain), and a thin arrow represents a variable mass transition. This symbolism is used here to show the feature types used in the rule generation and their relationships.

The feature types used in rule generation in the previous version of MAPS included lines in conventional spectra, lines in daughter spectra, neutral losses, and parent-to-daughter transitions. The neutral loss features are obtained from MS/MS spectra, which provide more definitive neutral loss data than MS spectra. Figure 5.1 shows the four feature types used in the previous version of MAPS using the Cooks symbolism (2). Each spectral feature type exploits specific patterns imbedded in MS/MS spectra. However, these four feature types do not take full advantage of all of the information available from MS/MS spectra. For example, the daughter spectrum of an ion containing a specific substructure may provide several neutral losses and/or daughter ions

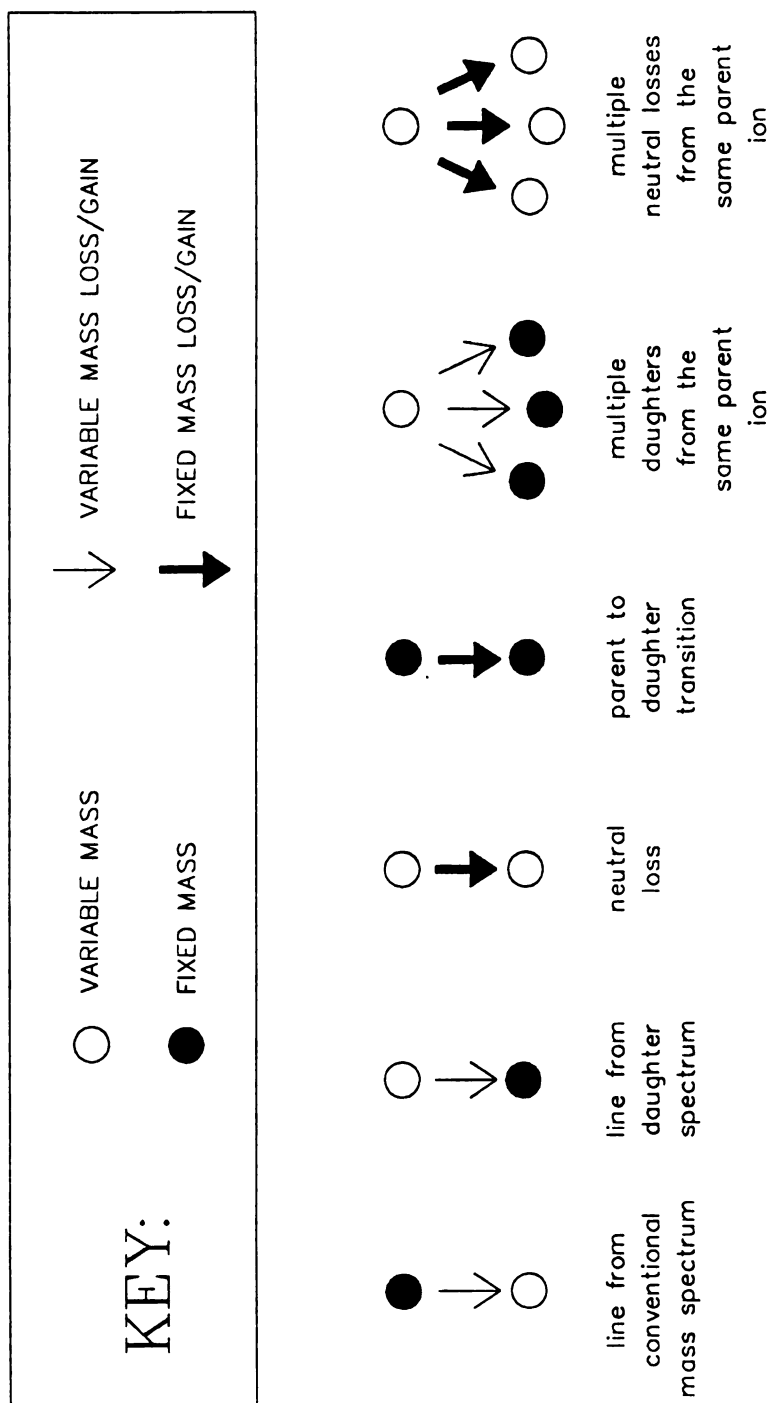


Figure 5.1 Feature types obtained from MS and MS/MS data for use in rule generation via MAPS.

characteristic of that substructure. For example, the daughter spectrum of m/z 135 from 4-*t*-butyl phenol (molecular weight of 150 u) shown in Figure 5.2 contains several daughter ions and neutral losses which are indicative of the substructures in the compound. The parent ion represents the loss of a methyl group from the molecular ion of this compound. The neutral losses of 15, 16, 28, 40, 44, 56, and 58 u, as well as the daughter ions at m/z 39, 41, and 55 can most likely be attributed to the *t*-butyl substructure, whereas the daughter ions at m/z 77, 79, and 91 are indicative of the benzyl substructure. It is evident that multiple daughter ions and neutral losses from the same parent ion can represent information that is characteristic of certain substructures. These new feature types, which are also shown in Figure 5.1 using the Cooks symbolism, have been implemented into the rule generation process. They represent *combinations of individual daughter ions and neutral losses with the added stipulation that these features must arise from the same parent ion*. Due to their higher dimensionality, multiple daughter ion and neutral loss features generally have higher uniqueness factors than single features.

Table 5.1 shows inclusion rules for the carbonyl, chloro, and ethyl substructures generated at a *C value* of 75%. The numbers in brackets for each clause represent the correlation and uniqueness factors for that feature. The intensity of each feature is represented by a s, m, or w letter next to the feature mass. These letters refer to intensity classes of strong, medium, and weak. The use of intensity data in the rules has been described in Chapter 4 and elsewhere (4). Although MAPS postulates candidate formulae for the fragment masses represented by

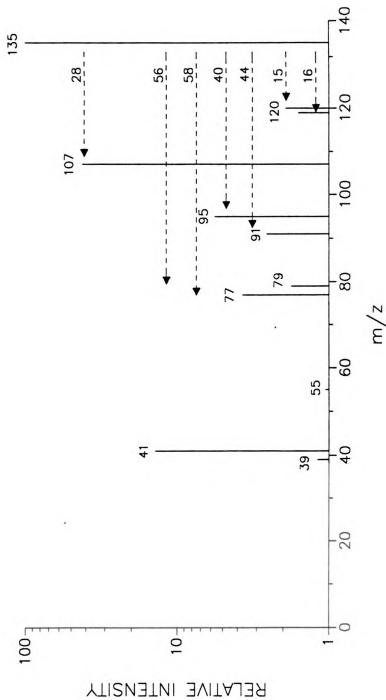


Figure 5.2 Daughter spectrum of m/z 135 from 4-t-butyl phenol.

Table 5.1 Inclusion rules for the carbonyl, chloro, and ethyl substructures generated at C = 75%.

IF [CF=91%,UF=33%] neutral loss of 28s amu
THEN the CARBONYL substructure is PRESENT

IF [CF=80%,UF=100%] neutral loss of 37m amu
OR [CF=80%,UF= 80%] neutral loss of 35m amu
THEN the CHLORO substructure is PRESENT

IF [CF=93%,UF=39%] neutral loss of 28s amu
THEN the ETHYL substructure is PRESENT

each clause in the rules, these are not shown here to save space. The rules shown in Table 5.1 contain only one or two clauses because few features correlate with these substructures at levels greater than 75%. Multiple daughter ion and neutral loss features do not appear in these rules since these feature types generally have low correlation factors.

Tables 5.2 and 5.3 show inclusion rules for the carbonyl and chloro substructures generated at a *U value* of 75%, respectively. The ethyl inclusion rule generated at U value of 75% is not shown since it contains 324 clauses compared to only 1 clause when this rule is generated at a C value of 75%. The use of a uniqueness filter rather than a correlation filter in the rule generation results in more clauses per rule. Note that the carbonyl and chloro rules mainly consist of multiple neutral loss features. The ethyl rule (not shown) contains 14 multiple daughter ion and 304 multiple neutral loss features. Since these features represent a combination of individual daughter ions and neutral losses from the same parent ion, they generally have higher uniqueness factors than the other feature types. Thus, these feature types dominate the rules when a uniqueness filter is used in the rule generation. Addition of multiple daughter ion and neutral loss feature types to the rule generation process more than doubles the number of clauses in each rule using a U value of 100%. The increase in the number of clauses through the use of these new feature types will be even greater at lower U values. Addition of these new feature types to the rules have improved their predictive capabilities.

Table 5.2 Inclusion rule for the carbonyl substructure generated at U = 75%.

```

IF [CF=26%,UF= 75%] neutral loss of 17s
OR [CF= 4%,UF=100%] parent giving neutral losses of 17m,28m
OR [CF= 4%,UF=100%] parent giving neutral losses of 17s,18m
OR [CF= 8%,UF=100%] parent giving neutral losses of 17s,28s
OR [CF= 4%,UF=100%] parent giving neutral losses of 17s,28s,
                        29s
OR [CF= 4%,UF=100%] parent giving neutral losses of 17s,29s
OR [CF= 4%,UF=100%] parent giving neutral losses of 18m,28m,
                        29m
OR [CF= 4%,UF=100%] parent giving neutral losses of 18m,28m,
                        29s
OR [CF= 4%,UF=100%] parent giving neutral losses of 18m,29m
OR [CF= 4%,UF=100%] parent giving neutral losses of 18m,29s
OR [CF= 4%,UF=100%] parent giving neutral losses of 18s,28s
OR [CF= 4%,UF=100%] parent giving neutral losses of 18s,29s
OR [CF=26%,UF= 75%] parent giving neutral losses of 28m,29m
OR [CF= 4%,UF=100%] parent giving neutral losses of 28m,29w
THEN the CARBONYL substructure is PRESENT

```

Table 5.3 Inclusion rule for the chloro substructure
generated at U = 75%.

```

IF [CF=80%,UF= 80%] neutral loss of 35m
OR [CF=40%,UF=100%] neutral loss of 35w
OR [CF=20%,UF=100%] neutral loss of 36w
OR [CF=80%,UF=100%] neutral loss of 37m
OR [CF=20%,UF=100%] neutral loss of 37s
OR [CF=20%,UF=100%] neutral loss of 37w
OR [CF=60%,UF= 75%] neutral loss of 38m
OR [CF=60%,UF=100%] neutral loss of 38s
OR [CF=40%,UF=100%] parent giving neutral losses of 35m,37m
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,36s
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,36s,
                                     37s
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,36s,
                                     37s, 38s
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,36s,
                                     38s
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,37s
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,37s,
                                     38s
OR [CF=20%,UF=100%] parent giving neutral losses of 35s,38s
OR [CF=20%,UF=100%] parent giving neutral losses of 35w,36m
OR [CF=20%,UF=100%] parent giving neutral losses of 35w,37w
OR [CF=40%,UF=100%] parent giving neutral losses of 36m,38m
OR [CF=20%,UF=100%] parent giving neutral losses of 36s,37s
OR [CF=20%,UF=100%] parent giving neutral losses of 36s,37s,
                                     38s
OR [CF=40%,UF=100%] parent giving neutral losses of 36s,38s
OR [CF=20%,UF=100%] parent giving neutral losses of 36w,38w
OR [CF=20%,UF=100%] parent giving neutral losses of 37s,38s
THEN the CHLORO substructure is PRESENT

```

Search Techniques

Problem solving and search techniques are fundamental to many fields of artificial intelligence (5). The use of these techniques is required to identify feature combinations which satisfy certain predefined criteria, so a brief review is apropos. The complete representation of a problem area and all of its possible solutions is referred to as its state space. A node is a specific point in the state space. The start node represents the initial point of entry into the state space and a terminal node is one which ends a specific path. The goal of the search is a node which satisfies given predicates. There may be several, one, or no solution(s) or goals in a given state space. A search algorithm progresses through the state space in an attempt to locate the goal. The directed graph of the nodes visited in a search is called the path. Many problems are typified by a large number of possibilities at each point in the state space. For these problems, search time increases exponentially as a function of the number of possibilities. This phenomenon is referred to as a combinatorial explosion. A rather illuminating example of this concept is provided by the game of chess. It has been estimated that the number of different complete plays in an average game is on the order of 10^{120} (6,7).

Search techniques vary in the order in which they examine nodes in the state space and can be classified as blind or heuristic. Blind or brute-force search techniques choose an arbitrary path through the state space and use no domain-specific information to judge where the solution lies. There are two different blind search strategies: depth-first

and breadth-first. A depth-first search explores each path to its end and backtracks to consider other paths. A breadth-first search checks each node on a given level in the state space before going to the next level. Although the search itself may be long, a breadth-first search will always find the shortest solution first. For most applications, some heuristics or rules-of-thumb are required to limit the search in large state spaces and prevent a combinatorial explosion. There are several different heuristic search strategies, including best-first, minimax, alpha-beta pruning, and hill-climbing. A best-first or ordered search selects the most promising node to be the next node to examine. The "promise" of a node can be evaluated several different ways. One way is to estimate the distance from a node to the goal state. More commonly, some evaluation function is used. Proper use of heuristics is required for rapid searching through large state spaces.

Finding Reliable Feature Combinations

There are several points regarding the search for combinations of MS and MS/MS spectral features which have specific performance characteristics which make this an interesting problem. Many applications which use search techniques require finding a path to the goal and thus involve identifying a *single solution*. However, several different combinations of MS and MS/MS spectral features may be required to identify a given substructure in various compounds since many different fragmentation patterns may be produced by the substructure in a variety of structural environments. Thus, it is not

sufficient to find a single solution in the state space of all possible feature combinations for identifying a substructure. The search algorithm must find all feature combinations which satisfy the search criteria. Figure 5.3 is a nonredundant representation of the state space for a set of four features. Note that 15 unique feature combinations are possible from this set of four features. The various feature combinations have different performance figures for substructure identification. Combinations which have few features may be exhibited in a variety of compounds and thus may produce false positives. Combinations which have many features may not be exhibited in any compound and thus may have low recall. Given the large number of combinations which can be obtained from a given set of features, some criteria must be defined to limit the search to find only those combinations with the desired performance characteristics. For this work, the criteria that have been established are that *feature combinations must have a nonzero recall and zero false positives*. These criteria mean that *the search will find only those feature combinations which are unique to each substructure based on the training set and that no redundant combinations will be identified*. Appropriate heuristics are required for the search algorithm to avoid combinatorial explosions, direct the search, prune out combinations which do not satisfy the search criteria, and identify combinations which do satisfy the search criteria.

Figure 5.4 is a schematic of the methodology used for generating feature combinations. Inclusion rules from MAPS are used to provide feature sets from which the feature combinations are generated. Using a rule generated at a lower C or U value allows more features into the rule

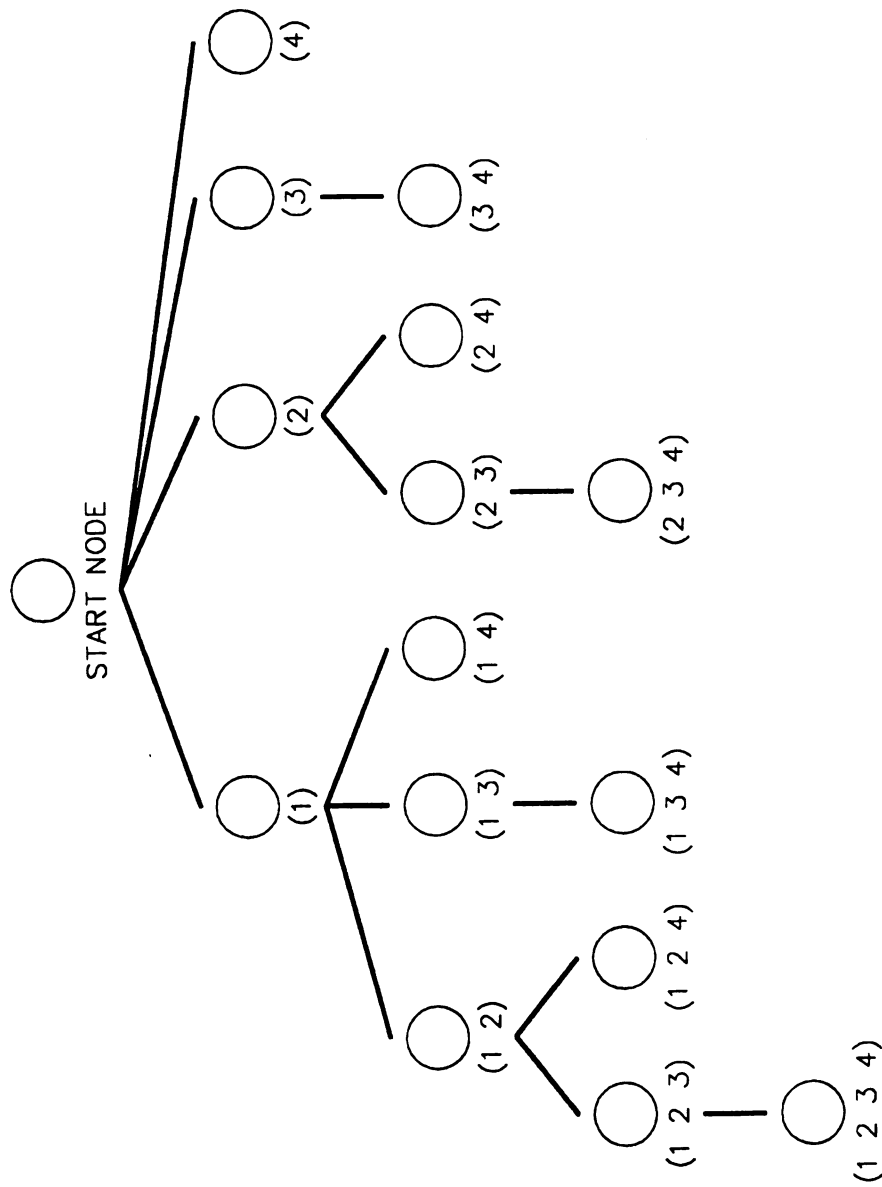


Figure 5.3 State space for a set of 4 features comprising 15 possible combinations of individual features.

