This is to certify that the

dissertation entitled

An Automated Structure Elucidation System for MS/MS Data:
Substructure Determination Through Spectral Matching

presented by

Kevin Patrick Cross

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Chemistry

C. G. Enke
Major professor

Date November 8, 1985

AN AUTOMATED STRUCTURE ELUCIDATION SYSTEM FOR MS/MS DATA: SUBSTRUCTURE DETERMINATION THROUGH SPECTRAL MATCHING

By

Kevin Patrick Cross

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Chemistry

1985

ABSTRACT

AN AUTOMATED STRUCTURE ELUCIDATION SYSTEM FOR MS/MS DATA: SUBSTRUCTURE DETERMINATION THROUGH SPECTRAL MATCHING

By

Kevin Patrick Cross

The building blocks for an automated structure elucidation system have been developed to evaluate the structural information contained in mass spectrometry/mass spectrometry (MS/MS) spectra. The system employs several software tools to assist in the determination of unknown organic structures. These include tools to: 1) match conventional and MS/MS spectra, 2) assist in the determination of spectra/substructures correlations, and 3) assist in the determination of structures from identified substructures.

Correlations of MS/MS daughter spectra with substructures are determined by matching daughter spectra of parent ions of identical masses. The molecular substructures giving rise to the parent ions with matching daughter spectra are then identified. This method of substructure determination is totally empirical and does not assume that the structural integrity of an ion is maintained in the ionization or fragmentation process. It does not, therefore, require the identification of ion structures.

The ability to group similar daughter spectra representing a common substructure is central to the substructure determination process. Therefore a program was developed to match an unknown MS/MS spectrum against either conventional or MS/MS spectra in a reference data base. It is an interactive, transparent program which employs several matching techniques for flexibility in matching criteria. An overall match factor is calculated which is a combination of "forward" and "reverse" searching techniques. This approach emphasizes common spectral features while de-emphasizing experimental conditions.

The ability of the search program to correctly discern identical compound MS/MS spectra taken under different operating conditions from similar spectra of other compounds has proven valuable in determining how instrumental conditions affect MS/MS spectra. By determining voltage and pressure ranges for each parameter that yield matching MS/MS spectra, acceptable standard conditions for acquiring MS/MS reference spectra were determined.

The MS/MS search program has been successfully applied to determine the substructures of the compound di-n-octylphthalate by grouping daughter spectra similar to those of the sample. These substructures were then combined to generate the complete phthalate molecule. The program has also been used to identify substructure/daughter spectrum correlations by matching daughter spectra of several known compounds to determine the molecular substructure associated with a particular daughter spectrum.

ACKNOWLEDGEMENTS

A project such as this results only from the cooperation and help of many individuals. They all deserve a word of thanks. My advisor, Chris Enke, deserves special acknowledgement for his patient guidance, innovative ideas, and pursuit of excellence. I am also grateful to the members of the "structure determination group" who provided an environment where all benefited from the ideas and work of others. These individuals include Phil Hoffman, Hugh Gregg, Anne Giordani, Pete Palmer, and Kevin Hart. A special thanks goes out to Tom Atkinson for all his efforts toward helping anyone with any computer problem at any time. I wish to thank Adam Schubert for the time he spent evaluating and discussing my work and the rest of the Enke group for their help and friendship. I would also like to thank all the professors at Lawrence University for the creative environment and intellectual stimulation they provided; especially my advisors, Dr. Robert M. Rosenberg and Dr. James S. Evans. Most of all, I would like to thank my wife, Carol, for all the love and endurance througout this period.

Financial support for this work was provided from the National Institutes of Health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Organic structure determination is one of the most sought after goals of chemical analysis. Positive identification of a compound whether in a clinical, academic, pharmaceutical, or industrial environment requires the determination of its structure. Compounds with minor structural differences that are often difficult to qualitatively analyze may have greatly differing effects on biological systems.

As the chemist seeks to identify complex compounds present in trace amounts, the number of possible