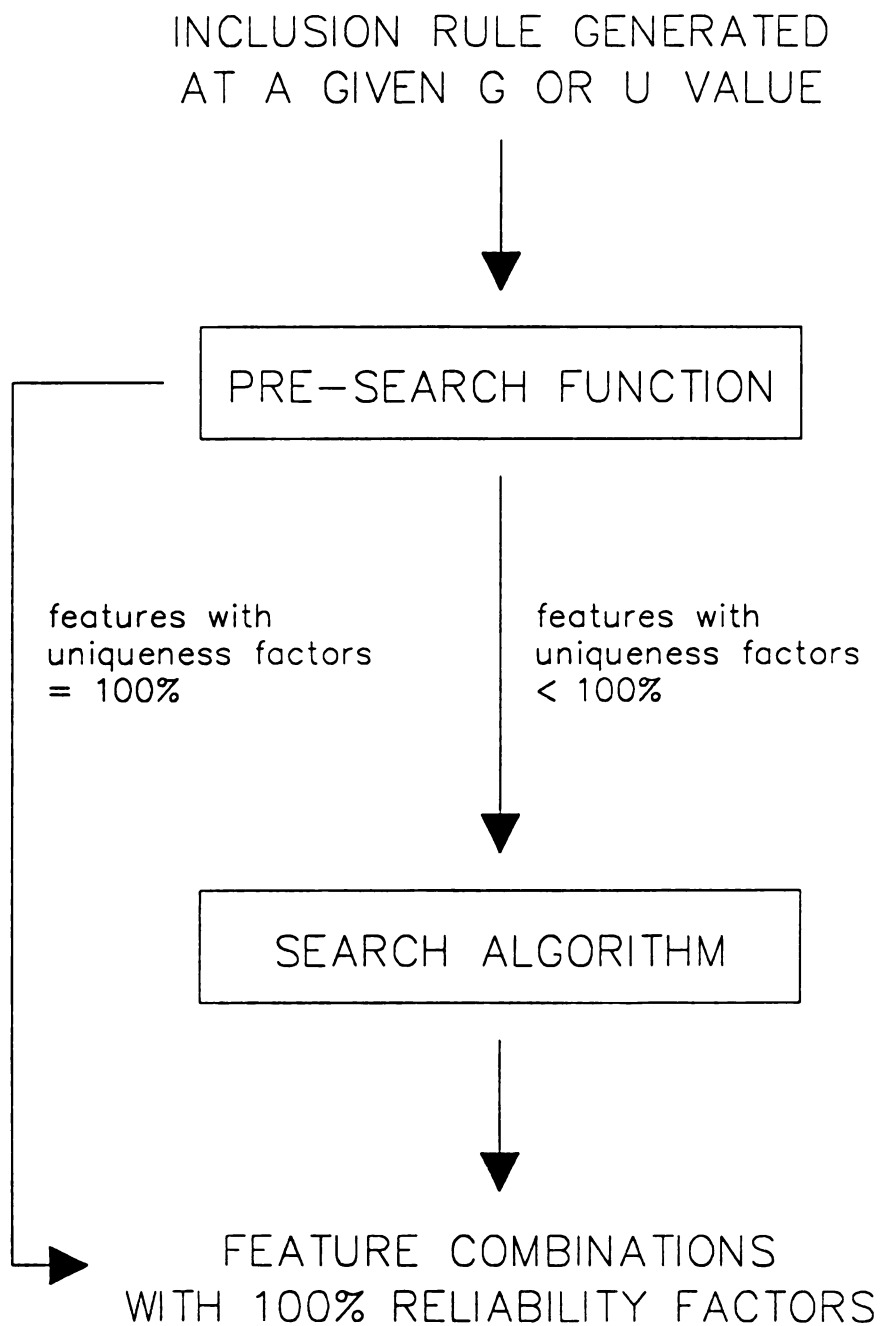


Figure 5.4 Schematic of procedure for identifying feature combinations for substructure identification.

and results in a larger state space and usually more feature combinations which have the desired performance characteristics. The pre-search function sets up parameters for the search and removes features which have uniqueness factors of 100% from the feature set. These features represent combinations that consist of single features which satisfy the search criteria and thus need not be used in the search. The search algorithm itself is a recursive function which performs a best-first, depth-first search. A breadth-first search was also developed but utilized more storage space in its execution and was correspondingly slower. The search algorithm operates on features with uniqueness factors less than 100%, since combinations of these may achieve the desired performance characteristics. The search algorithm generates new combinations by adding a feature onto the current combination and then calculates the performance figures (recall and false positives) for the new combination using training set data. Two situations may terminate a search down any path in the state space. These are nodes or feature combinations which have i) 0% false positives and nonzero recall or ii) 0% recall. When the first case is encountered, the combination is saved and the search is terminated since further searching down this path will lead to redundant combinations. When the second case is encountered, the search down this path is terminated since this combination will not identify any occurrences of the substructure in the training set. The search then backtracks to the next promising node. When the search turns up combinations which have nonzero recall and false positives, the search is allowed to continue to descendant nodes, since adding other features to that combination may

result in combinations which have the desired performance criteria. When the search continues from a given node, its successor nodes are examined in "best-first" order or increasing order of false positives. This guarantees that the most promising node will be examined first and also ensures that no redundant combinations will be identified.

Figure 5.5 is a partial representation of the state space for identifying the bromo substructure based on the inclusion rule for this substructure generated at a C value of 33% (shown in Table 5.4). The numbers in parentheses below the circles in this figure represent the numbers of the individual features from the bromo inclusion rule, and the other numbers refer to recall (RE) and false positives (FP). Many of the characteristics of the search algorithm are demonstrated in this figure. The solid circles represent combinations with nonzero recall and 0% false positives, the unfilled circles represent combinations which have zero recall, and the crosshatched circles represent combinations with nonzero recall and false positives. Note that only when a node has nonzero recall and false positives are its descendent nodes examined.

Once identified, these feature combinations may be employed independently of the training set for substructure identification. They may also be used to study fragmentation patterns of specific substructures in a variety of structural environments and instrumental operating conditions.

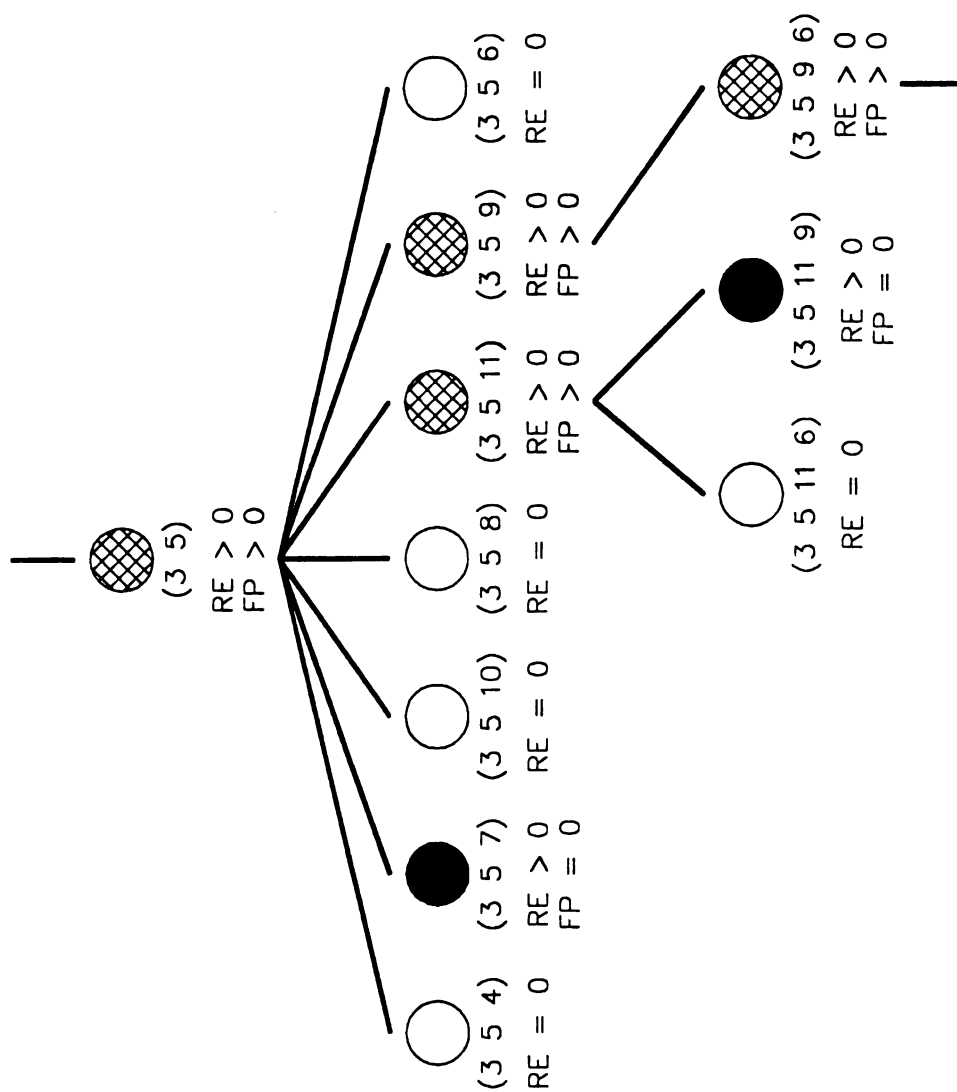


Figure 5.5 Portion of the state space for feature combinations indicative of the bromo substructure.

Results and Discussion

The number of nodes examined by the search algorithm has an upper limit of $2^n - 1$, where n is the number of features in the MAPS inclusion rule (not including the features with uniqueness factors of less than 100%, since these features are filtered out prior to the search). The actual number of nodes examined by the search algorithm is usually less than this number and depends on the uniqueness factors for each feature and the performance characteristics of the combinations themselves. The number of nodes in the state space divided by the number of nodes examined can be used as a measure of the efficiency of the search algorithm; this was typically 2.5 for large rules. This efficiency factor will naturally be unity when no combinations of 100% reliability are found, since all nodes must be examined to reach this conclusion.

The computation time required for finding feature combinations is directly proportional to the number of nodes examined by the search algorithm as it progresses through the state space. Figure 5.6 plots the computation time versus the number of nodes examined by the search algorithm. The data for each point were obtained by finding feature combinations for identifying the ethyl substructure using rules generated at different C values. The slope of this plot was used to determine the computation time required per node examined, which was found to be 31 msec/node. The feature combinations shown in this work are for mainly small substructures. There is good reason for this, as the rules for these substructures contain relatively few clauses. Since the rules for larger

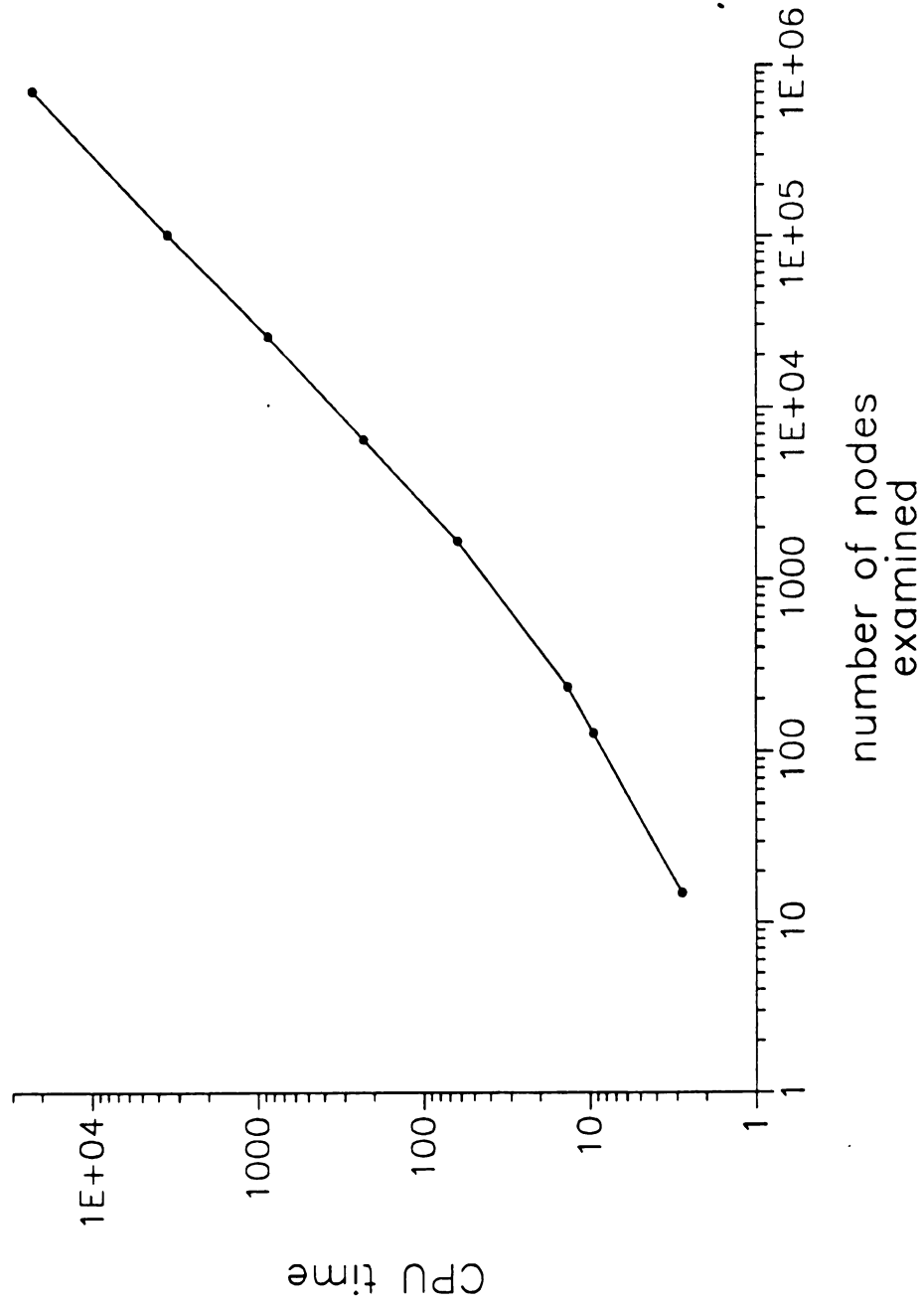


Figure 5.6 Plot of computation time versus the number of nodes examined by the search algorithm.

substructures contain many clauses, the computation times required for finding feature combinations for these substructures would be prohibitive on the Xerox 1108. Given a rule which contains 30 features and assuming that 50% of all possible nodes will be examined (approximately 537 million nodes), the computation time required would be 129 days!

The Xerox 1108 used in this work is limited by a small amount of memory (1.5 MBytes). Although the InterLISP-D software on this computer has many excellent features for facilitating applications development, these "extras" come at a substantial price in memory. Many LISP systems are notoriously memory intensive, and InterLISP-D is no exception. The current system uses 16 MBytes of virtual memory, 5 Mbytes of which are consumed by InterLISP-D alone! When the MAPS software and associated data structures are loaded, only 50% of the virtual memory space is free. The small amount of real memory compared to virtual memory often results in a large number of page faults during execution, which slows processing speeds down further. Obviously, further heuristics are required to reduce the number of nodes examined and faster computing resources are needed to reduce processing time. Parallel processors would provide an excellent solution to this problem since the search algorithm has imbedded parallelism.

The rule for the bromo substructure generated at a C value of 33% is shown along with the feature combinations derived from this rule in Table 5.4. This rule includes 12 features, 3 of which have uniqueness factors of 100%. Thus, the state space for this rule contains 2^{12-3} or 511 feature combinations. From this rule, 6 combinations which

Table 5.4 Inclusion rules for the bromo substructure
 a) generated at C = 33% and b) using feature combinations derived from this rule.

a) IF [CF=83%,UF=100%] neutral loss of 80s
 OR [CF=50%,UF=100%] neutral loss of 81s
 OR [CF=83%,UF= 50%] neutral loss of 82s
 OR [CF=50%,UF= 8%] parent ion at m/z 79m
 OR [CF=33%,UF= 18%] parent ion at m/z 79w
 OR [CF=33%,UF= 8%] parent ion at m/z 80m
 OR [CF=33%,UF= 11%] parent ion at m/z 80w
 OR [CF=50%,UF= 11%] parent ion at m/z 81m
 OR [CF=33%,UF= 14%] parent ion at m/z 81w
 OR [CF=50%,UF= 17%] parent ion at m/z 82m
 OR [CF=33%,UF= 7%] parent ion at m/z 82w
 OR [CF=50%,UF=100%] parent giving neutral losses of 81s,82s

THEN the BROMO substructure is present

b) IF neutral loss of 80s
 OR neutral loss of 81s
 OR parent giving neutral losses of 81s and 82s
 OR (neutral loss of 82s AND
 parent ion at m/z 79w AND
 parent ion at m/z 80w)
 OR (neutral loss of 82s AND
 parent ion at m/z 80m AND
 parent ion at m/z 81m)
 OR (parent ion at m/z 79w AND
 parent ion at m/z 80w AND
 parent ion at m/z 81w)

THEN the BROMO substructure is present

satisfied the search criteria were identified. Three of these are the single features with uniqueness factors of 100%. Even though the remaining features in the rule have uniqueness factors of less than 50%, taken collectively they produce three combinations which are unique to the bromo substructure. It is interesting to note that features with similar intensity classes are paired in these combinations. Note that one of the combinations represents weak intensity parent ion peaks at m/z 79, 80, and 81, while another combination represents medium intensity parent ions at m/z 80 and 81 and a strong intensity neutral loss of 82 u. This pairing of intensity classes may be caused by the different fragmentation efficiencies of the bromo-containing parent ions from different compounds.

The performance characteristics of the set of feature combinations identified by the search algorithm depend on the rule on which it operates. Tables 5.5 and 5.6 show the number of rule clauses, number of combinations identified, and overall recall for the chloro, bromo, hydroxyl, methyl, ethyl, and xphenyl substructures using inclusion rules generated at several C and U values, respectively. Recall is determined by applying the set of feature combinations identified for a given substructure against the training set. False positives will not occur due to the nature of the search criteria. Note that some of the entries in this table, such as those for the ethyl and xphenyl substructures at certain C values are not shown due to the large computation times required; these missing entries are denoted by asterisks.

As the C or U value decreases, more features are allowed into each rule, thus expanding the state space for feature combinations and

Table 5.5 Number of rule clauses, number of feature combinations identified, and overall recall for sets of feature combinations when applied to the training set using rules generated at several C values.

substructure	C value	number of rule clauses	number of combinations identified	overall recall
chloro	100	0	0	0
	75	2	1	80
	50	5	4	80
	33	5	4	80
bromo	100	0	0	0
	75	1	1	83
	50	7	3	83
	33	7	3	83
hydroxyl	100	0	0	0
	75	0	0	0
	50	4	2	57
	33	6	6	64
methyl	100	0	0	0
	75	0	0	0
	50	4	3	66
	33	8	8	85
ethyl	100	0	0	0
	75	1	0	0
	50	9	0	0
	33	28	*	*
xphenyl	100	0	0	0
	75	4	2	82
	50	26	*	*
	33	66	*	*

Table 5.6 Number of rule clauses, number of feature combinations identified, and overall recall for sets of feature combinations when applied to the training set using rules generated at several U values.

substructure	U value	number of rule clauses	number of combinations identified	overall recall
chloro	100	2	2	80
	75	4	3	80
	50	5	4	80
	33	5	4	80
bromo	100	3	3	83
	75	3	3	83
	50	4	3	83
	33	4	3	83
hydroxyl	100	5	5	43
	75	15	30	70
	50	19	85	75
	33	19	85	75
methyl	100	30	30	82
	75	38	59	87
	50	39	59	87
	33	39	59	87
ethyl	100	28	28	63
	75	49	130	67
	50	99	*	*
	33	*	*	*
xphenyl	100	78	78	100
	75	166	*	100
	50	225	*	100
	33	237	*	100

allowing the identification of more combinations with 100% reliability factors. The overall recall of a set of feature combinations will eventually approach a maximum at some C or U value. Some substructures obtain maximum recall from a combination which represents a single feature. For example, 80% of the chloro-containing compounds and 83% of the bromo-containing compounds in the training set are identified by a single diagnostic feature (medium intensity neutral loss of 37 u for the chloro substructure and strong intensity neutral loss of 80 u for the bromo substructure) using rules generated at a C value of 75%. The use of rules generated at lower C values and various U values for obtaining feature combinations does not improve recall for these two substructures. For rules generated at equal C and U values, recall is higher for feature combinations obtained from rules generated at a given U value. For example, note that recall for the methyl substructure is 79% using the rule generated at U = 100%, whereas recall is 0% using the rule generated at C = 100%. This is expected, as the use of a correlation filter in the rule generation results in rules which contain features which may not be unique. Note that the number of combinations is equal to the number of rule clauses when a U value of 100% is used. Since the features in such rules have uniqueness factors of 100%, each feature becomes a combination. The number of combinations identified is usually greater than or equal to the number of rule clauses using rules generated at any given U value. This is not the case for rules generated at any C value, since combinations of the features in such rules may not fulfill the search criteria.

Complete recall may not be achieved due to the fact that some compounds which contain a given substructure may not produce any of its characteristic features. Optimal recall seems to be achieved at some intermediate U value. Lowering the U value further will result in larger rules and more feature combinations, but these may not produce higher recall and the computation time required for identifying these combinations increases.

Conclusions

The use of combinations of individual MS and MS/MS spectral features for substructure identification has been found to *improve recall without an increase in false positives*; this was not possible using the version of MAPS described in Chapter 4. This methodology represents a completely new approach for substructure identification which *does not require the use of a match value*. The search algorithm produces combinations with a high overall recall when it uses sets of features from rules generated at an intermediate U value. The combinations generated by MAPS have reliabilities of 100% with respect to the training set and once identified may be applied independently of the MAPS code and training set.

There are several drawbacks to this approach. First of all, a substantial amount of computation time is required to identify these combinations for each substructure. The computation times involved in finding feature combinations for rule which contain many features are prohibitive. The use of more sophisticated heuristics in the search

algorithm and more modern computing resources will improve computation speeds. The search algorithm currently identifies only those combinations with 100% reliability. This was considered necessary to constrain the search and to produce predictions of high reliabilities for GENOA. However, predictions which have lower reliability factors may be desired and in many cases could be useful. Considering the almost unavoidable combinatorial explosion which results from identifying feature combinations from rules which contain more and more features and are derived from a continually expanding training set, more work and greater computing resources are needed to define the limits of this approach.

References

1. Palmer, P.T., Enke, C.G., in preparation for submission to Anal. Chem.
2. Louris, J.N., Wright, L.G., Cooks, R.G., Schoen, A.E., Anal. Chem., **57**, 2918 (1985).
3. Wade, A.P., Enke, C.G., Cooks, R.G., submitted to Int. J. Mass Spectrom. Ion Proc.
4. Palmer, P.T., Wade, A.P., Hart, K.J., Enke, C.G., in preparation for submission to Talanta.
5. Barr, A.B., Feigenbaum, E.A. (Eds), "The Handbook of Artificial Intelligence", Vol 1, Heuristech Press, Stanford, CA, 1981.
6. Shannon, C.E., Philosophical Magazine, **41**, 256 (1950).
7. Shannon, C.E., in Newman, J.R. (Ed.), "The World of Mathematics", Volume 4, Simon and Schuster, NY, 1956.

CHAPTER 6
PROGRAMS FOR MOLECULAR FORMULA DETERMINATION
EMPLOYING DATA FROM UNIT RESOLUTION MS/MS SPECTRA

Introduction

The molecular formula of an unknown compound is often a crucial piece of information in elucidating its structure. High resolution mass spectrometry allows measurement of the mass of a molecular ion with sufficient accuracy to define its molecular formula unambiguously. However, quadrupole mass spectrometers provide only unit mass resolution. Direct determination of molecular formulae cannot be accomplished from such unit resolution data, since a given nominal molecular weight may be consistent with many formulae.

Several groups have developed programs for determination of molecular formulae through the use of low resolution mass spectral data. The ELANAL program, developed by Kavanagh, generates all possible formulae corresponding to a given molecular weight and calculates the theoretical abundances of the isotopic and nonisotopic molecular ions for each formulae (1). Recently, Tenhosaari has developed a program which is similar in many respects, but differs in that it uses elemental composition data obtained from common fragment ions and assumed neutral losses to constrain the generation of candidate formulae (2).

Both methods use spectral matching to compare theoretical to experimental abundance patterns for ranking the list of candidate formulae. These methods require accurate intensity ratios and reasonable molecular ion abundances, which are often not possible for some compounds using electron impact (EI) ionization.

Several artificial intelligence and machine learning methodologies have been developed in this laboratory for automatic structure elucidation from unit resolution MS/MS data. Together, they form an integrated set of software tools called ACES (3) which is described in Chapter 1. One component of this system is an adaptation of GENOA, a constrained structure generator developed during the course of the DENDRAL project (4). An essential piece of information required by GENOA for structure generation is the molecular formula of the unknown compound. Thus, molecular formulae are *essential* for structure elucidation by ACES. This chapter discusses the methodology and software tools that have been developed for molecular formulae determination using data from a tandem quadrupole mass spectrometry (TQMS) instrument (5).

Several types of information can be derived from TQMS data as shown in Figure 6.1. Unit molecular weight information is usually obtained from chemical ionization (CI) mass spectra. Constraints on the elemental composition can be obtained from isotopic ratios and identified ions, neutral losses, and substructures. Several software tools have been developed in this laboratory to derive composition information from TQMS data for molecular formula determination. Software tools have been developed to determine elemental composition information from MS

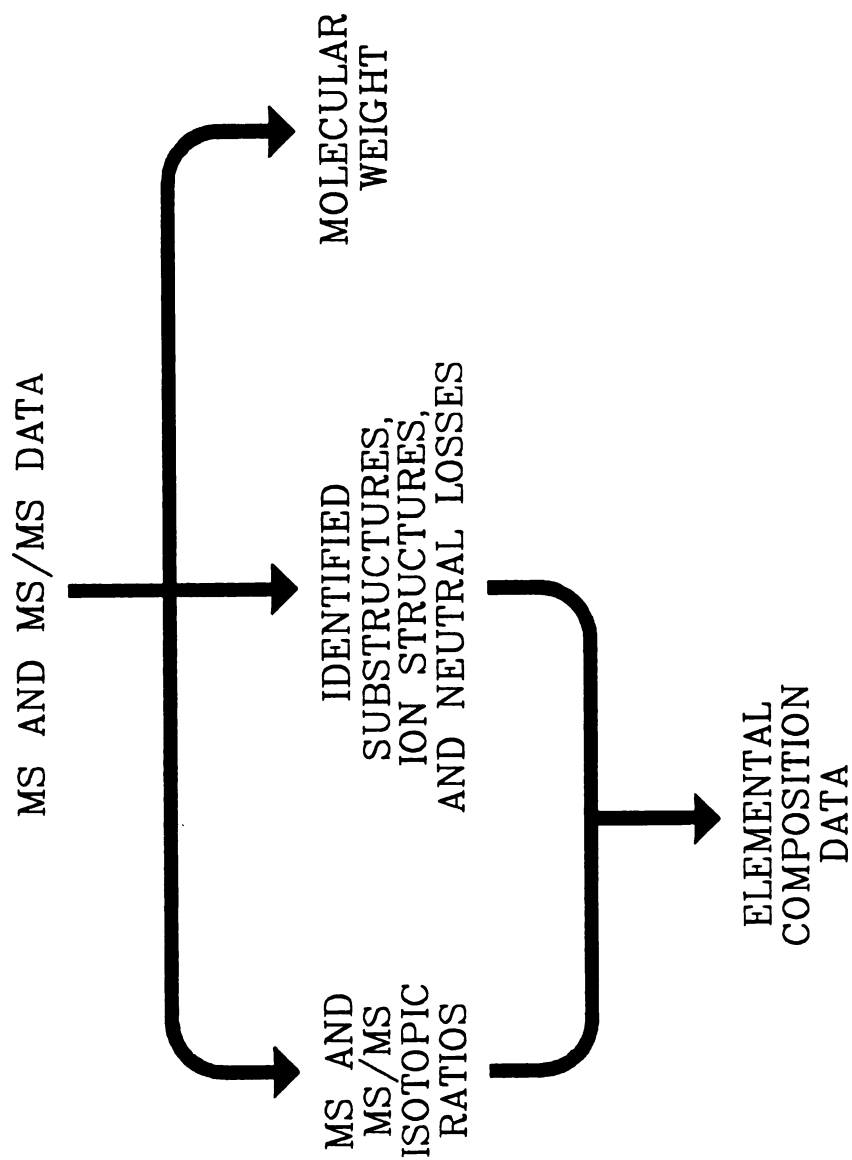


Figure 6.1 Information available from TQMS data.

and MS/MS isotopic ratio data. Additional software has been developed to identify substructures from MS and MS/MS data (MAPS) (6). Molecular weight information and composition data are entered as input to an molecular formula generator program, which identifies all formulae consistent with the constraints. This methodology for molecular formula determination is diagrammed in Figure 6.2 and differs from previous methods (1,2) in that MS/MS data is used to formulate elemental composition constraints prior to generation of candidate formulae, and spectral matching is not used to rank the list of candidate formulae. Examples of the great reduction in the number of candidate formulae through these software tools are provided later on in this chapter.

Experimental

All samples were obtained from a Chem Service compound kit and used without further purification in milligram quantities. They were introduced into an Extrel 400 series TQMS instrument on a solids probe. For optimum sensitivity, the abundance of the ^{13}C -containing molecular ion must be maximized. CI using methanol as a reagent gas was used to provide soft ionization with an abundant $(\text{M}+\text{H})^+$ ion for 1,2-benzene-dicarboxylic acid, di-n-octyl ester and 1,2-benzene-dicarboxylic acid, di-cyclohexyl ester (7). Methane CI was similarly used for eicosane. EI ionization can be used in obtaining these daughter spectra, but the sensitivity is reduced due to the lower abundance of the molecular ion. Argon was used as the collision gas at pressures of 1 mtorr or less to ensure single collision conditions. A 10 u region where the peak pairs

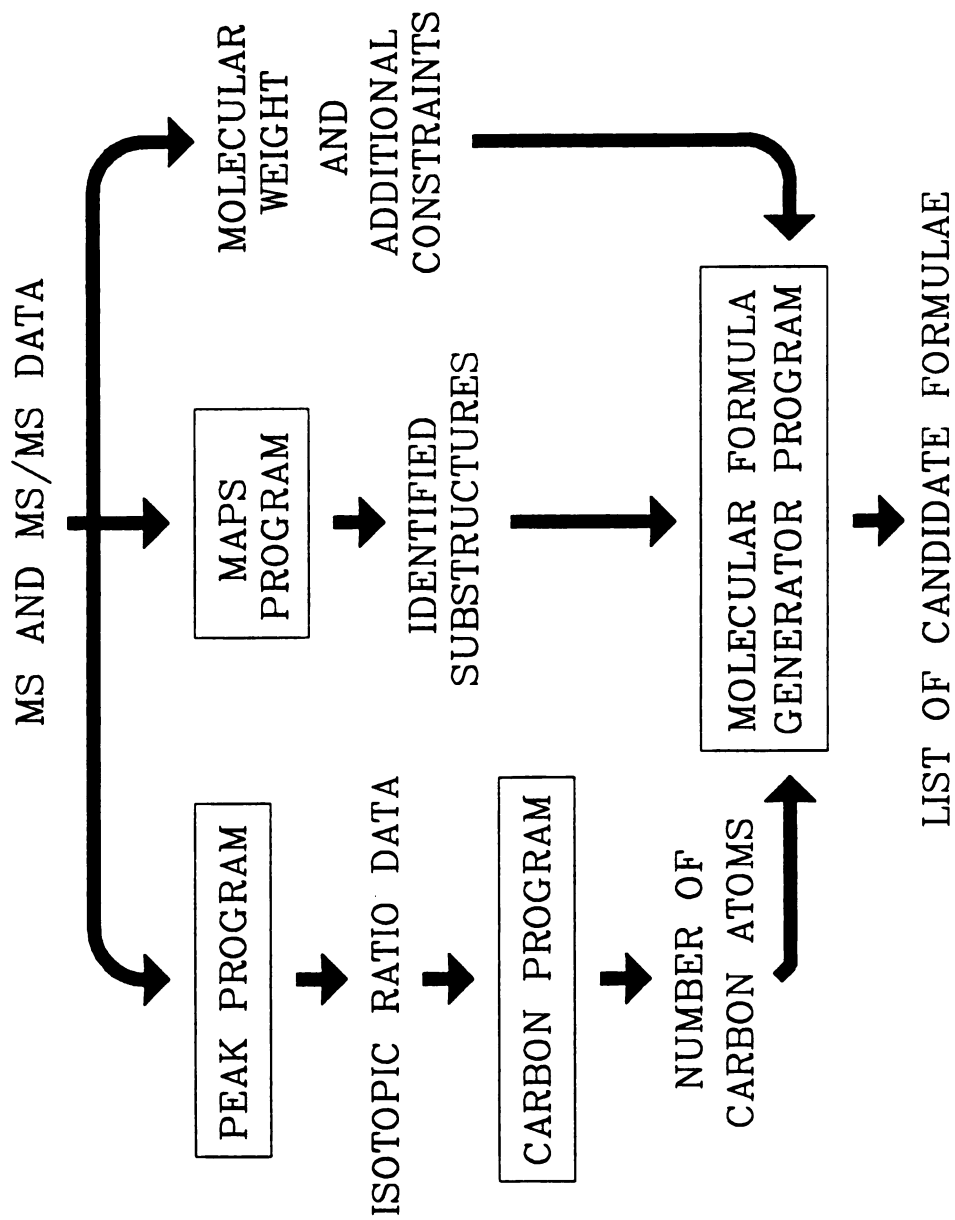


Figure 6.2 Schematic of molecular formula determination process.

appeared was swept with a resolution of 30 points/u. Peak area ratios were found to give better precision than peak height ratios. These ratios indicate the area of the peak at m/z m divided by the area of the peak at m/z $m+1$, and are averaged from 10 spectra. The daughter spectra of the m/z 149 ions were obtained using EI ionization. The PEAK, CARBON, and molecular formula generator programs shown in Figure 6.2 are written in FORTRAN-77 and are implemented on a MicroVAX II running the VMS operating system.

Algorithms

Applying Constraints to a Molecular Formula Generator. A program that adapts the standard "molecular formulae versus molecular weight" table has been developed. This program accepts constraints on the molecular weight, the precision to which it is known, the elements that are present, upper and lower limits on the numbers of these elements, and upper and lower limits on the degree of unsaturation of the compound. In addition, the program accepts input on the factor of the atom numbers of each element present, which is a constraint obtained from MS/MS isotopic ratios. Formulae can be generated for either neutral fragments or molecules, even-electron ions, or odd-electron ions.

Elemental composition data (elements present and the upper and lower limits on the numbers of these elements) are used to constrain the formula generation. The program exhaustively generates all nonredundant formulae that fall within the molecular weight window (molecular weight \pm precision). Formulae whose degree of

unsaturation are not within the specified rings plus double bond range are discarded. The nitrogen rule is applied to reject other implausible formulae. The program outputs the constraints used and the list of formulae consistent with these constraints. This list includes the mass and degree of unsaturation of each formula.

Determining the Number of Carbon Atoms. The presence of isotopic molecular ions in the conventional mass spectrum of a compound allows the partial elemental composition to be determined from the intensities and known natural abundances (1,8-12). Perhaps the most important constraint on the elemental composition of organic compounds is the number of carbon atoms. This can be estimated from the $M+1$ to M intensity ratio in conventional mass spectra. However, this measurement lacks precision due to the low natural abundance of ^{13}C . Also, inaccuracies may be caused by i) presence of another element in the compound, such as nitrogen, silicon, or sulfur, which has an isotope one mass unit greater than its most abundant isotope ($A+1$ element), ii) isotopic contributions to the $M+1$ intensity caused by the presence of an element, such as bromine or chlorine, in the $M-1$ ion which has an isotope two mass units greater than its most abundant isotope ($A+2$ element), iii) isotopic contributions to the M intensity caused by the presence of an $A+1$ element in the $M-1$ ion, and iv) variations in the natural abundance ratios.

The number of carbon atoms in a compound can also be determined from MS/MS data, or more specifically, from isotopic ratios from the daughter spectrum of a ^{13}C -containing molecular ion. When

used along with the molecular weight, this ratio data can yield a factor of the number of carbon atoms in a compound, or in some cases, the exact number of carbon atoms as will be demonstrated in several examples. This was first demonstrated by Bozorgzadeh and coworkers, who used a mass-analyzed ion kinetic energy spectrometry (MIKES) instrument to determine the number of carbon atoms in several compounds (13). This technique is not limited to fragmentation of the ^{13}C -containing molecular ion, and may be used to determine the elemental composition of any ion that contains elements with an isotope of sufficient natural abundance, such as chlorine and bromine (14-16). It can give better precision than MS molecular ion isotopic ratios because the ratios are generally closer to unity. Also, where several different ratios may be obtained from one daughter spectrum, these serve as confirmatory evidence. The principal limitation of this technique is sensitivity. The abundance of the ^{13}C -containing molecular ion is proportional to the number of carbon atoms in the compound. If the molecular ion is low and the number of carbon atoms is small, the abundance of the ^{13}C -containing molecular ion may be too small to generate reliable ratios in the daughter spectrum of this ion.

Our results demonstrate that the number of carbon atoms in a compound can be determined on a TQMS instrument through the use of conventional molecular ion isotopic ratio and daughter spectra isotopic ratio data. This instrument's advantage over the MIKES instrument for determining these MS/MS isotopic ratios is due to its ability to provide unit mass resolution in both analyzers. Due to the low natural abundance of ^{13}C , this determination of the number of carbon atoms is

certainly more difficult than for chlorine or bromine. When the results of this determination are employed by an molecular formula generator, a great reduction of the number of candidate formulae can be achieved.

The daughter spectrum of a molecular ion that contains one ^{13}C atom will give pairs of daughter ions at adjacent masses which represent either loss or retention of the ^{13}C atom in the fragment ion. The ratio of the relative intensities of these daughter ion pairs depends on the ratio of carbon atoms lost to the carbon atoms retained in forming that daughter ion. For the daughter spectrum of a molecular ion containing one ^{13}C atom, the ratio of peak heights at m/z m to $m+1$ is given by the relationship:

$$I_m / I_{m+1} = y / (x-y) \quad (1)$$

where I_m, I_{m+1} = intensities at m/z m and $m+1$, respectively
 x = total number of carbon atoms in the compound
 y = number of carbon atoms in the neutral fragment lost
 $(x-y)$ = number of carbon atoms in the daughter ion formed

Thus, measurement of these intensity ratios can often allow the number of carbon atoms in the parent ion, the daughter ion, and the neutral fragment to be deduced. Since the ratio I_m / I_{m+1} is a real number corresponding to the integer fraction $y/(x-y)$, a given intensity ratio can correspond to several integer fractions. For example, an intensity ratio of 0.5 can correspond to integer fractions of $1/2$, $2/4$, $3/6$, etc. Thus,

experimental ratios by themselves can be used only to estimate a *factor* of the number of carbon atoms present in a compound.

Determination of atom numbers by this technique assumes that the parent and daughter ions contain negligible isobaric interferences. Our study was limited to compounds containing only carbon, hydrogen, and oxygen. The contribution of ^2H and ^{17}O to the ^{13}C -containing molecular ion intensity was assumed to be negligible. The presence of other $A+1$ elements in a compound besides carbon in numbers greater than one complicate this determination. For example, the $M+1$ ion of a compound such as $\text{C}_8\text{H}_{22}\text{Si}_2$ may contain either the ^{13}C or ^{29}Si isotope. Thus, the daughter spectrum of this ion will be the linear superposition of the two daughter spectra which represent the two possible isotopic compositions of the parent ion. The number of carbon atoms may still be determined from this daughter spectrum if the presence and number of the interfering $A+1$ atoms are known. Bozorgzadeh has recently extended this methodology to determine the number of carbon, hydrogen, nitrogen, and oxygen atoms in parent and daughter ions using data from a MIKES instrument (17-19).

The number of carbon atoms in an unknown is determined as shown in the left side of Figure 6.2. The PEAK program analyzes raw intensity data from the daughter spectra of ions containing more than one isotope of an element, calculates baseline-corrected peak areas and heights, and determines isotopic ratios. The peak areas are obtained by a Simpson integration of the area underneath the peak profiles. The CARBON program calculates the number of carbon atoms in a compound from isotopic ratios in daughter spectra, ratios from

conventional mass spectra, and molecular weight information. First, the molecular weight is used to obtain an upper limit on the number of carbon atoms in the compound, which is x as shown in equation (1). Next, the program determines all integer fractions (y divided by $x-y$) that fall within a window specified by each experimental intensity ratio and its standard deviation. The estimate on the upper limit on the number of carbon atoms in the compound is used to constrain the number of integer fractions generated, since x cannot be greater than this limit. Factors of the number of carbon atoms in the compound are obtained by summing the numerator (y) and the denominator ($x-y$) for each fraction. The program calculates the largest common factor between the set of factors obtained from each experimental ratio, and uses this factor along with conventional isotopic ratios to determine the number or a factor of the number of carbon atoms in the compound. Naturally, larger common factors are more helpful for identifying the number of carbon atoms in a compound. Low factors can be attributed to the paucity of ion pairs in daughter spectra, specific magnitudes of intensity ratios (such as 0.5), and large standard deviations. Erroneous results may be obtained if any predicted ratio does not fall within the window of the experimental ratio plus or minus one standard deviation. The use of this methodology for determining the number of carbon atoms is demonstrated for three example compounds below.

Identifying Substructures, Ion Structures, and Neutral Losses.

Several types of structural information are available from MS and MS/MS data. Elemental composition data may be derived from these

data. Software has been developed to identify and characterize the neutral losses contained in daughter spectra (20). However, the identification of the composition of a neutral molecule from its mass alone becomes more uncertain as its mass increases. Likewise, the identification of an ion's structure becomes more difficult as its mass increases because of the larger number of possibilities for its composition. Certain characteristic ions in mass spectra are nevertheless highly indicative of particular structural features. Still less ambiguous information concerning the structure of specific ions can be obtained from their daughter spectra. Daughter spectra of isobaric parent ions can be used to differentiate between their corresponding substructures (21). The MAPS program, described in detail in Chapters 3 and 4, can be used to identify the presence of substructures in unknowns. In this context, the composition of the substructures identified by MAPS and from the masses of specific ions and neutral losses can be used as further constraints for the molecular formula generator.

Examples of Molecular Formula Determination

The molecular formula determination process occurs in two stages. First, the number of carbon atoms is determined and other constraints on the elemental composition are developed from MS and MS/MS data. The molecular formula generator is then applied using available constraints. These stages are now demonstrated for three examples.

Example 1 - 1,2-benzene-dicarboxylic acid, di-cyclohexyl ester. The daughter spectrum of the ^{13}C -containing protonated molecule from 1,2-benzene-dicarboxylic acid, di-cyclohexyl ester (mass 332) consists of daughters at m/z values of 149, 150, 247, and 248. Table 6.1 gives the isotopic peak area ratios of these daughter pairs. To determine a factor of the number of carbon atoms in this compound, the ratios from the daughter spectrum must first be converted to their nearest integer fractions. From the two ratios and equation (1), the following equations are obtained:

$$y_1 / (x - y_1) = 3 / 2$$

$$y_2 / (x - y_2) = 3 / 7$$

These may be rearranged to obtain:

$$y_1 = (3 / 5) x$$

$$y_2 = (3 / 10) x$$

Since y_1 and y_2 both must be integers, the number of carbon atoms in the compound, x , must be a multiple of 10. This constraint on the elemental composition substantially restricts the number of possibilities for the molecular formula of this compound. Given that the molecular weight of the compound is 330, the number of carbon atoms must be either 10 or 20.

Data from conventional isotopic ratios and the MS/MS technique mentioned here can be complementary. Data from the CI mass

Table 6.1 Comparison of experimental to predicted isotopic ratios for several compounds

COMPOUND	IONS	EXPERIMENTAL RATIO	STANDARD DEVIATION	PREDICTED RATIO
1,2-benzene- dicarboxylic acid, di-cyclohexyl ester	149/150 247/248	1.51 0.441	0.02 0.015	1.50 0.429
eicosane	71/72 85/86 99/100	3.04 2.32 1.89	0.15 0.09 0.16	3.00 2.33 1.86
1,2-benzene- dicarboxylic acid, di-n-octyl ester	149/150 261/262	1.99 0.502	0.03 0.009	2.00 0.500

spectrum of this compound showed the ratio of the intensities of M+1 to M to be 22.6%. This indicated that the number of carbon atoms in the compound was approximately 21, with an uncertainty of one or two carbons. Since the isotopic daughter spectrum had already limited the possible number to be a multiple of 10, the exact number of carbon atoms in this compound was found to be 20.

At this point, the molecular formula generator was invoked. The molecular weight of this compound was found to be 330 from its CI mass spectrum. Table 6.2 represents the partial output from the molecular formula generator for the case where no constraints were provided other than only atoms of carbon, hydrogen, and oxygen are present. In this case, 43 formulae were produced, some of which are highly unlikely, yet all obey the rules of valence. When the molecular formula generator was invoked again with the number of carbons specified as 20, only formulae 29-31 from Table 6.2 were produced. Additional information was supplied at this point to further reduce the number of candidate formulae. For example, the phthalate substructure was identified in this compound using the MAPS software, and thus two additional constraints were obtained: the number of oxygens must be at least four and the degree of unsaturation must be at least six. At this point, there are only two formulae from Table 6.2 consistent with all of these constraints: $C_{20}H_{26}O_4$ and $C_{20}H_{10}O_5$. From this example, it can be seen that when the number of carbon atoms in a compound are specified, the number of candidate formulae is substantially reduced. As more constraints are provided by MS/MS data, the number of candidate formulae is reduced even further. A unique molecular formula was not

Table 6.2 Molecular formula generator output for 1,2-benzene-dicarboxylic acid, di-cyclohexyl ester.

CONSTRAINTS:

SYMBOL	VALENCE	UPPER LIMIT	LOWER LIMIT	FACTOR	MASS
C	4	99	0	1	12.00000000
H	1	99	0	1	1.00782502
O	2	99	0	1	15.99491501

MASS = 330.00000000 +/- 0.500

LOWER LIMIT ON RINGS PLUS DOUBLE BONDS = 0.0

UPPER LIMIT ON RINGS PLUS DOUBLE BONDS = 20.0

NUMBER OF FORMULAE GENERATED = 43

CANDIDATE FORMULAE:

	MASS	DIFFERENCE	RDB	FORMULA
1)	329.919037	-0.08096	2.0	C ₂ H ₂ O ₁₉
2)	329.955414	-0.04459	1.0	C ₃ H ₆ O ₁₈
3)	329.991791	-0.00821	0.0	C ₄ H ₁₀ O ₁₇

8)	330.079834	0.07983	2.0	C ₁₀ H ₁₈ O ₁₂
9)	329.949554	-0.05045	10.0	C ₁₀ H ₂ O ₁₃
10)	330.116211	0.11621	1.0	C ₁₁ H ₂₂ O ₁₁

29)	330.313385	0.31339	0.0	C ₂₀ H ₄₂ O ₃
30)	330.183136	0.18314	8.0	C ₂₀ H ₂₆ O ₄
31)	330.052826	0.05283	16.0	C ₂₀ H ₁₀ O ₅

42)	330.104462	0.10446	19.0	C ₂₅ H ₁₄ O
43)	330.140839	0.14084	18.0	C ₂₆ H ₁₈

identified for this compound, but the number of possible formulae were reduced to two through the use of molecular weight data, MS and MS/MS isotopic ratios, and the identification of the phthalate substructure by the MAPS software.

Example 2 - Eicosane. The daughter spectrum of the ^{13}C -containing molecular ion of eicosane (mass 283) is shown in Figure 6.3. The low mass ion series appearing in this daughter spectrum provides several isotopic ratios. However, sensitivity limitations on this instrument limited the precise determination of isotopic ratio data to the most abundant ion pairs. Data for three such ion pairs are shown in Table 6.1. From these three ratios, equation (1), and converting the ratios to the nearest integer fractions, the following equations were derived:

$$y_1 = (3 / 4) x$$

$$y_2 = (7 / 10) x$$

$$y_3 = (13 / 20) x$$

It follows that the number of carbon atoms in the compound must be a multiple of 20. Since the molecular weight of this compound is 282, the actual number of carbon atoms in the compound must be 20, as no formulae with 40 or more carbons are consistent with this mass. Thus, in some cases, MS/MS ratio data used along with the molecular weight can give the exact number of carbon atoms in a compound without recourse to conventional M+1 to M intensity ratios or other data.

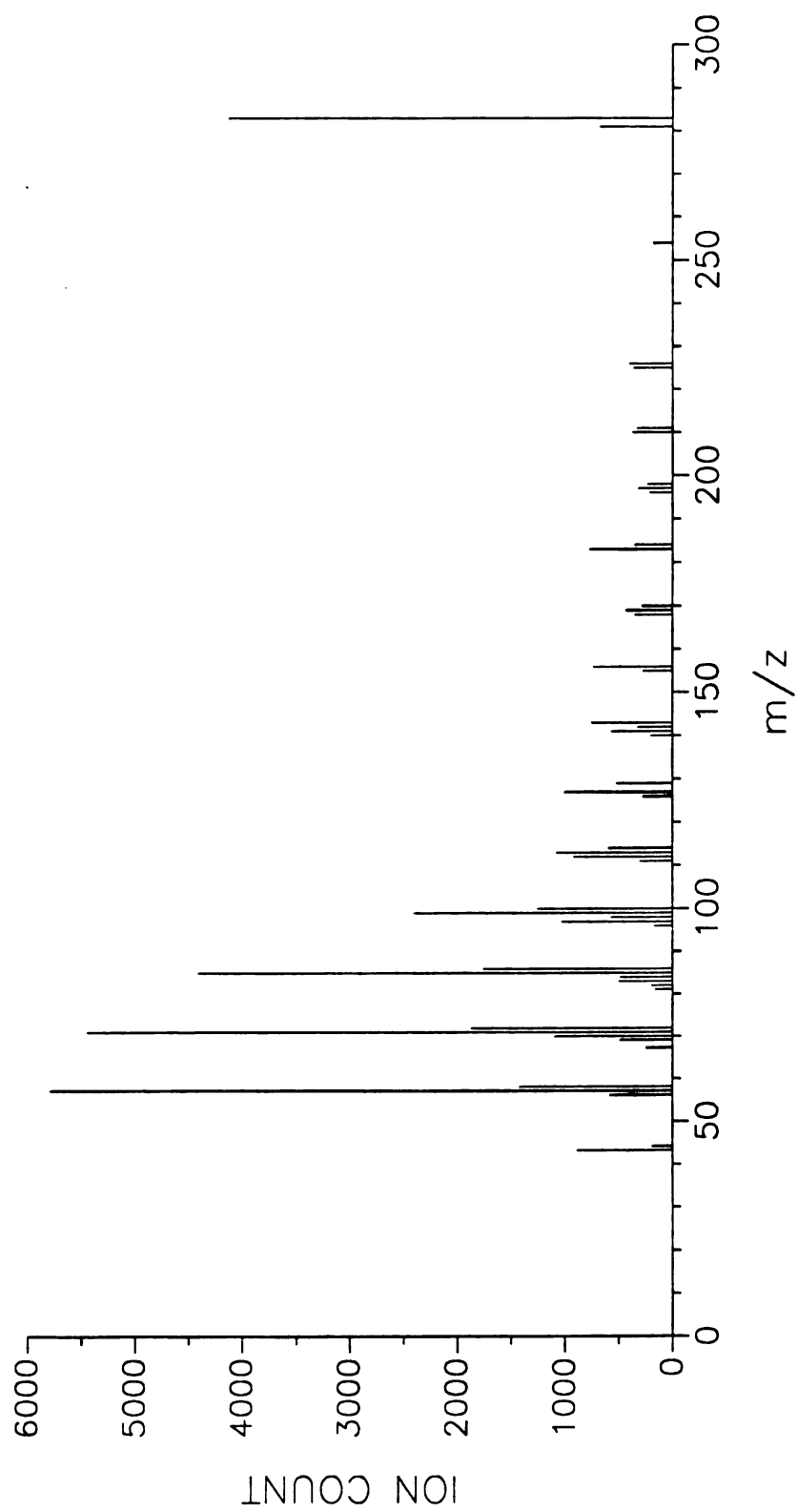


Figure 6.3 Daughter spectrum of the $(M+1)^+$ ion from eicosane.

Generally, it is advantageous to obtain as many ratios as possible from the daughter spectrum of an isotopic ion. For eicosane, five or more isotopic ratios could potentially be obtained from the daughter spectrum of the ^{13}C -containing molecular ion, sensitivity permitting. Multiple ratios serve to verify each other, or may increase the magnitude of the factor. The number of possibilities for the molecular formula of the compound decreases as the magnitude of this factor increases. In the worst case, daughter spectrum data would yield only that the number of carbons is a factor of two. This constraint rules out only molecular formulae which contain odd numbers of carbon atoms. However, if additional data were available, the ratio data may still be used to determine the exact number of carbon atoms in the compound. For example, if the molecular formula of one of the ions that appears in the daughter spectrum of the ^{13}C -containing was known, or if $M+1$ to M intensity ratios from conventional mass spectra are utilized, the exact number of carbon atoms can be determined.

Again, isotopic ratios from MS data can be used as confirmatory evidence for the number of carbon atoms. The CI mass spectrum of eicosane showed three ions in the molecular ion cluster, which indicated that two modes of ionization were competing; hydride abstraction and electron ionization. Thus, the ratio of the intensities of $M+1$ to M in the CI mass spectrum was useless as an indication of the number of carbon atoms in the compound. However, it was possible to calculate the corrected intensities of $M+1$ to M . A probability can be associated with both ionization modes, with the sum of the two probabilities equal to

one. The corrected intensities of M and M+1 can be calculated from the probabilities of each ionization mode and the intensities as shown below:

$$I_{m-1} = P * I'_m \quad (2)$$

$$I_m = ((1-P) * I'_m) + (P * I'_{m+1}) \quad (3)$$

$$I_{m+1} = (1-P) * I'_{m+1} \quad (4)$$

where P = probability of ionization occurring via hydride abstraction,

I'_m = corrected intensity for the molecular ion (M) assuming only one ionization mode,

I'_{m+1} = corrected intensity for the ^{13}C -containing molecular ion (M+1) assuming only one ionization mode, and

I_{m-1} , I_m , I_{m+1} = intensities at m/z m-1, m, and m+1, respectively

Software was written to calculate these corrected intensity ratios. Using the intensities at m/z's m-1, m, and m+1, and equations (2), (3), and (4) above, the corrected intensity ratio of m/z M to M+1 was found to be 19.7. This indicated that the number of carbon atoms in the compound was approximately 18, with an uncertainty of one or two carbons. This data supports the results obtained from the fragmentation of the ^{13}C -containing molecular ion, in which the number of carbons was determined to be 20.

The molecular weight of this compound was found to be 282 from the CI mass spectrum. When no constraints were provided other than only atoms of carbon, hydrogen, and oxygen are present, 34 formulae were produced. In Table 6.3, we see that when the constraint that the

number of carbon must be 20 is added, the number of candidate formulae are reduced from 34 to 3. The presence of the low mass ion series in Figure 6.3 indicates that this compound is most likely an alkyl compound, with a degree of unsaturation certainly less than 8. Thus, with this additional information, the exact molecular formula, $C_{20}H_{42}$, can be determined.

Example 3 - 1,2-benzene-dicarboxylic acid, di-n-octyl ester. The daughter spectrum of the ^{13}C -containing protonated molecule of 1,2-benzene-dicarboxylic acid, di-n-octyl ester (mass 392) contains the daughter pairs given in Table 6.1. Using the ratio data from Table 6.1 and the process already described, the following equations were derived:

$$y_1 = (2 / 3) x$$

$$y_2 = (1 / 3) x$$

The data thus constrain the number of carbon atoms in this compound to be a factor of three. In determining the molecular formula of this compound, this factor will not be as useful as a larger one. This low factor results from the paucity of daughters and the magnitude of the experimental ratios obtained from this compound. Thus, additional information was desired to further constrain the number of carbon atoms. The exact number of carbon atoms in this compound may be determined if the elemental composition of one of the daughter ions is known. In this example, the identity of the m/z 149 ion was determined by matching its daughter spectrum to a data base of m/z 149 daughter

Table 6.3 Molecular formula generator output for eicosane.

CONSTRAINTS:

SYMBOL	VALENCE	UPPER LIMIT	LOWER LIMIT	FACTOR	MASS
C	4	20	20	1	12.00000000
H	1	99	0	1	1.00782502
O	2	99	0	1	15.99491501

MASS = 282.00000000 +/- 0.500

LOWER LIMIT ON RINGS PLUS DOUBLE BONDS = 0.0

UPPER LIMIT ON RINGS PLUS DOUBLE BONDS = 20.0

NUMBER OF FORMULAE GENERATED = 3

CANDIDATE FORMULAE:

	MASS	DIFFERENCE	RDB	FORMULA
1)	282.328644	0.32864	0.0	C ₂₀ H ₄₂
2)	282.198364	0.19836	8.0	C ₂₀ H ₂₆ O
3)	282.068085	0.06808	16.0	C ₂₀ H ₁₀ O ₂

spectra which were obtained using EI ionization, some of which are shown in Figure 6.4 (21). This ion is the well-known phthalate anhydride ion, whose elemental composition is $C_8H_5O_3$. Using these data, only one ratio, that of m/z 149 to m/z 150, and equation (1) the following equation was derived:

$$y_1 / (x - y_1) = 2$$

Since $(x - y_1)$ represents the number of carbon atoms in the daughter ion formed, the quantity $(x - y_1) = 8$. Solving these two equations for x gives the number of carbon atoms in this compound, which is 24.

The exact number of carbon atoms in this compound can also be determined with additional data from conventional isotopic ratios as demonstrated above. Data from the conventional mass spectra of this compound showed the ratio of the relative intensities of $M+1$ to M to be 25.9. Thus the number of carbon atoms was approximately 24, with an uncertainty of one or two carbon atoms. These data along with MS/MS isotopic ratios, in which the number of carbon atoms was found to be factor of 3, are used to identify the exact number of carbon atoms in the compound, which was found to be 24.

The molecular weight of this compound was found to be 390 from the CI mass spectrum. The output of the molecular formula generator is given in Table 6.4 for the case where the number of carbon atoms was constrained to be 24. When no constraints were provided other than that only atoms of carbon, hydrogen, and oxygen are present, the program produced 54 formulae that were consistent with the molecular

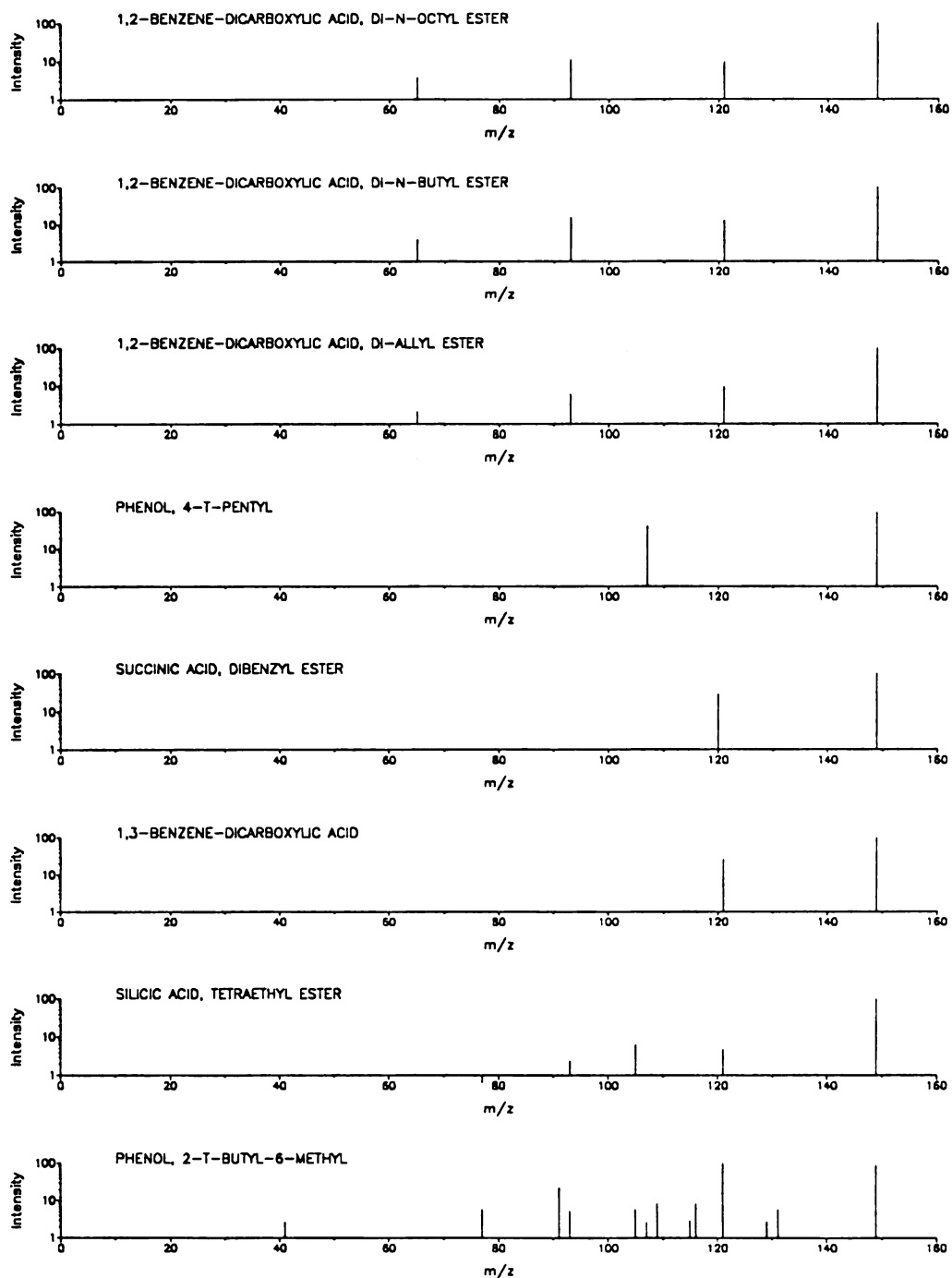


Figure 6.4 Selected daughter spectra of the m/z 149 ion from several different compounds.

Table 6.4 Molecular formula generator output for 1,2-benzene-dicarboxylic acid, di-n-octyl ester.

CONSTRAINTS:

SYMBOL	VALENCE	UPPER LIMIT	LOWER LIMIT	FACTOR	MASS
C	4	24	24	1	12.00000000
H	1	99	0	1	1.00782502
O	2	99	0	1	15.99491501

MASS = 390.00000000 +/- 0.500

LOWER LIMIT ON RINGS PLUS DOUBLE BONDS = 0.0

UPPER LIMIT ON RINGS PLUS DOUBLE BONDS = 20.0

NUMBER OF FORMULAE GENERATED = 2

CANDIDATE FORMULAE:

	MASS	DIFFERENCE	RDB	FORMULA
1)	390.277039	0.27704	6.0	C ₂₄ H ₃₈ O ₄
2)	390.146729	0.14673	14.0	C ₂₄ H ₂₂ O ₅

weight. When the number of carbon atoms was specified, only two formulae were produced.

Conclusions

MS/MS data can provide a substantial amount of information on the elemental composition of unknowns. The number of carbon atoms can be determined from MS and MS/MS isotopic ratios. Identified substructures allow additional specification on the elemental composition. Software has been developed to accomplish each of these tasks. Composition data are then provided to a molecular formula generator which generates a list of formulae consistent with these constraints. As clues to the elemental composition of a compound are determined, the list of candidate formulae becomes substantially reduced. An experienced mass spectroscopist can often use identified ion structures and neutral losses, conventional isotopic ratios, and other techniques to provide additional specification on the types and numbers of each element in the compound. With sufficient composition information, the list can often be reduced to a single formula.

References

1. Kavanaugh, P.E., Org. Mass Spectrom., **15**, 334 (1980).
2. Tenhosaari, A., Org. Mass Spectrom., **23**, 236 (1988).
3. Enke, C.G., Wade, A.P., Palmer, P.T., Hart, K.J., Anal. Chem., **59**, 1363A (1987).
4. Carhart, R.E., Smith, D.H., Gray, N.A.B., Nourse, J.G., Djerassi, C., J.A.C.S., **46**, 1708 (1981).
5. Palmer, P.T., Enke, C.G., Int. J. Mass Spectrom. Ion Proc., in press.
6. Wade, A.P., Palmer, P.T., Hart, K.J., Enke, C.G., Anal. Chim. Acta, in press.
7. Bauer, M.B., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1987.
8. Brauman, J.I., Anal. Chem., **38**, 607 (1966).
9. Boone, B., Mitchum, R.K., Scheppele, S.E., Int. J. Mass Spectrom. Ion Phys., **5**, 21 (1970).
10. Crawford, L.R., Int. J. Mass Spectrom. Ion Phys., **10**, 279 (1973).
11. Evans, J.E., Jurinski, N.B., Anal. Chem., **47**, 961 (1975).
12. McLafferty, F.W., "Interpretation of Mass Spectra", University Science Books, Mill Valley, CA, 1980.
13. Bozorgzadeh, M.H., Morgan, R.P., Beynon, J.H., Analyst, **103**, 613 (1978).
14. Todd, P.J., Barbalas, M.P., McLafferty, F.W., Org. Mass Spectrom., **17**, 79 (1982).
15. Tou, J.C., Anal. Chem., **55**, 367 (1983).
16. Singleton, K.E., Cooks, R.G., Wood, K.V., Anal. Chem., **55**, 762 (1983).
17. Bozorgzadeh, M.H., Lapp, R.L., 34th Annual Conference on Mass Spectrometry and Allied Topics, 1986, p. 428.
18. Bozorgzadeh, M.H., 35th Annual Conference on Mass Spectrometry and Allied Topics, 1987, p. 395.

19. Bozorgzadeh, M.H, Rapid Comm. Mass Spectrom., **2**, 61 (1988).
20. Gregg, H.R., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1986.
21. Cross, K.C., Palmer, P.T., Giordani, A.B., Beckner, C.F., Hoffman, P.A., Gregg, H.R., Enke, C.G., in "Artificial Intelligence Applications in Chemistry", Pierce, T.H., Hohne, B.A. (Eds.), ACS Symposium Series 306, American Chemical Society, Washington, DC, 1986, p. 321.

CHAPTER 7

APPLICATION OF ACES TO SEVERAL TEST COMPOUNDS

Introduction

The individual ACES tools (MAPS, MFG, and GENOA) have been applied to a set of test compounds to provide an objective evaluation of their performance. The registry name, IUPAC name, molecular formula, and molecular weight for each test compound are shown in Table 7.1. The structure of each test compound is shown in Figure 7.1. These test compounds represent a variety of structures, ranging from small, simple compounds such as m-cresol (molecular weight of 108 u) to larger, more complex compounds such as cinnamic acid, 3,5-di-t-butyl-4-hydroxy-, octadecyl ester (molecular weight of 530 u). One test compound contains a bromo atom, another contains several chlorine atoms, three contain a nitrogen atom, and the remaining test compounds contain only carbon, hydrogen, and oxygen atoms. Each test compound is hereafter referred to by its registry name.

MAPS Results

Chapter 5 describes a very reliable and powerful method for identifying the presence and absence of substructures through the use of

Table 7.1 Registry names, IUPAC names, molecular formulae, and molecular weights for 16 test compounds.

registry name	IUPAC name	molecular formula	molecular weight
AMPH4T	phenol, 4-t-aryl	C ₁₁ H ₁₆ O	164
BZTBOL	propanol, 2,2-dimethyl-3-phenyl	C ₁₁ H ₁₆ O	164
CS243	benzaldehyde, 4-methyl	C ₈ H ₈ O	120
CS244	benzaldehyde, 4-(dimethylamino)	C ₉ H ₁₁ NO	149
CS341	benzenamine, N,N-dimethyl	C ₈ H ₁₁ N	121
CS46	acetic acid, trichloro	C ₂ HCl ₃ O ₂	166
CS529	1,2-benzenedicarboxylic acid, di-n-octyl ester	C ₂₄ H ₃₈ O ₄	390
CS608	decane, 1-bromo	C ₁₀ H ₂₁ Br	220
CS86	1,3-benzenedicarboxylic acid	C ₈ H ₆ O ₄	166
CS871	phenol, 3-methyl	C ₇ H ₈ O	108
GMR1	butane, 2,2-bis-(4-hydroxyphenyl)	C ₁₆ H ₁₈ O ₂	242
GMR14	cinnamic acid, 3,5-di-t-butyl- 4-hydroxy, octadecyl ester	C ₃₅ H ₆₂ O ₃	530
GMR21	benzenamine, 4-hydroxy-N-octadecyl	C ₂₄ H ₄₃ NO	361
GMR27	benzaldehyde, 4-hydroxy-3-methoxy	C ₈ H ₈ O ₃	152
GMR8	benzene, 1,4-dihydroxy, monobenzyl ether	C ₁₃ H ₁₂ O ₂	200
TBMP24	phenol, 2-t-butyl-4-methyl	C ₁₁ H ₁₆ O	164

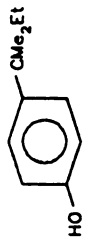
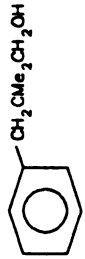


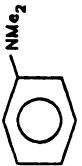
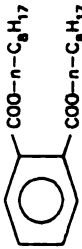


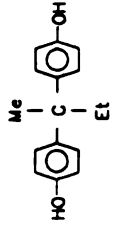
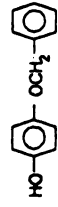
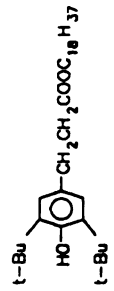



AMPH4T 	BZTBOL 	CS243 	CS244 
CS341 	CS46 CCl_3COOH	CS529 	CS608 $n\text{-C}_{10}\text{H}_{21}\text{Br}$
CS86 	CS871 	GMR1 	GMR8 
GMR14 	GMR21 	GMR27 	TBMP24 

Figure 7.1 Registry names and structures for the set of 16 test compounds.

combinations of MS and MS/MS spectral features. However, this method requires tremendous amounts of computing power and time for rule generation. Thus, it has not been used here for the analysis of the test compounds. Instead, the version of MAPS described in Chapter 4 were used to obtain the results shown here. The test compounds were left out of the training set prior to rule generation so that they could be treated as "unknowns". The inclusion rules were generated at a U value of 100% and the exclusion rules were generated at a C value of 100%. These rules were then applied to each test compound and the resulting predictions were categorized as correct or incorrect.

Table 7.2 summarizes the results from the application of the inclusion rules to each test compound. This table shows the number of recognized substructures present and the number of correct and incorrect inclusions, recall, and false positives for each test compound. An average of 4 substructures were correctly identified as present in each test compound. The number of incorrect inclusions ranged from 0 to 4 with an average of just less than one incorrect inclusion per test compound. Average recall was 53% and average false positives was 5%.

Table 7.3 summarizes the results from the application of the exclusion rules to each test compound. This table shows the number of recognized substructures absent, number of correct and incorrect exclusions, recall, and false positives for each test compound. An average of 9 substructures were correctly identified as absent from each test compound. There were no incorrect inclusions. Average recall was 25% and false positives was 0%.

Table 7.2 Summary of MAPS inclusion results for 16 test compounds.

registry name	# recognized substructures present	# correct inclusions	# incorrect inclusions	recall	false positives
AMPH4T	7	1	0	14	0
BZTBOL	5	4	1	80	4
CS243	5	2	0	40	0
CS244	6	2	1	33	4
CS341	4	3	4	75	16
CS46	3	1	1	33	4
CS529	15	10	4	67	29
CS608	11	8	0	73	0
CS86	5	2	0	40	0
CS871	5	3	0	60	0
GMR1	7	3	1	43	5
GMR14	19	8	0	42	0
GMR21	18	14	1	78	9
GMR27	6	4	0	67	0
GMR8	7	3	1	43	5
TBMP24	6	3	0	50	0

Table 7.3 Summary of MAPS exclusion results for 16 test compounds.

registry name	# recognized substructures absent	# correct exclusions	# incorrect exclusions	recall	false positives
AMPH4T	38	12	0	32	0
BZTBOL	40	13	0	32	0
CS243	39	14	0	36	0
CS244	37	9	0	24	0
CS341	40	12	0	30	0
CS46	41	15	0	37	0
CS529	29	10	0	35	0
CS608	34	8	0	24	0
CS86	39	9	0	23	0
CS871	39	11	0	28	0
GMR1	38	6	0	16	0
GMR14	21	4	0	19	0
GMR21	23	2	0	9	0
GMR27	37	5	0	14	0
GMR8	37	8	0	22	0
TBMP24	39	5	0	13	0

Table 7.4 shows the correct and incorrect inclusions for each test compound. Results from this table and Table 7.2 show that many of the substructures which are present in the test compounds were not identified as present by MAPS. Although the inclusion rules contain features which are unique to each substructure, these features will not always be exhibited in the MS and MS/MS data space whenever that substructure is present in a compound. Stated differently, features with uniqueness factors of 100% almost always have correlation factors which are much less than 100%. Hence, these substructures may not always be identified in unknowns and recall suffers. Recall can be improved by using combinations of MS and MS/MS spectral features for substructure identification as shown in Chapter 5.

Using this version of MAPS, an inclusion prediction is made if at least one rule clause is present in the MS and MS/MS data for an unknown. Many of the incorrect inclusions result from the presence of a particular feature, which previously had been found to be unique to a substructure, in a compound which does not contain that substructure. For example, parent-to-daughter transitions of m/z 67 to m/z 39 (medium intensity), m/z 55 to m/z 27 (strong intensity), and m/z 81 to m/z 41 (strong intensity) produced incorrect inclusions of the benzyl substructure for CS529, GMR1, and GMR14, respectively. A strong intensity parent ion at m/z 104 produced incorrect inclusions for the phthalate and phthalate-ester substructures for CS341. This type of false positive is due to the small size of the training set used. A medium intensity parent-to-daughter transition of m/z 69 to m/z 55 produces incorrect inclusions for the nonyl and decyl substructures for CS529.

Table 7.4 Correct and incorrect inclusions for 16 test compounds.

registry name	correct inclusions	incorrect inclusions
AMPH4T	xphenyl	
BZTBOL	methyl, phenyl, xphenyl, benzyl	phenol
CS243	xphenyl, benzyl	
CS244	methyl, xphenyl	phenol
CS341	methyl, phenyl, xphenyl	benzyl, benzoyl, phthalate, phthalate-ester
CS46	chloro	xphenyl
CS529	methyl, pentyl, hexyl, heptyl, octyl, ester, xphenyl, benzoyl, phthalate, phthalate-ester	nonyl, decyl, benzyl, phenol
CS608	methyl, pentyl, hexyl, heptyl, octyl, nonyl, decyl, bromo	
CS86	xphenyl, phthalate	
CS871	xphenyl, phenol, benzyl	
GMR1	methyl, xphenyl, phenol	benzyl
GMR14	methyl, heptyl, octyl, nonyl, decyl, xphenyl, phenol, benzyl	
GMR21	propyl, butyl, pentyl, hexyl, heptyl, octyl, nonyl, decyl, undecyl, dodecyl, xphenyl, phenol, phenylamine	benzyl
GMR27	methyl, xphenyl, phenol, benzyl	
GMR8	xphenyl, phenol, benzyl	methyl
TBMP24	methyl, xphenyl, phenol	

This type of false positive shows the limitations of basing the identification of a substructure on the presence of one feature. A medium intensity daughter ion at m/z 15 (appearing in a daughter scan of m/z 41) produced an incorrect inclusion for the methyl substructure for GMR8. This false positive may be due to contamination in the ion source or impurities in the sample as this daughter ion's formation from GMR8 is unlikely. Expansion of the training set for characterization of a wider variety of compounds and more accurate characterization of each substructure will produce more reliable rules and reduce false positives. Use of the methodology described in Chapter 5, which bases its predictions on the presence of combinations of features in unknowns, is expected to reduce false positives further.

The same arguments apply to exclusions. Recall will be improved and false positives will be decreased by using the methods described in Chapter 5, and again, a large and well-characterized training set is the best defense against false positives.

MFG Results

The MFG program, described in Chapter 6, uses molecular weight information and elemental composition data for generation of molecular formulae. EI and CI mass spectra were used to obtain molecular weight information for each test compound. Elemental composition data were derived from the list of substructures identified as present by MAPS in each test compound. Although additional composition data can be obtained from MS and MS/MS isotopic ratios, this has not been done for

these test compounds. The default constraints provided to the MFG program for these test compounds were that carbon, hydrogen, oxygen, and nitrogen were all possibly present. One exception to this was GMR14, for which only carbon, hydrogen, and oxygen were used for formula generation, since the addition of nitrogen leads to the generation of an extremely large number of formulae. Chlorine and bromine were used in formula generation only if these atoms were indicated as present in a test compound from the MAPS inclusion results. The limits on the degree of unsaturation for each test compound were 0 and 10. The tolerance of molecular weight data was specified as 0.5 u.

Table 7.5 shows the molecular formula, elemental composition data obtained from the inclusion results, and the number of formulae generated for each test compound. The number of formulae generated ranged from 1 to 35, with an average of 11 per compound. The correct formula will always be present in the list of generated formulae when correct elemental composition constraints are provided. Incorrect formulae result from insufficient or erroneous elemental composition data. The correct formula was present in the list of candidate formulae for 14 of the 16 test compounds. A single, correct molecular formula was identified for 3 of the 16 test compounds. The correct formulae were not present in the list of generated formulae for CS341 and CS46 due to erroneous composition data derived from incorrect inclusions. In some cases, incorrect inclusions do not exclude the correct formula from the list of formulae generated by the MFG program.

These data show the utility of MAPS results for molecular formula generation. MAPS results alone are usually insufficient for reducing the

Table 7.5 MFG results for 16 test compounds.

registry name	molecular formula	elemental composition constraints	# generated formulae
AMPH4T	C ₁₁ H ₁₆ O	C ₆	19
BZTBOL	C ₁₁ H ₁₆ O	C ₈ H ₅ O	5
CS243	C ₈ H ₈ O	C ₇ H ₂	4
CS244	C ₉ H ₁₁ NO	C ₇ H ₃ O	4
CS341	C ₈ H ₁₁ N	C ₁₁ H ₇ O ₄	0
CS46	C ₂ HCl ₃ O ₂	C ₆ Cl	12
CS529	C ₂₄ H ₃₈ O ₄	C ₂₀ H ₂₅ O ₄	13
CS608	C ₁₀ H ₂₁ Br	C ₁₀ H ₂₁ Br	1
CS86	C ₈ H ₆ O ₄	C ₈ H ₄ O ₄	1
CS871	C ₇ H ₈ O	C ₇ H ₂ O	1
GMR1	C ₁₆ H ₁₈ O ₂	C ₈ H ₅ O	35
GMR14	C ₃₅ H ₆₂ O ₃	C ₁₇ H ₂₃ O	26
GMR21	C ₂₄ H ₄₃ NO	C ₁₉ H ₂₇ NO	14
GMR27	C ₈ H ₈ O ₃	C ₈ H ₅ O	4
GMR8	C ₁₃ H ₁₂ O ₂	C ₈ H ₅ O	18
TBMP24	C ₁₁ H ₁₆ O	C ₆ H ₃ O	12

number of candidate formulae to one and thus additional elemental composition data are required. This can be obtained from further improved rules and MS and MS/MS isotopic ratio data. Incorrect inclusions from MAPS can cause the MFG program to generate a list of erroneous formula. This again underlines the need for reliable rules.

GENOA Results

When insufficient substructural constraints are provided to GENOA, many structures are generated. When several molecular formulae are consistent with the data, even more structures are possible.

For these reasons, structure generation was not attempted for 13 of the 16 test compounds. This failure to identify the structures for these 13 compounds cannot be attributed to the inability of MAPS to identify sufficient substructural constraints, but to the generation of several candidate formulae for these test compounds. Single formulae were identified by the MFG program for three test compounds: CS608, CS86, and CS871. The formulae identified by the MFG program and the substructural constraints identified by MAPS were used by GENOA for structure generation for each of these three compounds. When this was done, correct identification of the complete structure was achieved for CS608 and CS86. Three structures were identified for CS871, which correspond to the ortho, meta, and para isomers of cresol. The correct results hinged upon the identification of large portions of the complete structures. For CS608, the decyl and bromo substructures were identified, thus specifying all of the nonredundant substructures in this compound. For CS86, the phthalate substructure was identified, which specifies all of the nonredundant substructures in this compound except for the carboxyl substructure. For CS871, the identification of the benzyl and phenol substructures indicates that the compound is most likely a cresol but does not indicate which isomer.

Conclusions

This chapter has demonstrated the utility of ACES for structure elucidation, the performance and interaction of the individual ACES tools, and the situations which can lead to false positives. The results

shown in this chapter also underline the danger of false positives -- *a false positive guarantees that the correct structure will not be contained in the set of structures generated by GENOA.*

It is important to realize that the results for these test compounds are preliminary. Substructural constraints were obtained from an earlier version of MAPS which is described in Chapter 4. The use of multiple daughter ion and neutral loss features in the rule generation and the use of feature combinations would yield improved recall and fewer false positives for these test compounds. However, later versions of MAPS which incorporate these methodologies could not be used due to the long computer processing times required for rule generation. In addition, the complete scheme for molecular formula determination described in Chapter 6 was not employed for these test compounds. Elemental composition data was obtained only from MAPS results, and not from MS and MS/MS isotopic ratios. Thus, use of the MFG program yielded multiple molecular formulae for all but three of the test compounds.

Although complete structure elucidation was achieved for only 2 of the 16 test compounds, the results were still encouraging. Useful information was obtained for even those compounds for which structure generation was not attempted. This should be contrasted with spectral matching methods which provide either a right or wrong answer. Since ACES is an interpretive system, results from the ACES tools can provide at least a partial answer to the identity of an unknown.

CHAPTER 8

SUGGESTIONS FOR FUTURE WORK

Introduction

Beynon and coworkers demonstrated that complete structure elucidation could often be achieved using MS/MS data obtained from MIKES instrument (1). The TQMS instrument, developed by Rick Yost and Dr. Enke, has many advantages over the MIKES instrument, including unit resolution in both mass analyzers, greater CAD efficiencies, greater ease of computer control, and availability of parent, daughter, and neutral loss spectra (2). Up to now, MS/MS has been used mainly for target compound identification in complex mixtures and has not been fully exploited for structure elucidation (3,4). There are several reasons for this, mainly the lack of a generic database of CAD spectra and automated systems for interpretation of MS/MS spectra.

ACES represents the *first automated system for structure elucidation using MS/MS data* and has addressed a serious deficiency in this research area. It represents the culmination of the work of sundry individuals over a period of several years and its development has opened many doors for further research. This chapter outlines some of the modifications which can be made to improve the performance of ACES. The computer systems and instrumentation available to this research

group, combined with the group's experience in MS/MS, artificial intelligence, and computer programming should provide ample opportunities to further the development of ACES.

Expansion of the Training Set

The training set for MAPS needs to be expanded to cover a wider variety of substructures in different structural environments. This task is crucial for the development of a large rule base, which is required for determining the structures of a wide range of organic compounds. In addition, expansion of the training set is required for making valid comparisons between ACES and other structure elucidation techniques (such as DENDRAL, STIRS, and spectral matching programs). I was responsible for collection of the MS and MS/MS data which comprises the current training set using the Extrel TQMS instrument in our laboratory. This represented many months worth of work. This research group has recently purchased a Finnigan TSQ-70, a second generation TQMS instrument. The control language for this instrument allows facile development of specialized data acquisition routines. Hence, expansion of the training set can proceed much more quickly and easily with this instrument. In addition, the TSQ-70 has a higher signal to noise ratio due to a novel bent collision chamber, hyperbolic quadrupoles, a higher mass range, better ion transmission characteristics, and greater sensitivity than the Extrel instrument. These factors will contribute to the acquisition of higher quality spectra on the TSQ-70 compared to the Extrel instrument.

The mass spectrometry community has still not reached an agreement on standard operating conditions for collecting a generic, instrument-independent database of CAD spectra (5). Martinez has identified some of the prerequisites for collecting an database of such spectra (6,7). Once the criteria for collecting instrument-independent CAD spectra are *fully* defined, the development of a database of such spectra should be organized and coordinated by workers in industry, academia, and government. Until this occurs, a worthwhile goal for this research group would be to expand the training set using instrumental operating conditions consistent with the recommendations of Martinez to increase the quantity and, more importantly, improve the quality of the rules developed by MAPS.

Improvements to MAPS

I foresee many modifications which can be made to MAPS to improve the predictive capabilities of the rules. The quality of the rules is of crucial importance to their performance. In the process of generating a rule, the elemental composition of the substructure and valence rules are used to generate all possible fragments which may be attributed to that substructure. This information is used to create a list of fragment masses which may arise from that substructure. Any feature in the rule which has a mass that is not on the list of possible fragment masses for that substructure is removed from the rule. This "molecular formula checking" of feature masses removes many spurious clauses from the rules.

However, spurious clauses may still appear in the rules in some instances. For example, there are several clauses in the phenol inclusion rule which are suspect (i.e., parent and daughter ions at m/z 43 and 57, neutral losses of 44 and 58 u, and parent to daughter transitions from parent masses of m/z 57 and 71). These clauses may be caused by the concomitant presence of alkyl substructures in the phenol compounds in the training set. Although the features represented by these clauses can arise from the phenol substructure (based on the elemental composition of this substructure and valence rules), they may be more accurately attributed to alkyl substructures. This problem of spurious clauses in the rules may be somewhat alleviated by ensuring that each substructure in the training set is well-characterized or by using higher C values, which will filter out features with low levels of correlation from the rules.

It is possible to develop an additional filter to identify and remove such spurious clauses. Spurious clauses may be indicated by cross correlation factors between substructures, cross correlation factors between rules, and the levels of correlation for common clauses between rules. For example, the level of correlation between the phenol and t-butyl substructures is 52%. This indicates that 52% of the phenol compounds in the training set also contain a t-butyl substructure. Thus, it is possible that the phenol rule may be contaminated with clauses which may be more accurately attributed to the t-butyl substructure. Cross correlation of these two rules indicates that 81 out of the 148 clauses in the phenol rule are also present in the t-butyl rule. Such a high degree of overlap between rules is usually indicative of spurious

clauses. The clauses in common between these two rules are shown in Table 8.1 along with their respective levels of correlation and fragment formulae (for all but parent-to-daughter, multiple daughter ion, and multiple neutral loss features). Using such data, spurious clauses in the rules can be identified for removal. An algorithm incorporating this methodology should be implemented to identify and remove spurious clauses in the rules as a post-filter in the rule generation.

Currently, the exclusion rules use only lines in conventional mass spectra and daughter spectra, neutral losses, and parent-to-daughter transitions features in the rule generation. Recently, the MAPS code was modified to implement additional feature types into the *inclusion* rule generation process: multiple daughter ions and neutral losses from the same parent ion. This substantially improved their performance. These feature types are currently *not* used in the *exclusion* rule generation process. At the time this code was developed, it did not appear that multiple daughter and neutral loss features would possess the necessary level of correlation (100%) to be included in the exclusion rules. This assumption was found to be erroneous. Thus, these features types should be implemented into the exclusion rule generation process. Since they are generally more unique than most other features, they are more likely to be absent when the corresponding substructure is missing in an unknown. Thus, addition of multiple daughters and neutral loss features to the exclusion rules will improve their recall.

MAPS currently uses both MS and MS/MS data in generating rules. It would be interesting to study rule performance when MS or MS/MS data alone are used. Such a study should certainly prove the

Table 8.1 Fragment formulae and correlation factors for clauses common to both the phenol and t-butyl inclusion rules (generated at a C value of 33%)

CLAUSE			PHENOL RULE		T-BUTYL RULE	
			Corr	Formulae	Corr	Formulae
medium	intensity	daughter ion at m/z 15	68%	CH ₃	74%	CH ₃
medium	intensity	daughter ion at m/z 27	74%	C ₂ H ₃	79%	C ₂ H ₃
strong	intensity	daughter ion at m/z 27	74%	C ₂ H ₃	63%	C ₂ H ₃
medium	intensity	daughter ion at m/z 29	42%	C ₂ H ₅ , CHO	58%	C ₂ H ₅
strong	intensity	daughter ion at m/z 29	61%	C ₂ H ₅ , CHO	74%	C ₂ H ₅
medium	intensity	daughter ion at m/z 39	81%	C ₃ H ₃	84%	C ₃ H ₃
strong	intensity	daughter ion at m/z 39	77%	C ₃ H ₃	68%	C ₃ H ₃
medium	intensity	daughter ion at m/z 41	77%	C ₃ H ₅ , C ₂ HO	89%	C ₃ H ₅
strong	intensity	daughter ion at m/z 41	68%	C ₃ H ₅ , C ₂ HO	74%	C ₃ H ₅
medium	intensity	daughter ion at m/z 43	35%	C₃H₇, C₂H₃O	58%	C₃H₇
strong	intensity	daughter ion at m/z 43	39%	C₃H₇, C₂H₃O	42%	C₃H₇
strong	intensity	daughter ion at m/z 51	94%	C ₄ H ₃	89%	C ₄ H ₃
medium	intensity	daughter ion at m/z 55	65%	C ₄ H ₇ , C ₃ H ₃ O	58%	C ₄ H ₇
strong	intensity	daughter ion at m/z 55	45%	C ₄ H ₇ , C ₃ H ₃ O	47%	C ₄ H ₇
strong	intensity	daughter ion at m/z 57	42%	C₃H₅O	47%	C₄H₉
medium	intensity	neutral loss of 2 amu	84%	H ₂	84%	H ₂
strong	intensity	neutral loss of 2 amu	84%	H ₂	89%	H ₂
strong	intensity	neutral loss of 14 amu	35%	CH ₂	53%	CH ₂
medium	intensity	neutral loss of 15 amu	45%	CH ₃	53%	CH ₃
strong	intensity	neutral loss of 15 amu	81%	CH ₃	89%	CH ₃
medium	intensity	neutral loss of 16 amu	61%	CH ₄ , 0	84%	CH ₄
strong	intensity	neutral loss of 16 amu	55%	CH ₄ , 0	68%	CH ₄
medium	intensity	neutral loss of 18 amu	77%	H ₂ O, CH ₄ +H ₂	79%	CH ₄ +H ₂
medium	intensity	neutral loss of 26 amu	90%	C ₂ H ₂	84%	C ₂ H ₂
strong	intensity	neutral loss of 26 amu	100	C ₂ H ₂	95%	C ₂ H ₂
medium	intensity	neutral loss of 28 amu	81%	C ₂ H ₄ , CO	84%	C ₂ H ₄
strong	intensity	neutral loss of 28 amu	90%	C ₂ H ₄ , CO	95%	C ₂ H ₄
medium	intensity	neutral loss of 30 amu	71%	C ₂ H ₆ , CH ₂ O	79%	C ₂ H ₆

Table 8.1 (continued)

CLAUSE	PHENOL RULE		T-BUTYL RULE	
	Corr	Formulae	Corr	Formulae
strong intensity neutral loss of 30 amu	65%	C ₂ H ₆ , CH ₂ O	74%	C ₂ H ₆
medium intensity neutral loss of 40 amu	52%	C ₃ H ₄ , C ₂ O	53%	C ₃ H ₄
strong intensity neutral loss of 40 amu	55%	C ₃ H ₄ , C ₂ O	68%	C ₃ H ₄
medium intensity neutral loss of 42 amu	48%	C ₃ H ₆ , C ₂ H ₂ O	63%	C ₃ H ₆
strong intensity neutral loss of 42 amu	55%	C ₃ H ₆ , C ₂ H ₂ O	68%	C ₃ H ₆
medium intensity neutral loss of 43 amu	42%	C ₃ H ₇ , C ₂ H ₃ O	42%	C ₃ H ₇
medium intensity neutral loss of 44 amu	48%	C₃H₈, C₂H₄O	68%	C₃H₈
strong intensity neutral loss of 44 amu	58%	C₃H₈, C₂H₄O	79%	C₃H₈
medium intensity neutral loss of 50 amu	48%	C ₄ H ₂	53%	C ₄ H ₂
strong intensity neutral loss of 50 amu	35%	C ₄ H ₂	37%	C ₄ H ₂
medium intensity neutral loss of 52 amu	45%	C ₄ H ₄ , C ₃ O	42%	C ₄ H ₄
medium intensity neutral loss of 54 amu	45%	C ₄ H ₆ , C ₃ H ₂ O	47%	C ₄ H ₆
strong intensity neutral loss of 54 amu	65%	C ₄ H ₆ , C ₃ H ₂ O	58%	C ₄ H ₆
medium intensity neutral loss of 56 amu	52%	C ₄ H ₈ , C ₃ H ₄ O	63%	C ₄ H ₈
strong intensity neutral loss of 56 amu	58%	C ₄ H ₈ , C ₃ H ₄ O	63%	C ₄ H ₈
strong intensity neutral loss of 58 amu	39%	C₃H₆O	58%	C₄H₁₀
strong intensity parent ion at m/z 41	68%	C ₃ H ₅ , C ₂ H ₂ O	95%	C ₃ H ₅
medium intensity parent ion at m/z 42	48%	C ₃ H ₆ , C ₂ H ₂ O	53%	C ₃ H ₆
medium intensity parent ion at m/z 43	55%	C₃H₇, C₂H₃O	58%	C₃H₇
weak intensity parent ion at m/z 44	45%	C ₃ H ₈ , C ₂ H ₄ O	37%	C ₃ H ₈
medium intensity parent ion at m/z 50	39%	C ₄ H ₂	37%	C ₄ H ₂
medium intensity parent ion at m/z 51	55%	C ₄ H ₃	58%	C ₄ H ₃
medium intensity parent ion at m/z 52	48%	C ₄ H ₄ , C ₃ O	42%	C ₄ H ₄
medium intensity parent ion at m/z 53	74%	C ₄ H ₅ , C ₃ H ₂ O	84%	C ₄ H ₅
medium intensity parent ion at m/z 55	71%	C ₄ H ₇ , C ₃ H ₃ O	58%	C ₄ H ₇
strong intensity parent ion at m/z 57	45%	C₃H₅O	68%	C₄H₉
medium intensity parent ion at m/z 58	42%	C ₃ H ₆ O	74%	C ₄ H ₁₀
medium intensity parent to daughter transition of m/z 41 to 15	68%		74%	

Table 8.1 (continued)

CLAUSE	PHENOL RULE		T-BUTYL RULE	
	Corr	Formulae	Corr	Formulae
medium intensity parent to daughter transition of m/z 41 to 39	68%		79%	
medium intensity parent to daughter transition of m/z 43 to 15	52%		53%	
medium intensity parent to daughter transition of m/z 43 to 27	55%		68%	
medium intensity parent to daughter transition of m/z 43 to 41	42%		47%	
strong intensity parent to daughter transition of m/z 53 to 27	55%		42%	
strong intensity parent to daughter transition of m/z 55 to 29	45%		58%	
medium intensity parent to daughter transition of m/z 55 to 39	48%		53%	
medium intensity parent to daughter transition of m/z 57 to 27	39%		42%	
strong intensity parent to daughter transition of m/z 57 to 29	58%		68%	
medium intensity parent to daughter transition of m/z 57 to 39	48%		58%	
parent giving daughter ions of m/z 41m, 55m	35%		52%	
parent giving neutral losses of 2s, 26s	35%		52%	
parent giving neutral losses of 2s, 28s	38%		63%	
parent giving neutral losses of 15s, 28s	35%		52%	
parent giving neutral losses of 26s, 28s	45%		73%	
parent giving neutral losses of 28m, 56m	35%		47%	
parent giving neutral losses of 28s, 30s	48%		57%	
parent giving neutral losses of 28s, 40m	38%		52%	
parent giving neutral losses of 28s, 40s	35%		52%	
parent giving neutral losses of 28s, 42s	38%		57%	

Table 8.1 (continued)

CLAUSE	PHENOL RULE		T-BUTYL RULE	
		Corr Formulae	Corr Formulae	
parent giving neutral losses of 28s, 42s, 56s		35%		47%
parent giving neutral losses of 28s, 44s		38%		63%
parent giving neutral losses of 28s, 54s		48%		57%
parent giving neutral losses of 28s, 56s		58%		63%
parent giving neutral losses of 42s, 56s		35%		69%

greater utility of MS/MS data for substructure identification. Currently, five different types of features derived from MS/MS data are used in the development of inclusion rules by MAPS. The development of hybrid mass spectrometers, such as the BEQQ instrument (8,9) and penta-quadrupole instruments (10,11), has made it possible to perform MS/MS/MS (MS^3). Five consecutive stages of mass spectrometry have been followed using an FTMS instrument (12). Wade et. al. have employed morphological analysis techniques to identify the scan modes associated with multiple stage mass spectrometry (13). *All MS^n techniques for which n is greater than one can be used to inspect various portions of a fragmentation map and monitor specific fragmentation patterns.* For example, Louris and coworkers demonstrated that a consecutive neutral loss scan of 30 u followed by 28 u could be used as a highly selective diagnostic feature for nitroaromatic compounds (14). Such consecutive fragmentation patterns may only be *suggested* by data from MS/MS instruments. The higher dimensionality features obtained from MS^n instrumentation are even more characteristic of substructures than MS/MS features. The MAPS code can be readily extended to include such higher dimensionality features. This will greatly improve the predictive capabilities of the rules.

A fuzzy-logic intensity matching system was developed for matching the real intensities in unknowns to the intensity classes in rule clauses. The use of this in the rule application process did not improve rule performance, most likely due to the nature of the function defining the degree of match between real intensities and intensity classes. The use of a modified fuzzy-logic intensity matching system as well as the

creation of new intensity classes may yet improve rule performance. However, the variability of intensity data from MS/MS instruments must be kept in mind when making these modifications.

The approach described in Chapter 5 shows great promise for reliable substructure identification and addresses many of the deficiencies in previous versions of MAPS. This approach involves finding *combinations* of individual MS and MS/MS spectral features which have reliabilities of 100% for substructure identification based on training set data. This approach can be extended to find combinations of features which are indicative of the absence of substructures. Combinatorial explosion problems limited this approach to rules which contained less than 30 clauses. Currently, the search algorithm can be described as forward chaining, that is, it generates new combinations by adding features onto an existing combination. Each training set compound which possesses a given substructure exhibits a combination of features which is indicative of that substructure. Such information can be used to constrain the search. Perhaps a backward chaining approach can be developed to generate combinations by subtracting features from combinations which are present in training set compounds. The addition of further heuristics to the search algorithm, the use of *forward and backward* chaining techniques, and the use of more powerful or parallel computing facilities may improve the attractiveness of this approach. It may be very helpful to obtain the advice of persons knowledgeable in the area of search techniques to further the development of this approach, as this problem has many analogues in the field of artificial intelligence.

MAPS can accept a training set of MS/MS data regardless of the type of MS/MS instrument used for data collection. It would be interesting to see how the rules perform with MS/MS data obtained from the TRIMS instrument developed in this laboratory (15,16). The MAPS code could also be modified to utilize high resolution MS/MS data obtained from FTMS and BEBE instruments. Such high resolution data would also greatly improve the predictive capabilities of the rules.

The rules are currently applied to unknowns in alphabetic order. This is computationally wasteful. Forward and backward chaining techniques commonly used in expert systems can be used to direct the rule application process. The substructure hierarchy, which defines inheritance relationships between substructures as described in Chapter 4, can be used to control the order in which the rules are applied. The exclusion rules should be applied to an unknown first to identify the substructures which are absent. Assume that the alkyl exclusion rules are applied in the following order: hexadecyl, pentadecyl, ... methyl. If the pentadecyl substructure is identified as absent, then there is no reason to apply the exclusion rules for the substructures lower in the hierarchy (i.e., methyl, butyl, ... , tetradecyl) since these substructures are contained in the pentadecyl substructure and thus must also be absent. The results from the application of the exclusion rules can be used to eliminate specific substructures from the inclusion rule application. If the phenyl substructure is known to be absent, there is no need to apply the inclusion rules for substructures containing the phenyl substructure. The hierarchy can be used in a similar manner to control the control the order in which the inclusion rules are applied.

Assume that the alkyl inclusion rules are applied in the following order: hexadecyl, pentadecyl, ... methyl. If the pentadecyl substructure is identified as present, then there is no reason to apply the inclusion rules for the substructures lower in the hierarchy (i.e., methyl, butyl, ... , tetradecyl) since these substructures are again contained in the pentadecyl substructure and thus will also be identified as present. The substructure hierarchy has already been implemented and additional software should be written to accomplish the application of the rules in a tree-type format. This will very effectively improve the speed at which substructures are identified as present or absent in unknowns.

Beynon and coworkers demonstrated that it is possible to deduce the presence of multiple occurrences of a given substructure in a molecule from MS/MS data (1). MAPS and ACES would certainly benefit from such information. More work needs to be done to formulate some methodology to achieve this automatically.

The rules identified by MAPS may find many applications. Chemical interpretation of the rules and feature combinations can be used to study the fragmentation patterns of substructures in different operating conditions and different structural environments. A test which monitors rule content and performance as a function of instrumental operating conditions would define the limitations of the rules. The rules themselves may be used as highly diagnostic tests in the analysis of mixtures for specific substructures or compound classes.

MAPS can be considered to be an expert system for several reasons. Prior to the development of this software, little work had been done to deduce the MS/MS fragmentation patterns characteristic of

substructures. MAPS *automatically* deduces the relationships between MS and MS/MS spectral features and substructures. In doing so, it creates expertise where little or none previously existed. The rules identified by MAPS can be used to train mass spectroscopists to recognize the MS/MS fragmentation patterns characteristic of specific substructures. Several modifications will be necessary to make MAPS a "true" expert system. The rule application process currently includes "hooks" to display the features in an unknown which cause specific substructures to be identified as present and absent. This code should be modified to allow the user to query MAPS for the reasoning behind its decisions.

Further development of the MAPS software has encountered a bottleneck in the Xerox 1108 hardware. Although the InterLISP-D system on this computer has many excellent features and additional software to facilitate applications development, these "extras" come at a substantial price in memory. LISP systems are notoriously memory intensive and InterLISP-D is no exception. The current system has 16 Mbytes of virtual memory. The InterLISP-D software alone consumes almost 5 Mbytes of this virtual memory! When the MAPS software, training set, rules, and the associated data structures are loaded, only 50% of the available virtual memory space is free! The problem of limited memory will only become worse as the training set grows and the software increases in complexity. In addition, execution speeds on this computer are often excessively long. For example, generation of the inclusion and exclusion rules takes two days of processing time. Generation of the features combinations from a set of 30 features make

take more than 130 days of computation time. These long execution times may be attributed to limited memory (1.5 Mbytes) and limited disk space (43 MBytes) on the Xerox 1108. Many page faults occur with this computer due to the limited amount of real memory. In defense of this computer and the software I have written, I must concede that such long computation times are to some extent necessitated by the magnitude and complexity of the problem. However, there is no disputing the fact that this computer has become obsolete by modern standards.

For these reasons, alternatives to the Xerox 1108 for continued development of the MAPS software must be considered. The Xerox 1186 AI workstation is upwardly compatible with the Xerox 1108, has increased processing power, and includes 3.7 Mbytes of memory. Personal computers represent another option for replacing the Xerox 1108. Processing speeds for PC's have been greatly increased over the last few years. A PC running with an Intel 80386 microprocessor can achieve 3 million instructions per second. Software called POWERLISP is available for IBM-compatible PC's (17). The MAPS code could be implemented on a PC running POWERLISP with few modifications since this version of LISP is compatible with InterLISP-D.

Chris Weaver has been translating the MAPS code to Common LISP. It appears that the computer industry is moving towards Common LISP as a standard (18). There are several advantages to having MAPS translated to Common LISP. This language runs on a variety of minicomputers and PC's. More importantly, a version of MAPS written in Common LISP will run on the MicroVAX II computer in our laboratory. There are several advantages to this. All other ACES tools, including

GENOA and the MFG program are currently running on this computer. The VMS operating system on the MicroVAX II allows communication between software modules written in different languages and will be ideal for ACES, since GENOA, MAPS, and the MFG software are written in three different programming languages. Benchmarks done in this laboratory have shown that MAPS will run faster on the MicroVAX using a commercial version of Common LISP (19) compared to the Xerox 1108. The faster execution speeds can be attributed to the greater memory resources (9 MBytes of real memory and over 300 MBytes of virtual memory) on the MicroVAX, as well as a more modern, efficient LISP compiler. The translation of MAPS from InterLISP-D to Common LISP will allow the ACES tools to be completely integrated on one computer and will greatly enhance its portability to other VAX computers.

Improvements to the MFG program

As mentioned in Chapter 6, daughter spectra of isotopic ions allows their elemental composition to be determined. I have used such data to determine the number of carbon atoms in molecular ions. Recent work by Bozorgzadeh has made possible the determination of the *complete* elemental composition of parent and daughter ions from daughter spectra of isotopic ions (20-22). This methodology and the associated software should be implemented into the molecular formula determination process in ACES. This will certainly aid in the determination of molecular formulae from unit resolution MS and

MS/MS data and should result in a further reduction in the number of candidate formulae generated by the MFG program.

Many programs have been written for calculating theoretical intensity ratios of isotope patterns in unit resolution mass spectra. Kavanagh and Tenhosaari have developed algorithms for ranking candidate formulae on the basis of the similarity of their theoretical isotope patterns to experimental isotope patterns (23,24). The experimental intensity ratios, however, must be measured very accurately so that the correct formula is ranked first. This methodology should be implemented in the MFG program to achieve further reduction of the list of candidate formulae.

Improvements to GENOA

The source code the GENOA is being modified by Kevin Hart to automate substructure searching and structure generation sessions. Kevin has also developed libraries of structures and substructures. The modifications and improvements to GENOA will be addressed in Kevin's thesis and will not be discussed further here.

Further Development of ACES

Structure elucidation of true unknowns is often a very complex problem requiring human expertise. ACES is an expert system which uses substructure identification rules, a molecular formula generator, and a structure generator to solve structure elucidation problems. It

requires several additional components to make it a "true" expert system. It should include software to justify its reasoning. A natural language interface would certainly facilitate communication with users. ACES also would benefit from some methodology for verifying proposed structures. Perhaps it would be useful to collaborate with researchers in expert system development to obtain recommendations for improvements to ACES, since there is little expertise in this area in the chemistry department at Michigan State University.

One additional component to ACES has been envisioned - an intelligent controller (IC). Its main task would be to guide and direct the structure elucidation process. It would completely automate data transfers between the instrument and ACES, and communication between software modules. The IC and ACES would greatly benefit from the use of a blackboard to facilitate communication between software modules. Blackboards are a common component of expert systems and are used as global databases for storage of available facts and the decisions made by individual experts. An additional task for the IC would be a module for identifying additional experimentation to confirm known portions of the unknown structure and identify as yet unknown portions. This additional experimentation may involve using chemical ionization to determine the molecular weight, collecting daughter spectra using different collision gas pressures and collision energies for substructure identification, or using ion-molecule reactions in the collision chamber to differentiate between isomers. The IC should contain rules and procedures for performing these types of experiments.

A long-sought goal of many researchers in analytical chemistry is the development of "intelligent" instrumentation. Such instrumentation is capable of not only automated optimization of instrumental parameters and data acquisition but can also perform data analysis and interpretation. This goal is now within reach for TQMS instrumentation. A diagram of an intelligent TQMS instrument is shown in Figure 8.1. The TQMS instrument in this figure could be the Finnigan TSQ-70, which includes software modules for data processing, instrument control, and diagnostics. The expert system would contain all of the software tools from ACES along with the IC. This intelligent instrument includes a feedback loop between the interpretive tools and the instrument. The data obtained from the instrument would be processed by the ACES tools. Results from GENOA would be summarized to show the known and unknown portions of the complete structure. From these results, the IC should be able to suggest and initiate additional experimentation to resolve the identity of the unknown to the experimenter's satisfaction. These tasks can be automated since the Finnigan TSQ-70 is under complete computer control and has a very flexible and powerful instrument control language. On each pass through the feedback loop, the results from the expert interpretive tools would be fed back to the user.

While MS/MS is a very powerful technique for structure elucidation, it is important to realize its limitations. Functionalities which are often difficult to identify by MS or MS/MS, such as the hydroxyl, carbonyl, and carboxyl substructures, can be often be readily identified by i.r. data. PAIRS, an expert system for interpretation of i.r.

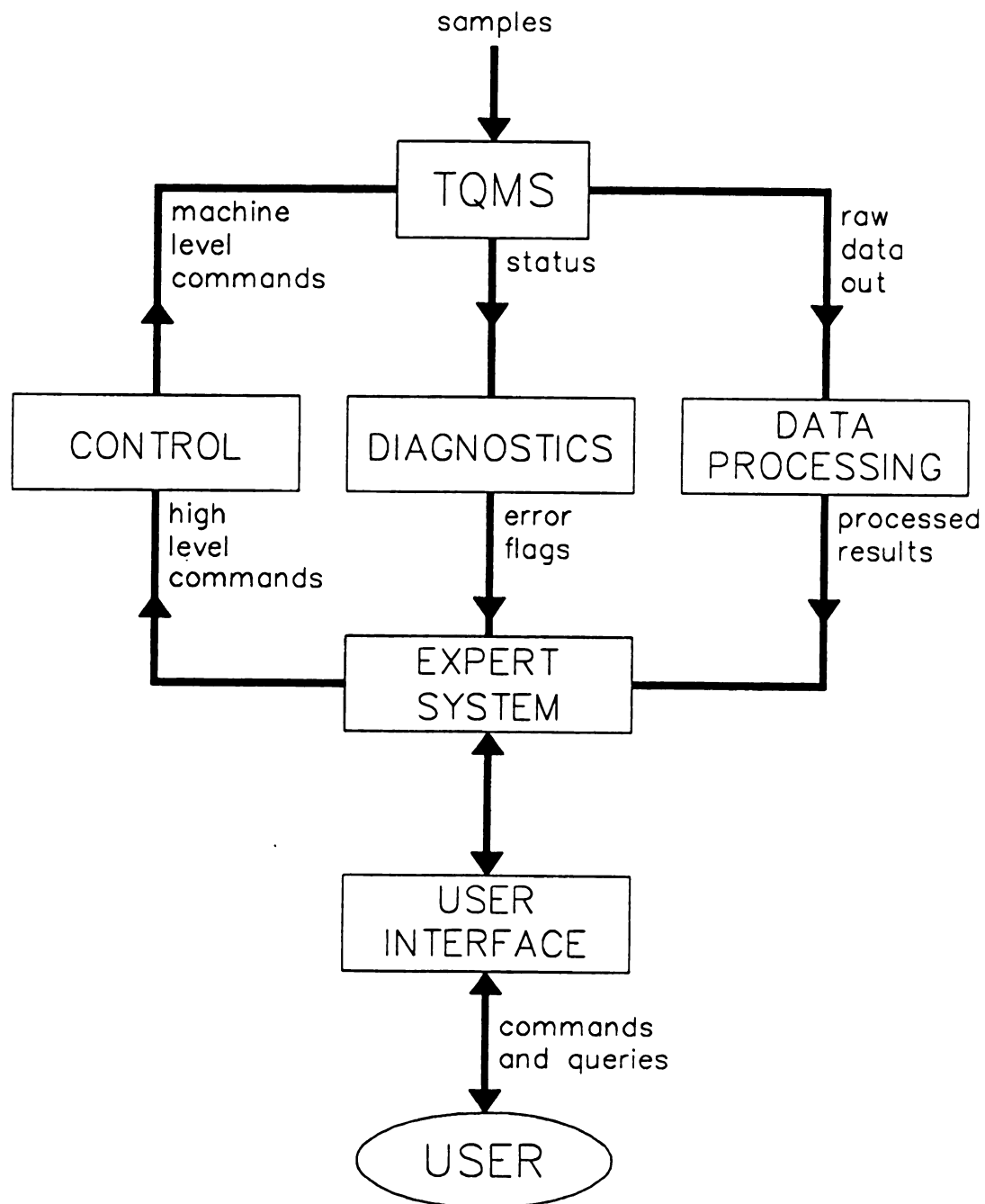


Figure 8.1 An intelligent TQMS instrument incorporating a feedback loop from expert interpretive tools to the instrument.

data (25,26), could be implemented into ACES to produce an even more powerful structure elucidation system. Likewise, ACES would certainly benefit from the connectivity and substructure information which can be obtained from n.m.r. data. Spectroscopic data from such different techniques is complementary and can be used to provide confirmatory information or formulate multiple competing hypotheses. Currently, only a few expert systems use MS, i.r., and n.m.r. data for structure elucidation, the most well-known of which is the CHEMICS system (27). ACES is the first structure elucidation system to employ MS/MS data. When combined with other expert systems for structure elucidation employing spectroscopic data from other techniques, it should prove to be an extremely powerful tool.

This thesis has demonstrated the utility of ACES for structure elucidation. However, a tool cannot be used unless it is generally available. ACES should be made available to industry to promote its use and acceptance. Perhaps the rules themselves could be made available to the public for now, since ACES is still under development. Eventually, users should have access to the entire ACES system. This would allow them to develop their own training sets and rules for specific applications, and use these tools to aid in the structure elucidation of unknowns. In addition, ACES should be compared to spectral matching techniques (such as PBM) and interpretive techniques (such as STIRS and DENDRAL). Such a study would provide an objective evaluation of its performance.

ACES combines techniques of pattern recognition and artificial intelligence with the power of MS/MS for structure elucidation and

shows great promise. Its development has opened many avenues for additional research. Computer, software, and artificial intelligence technologies have undergone tremendous leaps over the last decade. The time is ripe for the development of an expert system for structure elucidation using data from different spectrometric techniques. ACES represents one step towards this goal.

References

1. Bozorgzadeh, M.H., Morgan, R.P., Beynon, J.H., Analyst, **103**, 613 (1978).
2. Yost, R.A., Enke, C.G., Anal. Chem., **51**, 1251A (1979).
3. Hunt, D.F., Shabinowitz, J., Harvey, T.M., Coates, M., Anal. Chem., **57**, 525 (1985).
4. Bauer, M.R., Ph.D. Dissertation, Michigan State University, East Lansing, MI, 1987.
5. Martinez, R.I., Cooks, R.G., 35th Annual Conference on Mass Spectrometry and Allied Topics, Denver, CO, May 1987, p. 1175.
6. Martinez, R.I., Rapid Comm. Mass Spectrom., **1**, 8 (1988).
7. Martinez, R.I., Dheandhanoo, S., J. Res. Natl. Bur. Stand., **92**, 229 (1987).
8. Schoen, A.E., Amy, J.W., Clupek, J.D., Cooks, R.G., Dobberstein, P., Jung, G., Int. J. Mass Spectrom. Ion Proc., **65**, 125 (1985).
9. Clupek, J.D., Amy, J.W., Cooks, R.G., Schoen, A.E., Int. J. Mass Spectrom. Ion Proc., **65**, 141 (1985).
10. Morrison, J.D., Stanney, K.A., Tedder, J., 34th Annual Conference on Mass Spectrometry and Allied Topics, 1986, p. 222.
11. Beaugrand, C., Devant, G., Rolando, C., 34th Annual Conference on Mass Spectrometry and Allied Topics, 1986, p. 220.
12. Laukien, F.H., Abstract 524, 14th Annual FACSS Meeting, Detroit, MI, 1987.
13. Wade, A.P., Cooks, R.G., Enke, C.G., submitted to Int. J. Mass Spectrom. Ion Proc.
14. Louris, J.N., Wright, L.G., Cooks, R.G., Schoen, A.E., Anal. Chem., **57**, 2918 (1985).
15. Stults, J.T., Enke, C.G., Holland, J.F., Anal. Chem., **55**, 1323 (1983).
16. Eckenrode, B., Newcome, B.H., Holland, J.F., Enke, C.G., 35th Annual Conf. on Mass Spectrometry and Allied Topics, 1987, p. 277.

17. Powerlisp Literature, MicroProducts, 370 W. Camino Gardens Blvd., Boca Raton, FL, 33432.
18. Allen, J.R., AI Expert, **2**, 48 (1987).
19. Lucid Common LISP, Lucid Inc., 707 Laurel St., Menlo Park, CA 94025.
20. Bozorgzadeh, M.H., Lapp, R.L., 34th Annual Conference on Mass Spectrometry and Allied Topics, 1986, p. 428.
21. Bozorgzadeh, M.H., 35th Annual Conference on Mass Spectrometry and Allied Topics, 1987, p. 395.
22. Bozorgzadeh, M.H., Rapid Comm. Mass Spectrom., **2**, 61 (1988).
23. Kavanagh, P.E., Org. Mass Spectrom., **15**, 334 (1980).
24. Tenhosaari, A., Org. Mass Spectrom., **23**, 236 (1988).
25. Woodruff, H.B., Smith, G.M., Anal. Chem., **52**, 2321 (1980).
26. Woodruff, H.B., Smith, G.M., Anal. Chem., **53**, 543 (1981).
27. Yamasaki, T., Abe, H., Kudo, Y., Sasaki, S., in Smith, D.H. (Ed.), "Computer-Assited Structure Elucidation", American Chemical Society, Washington, DC, 1977, p. 108.

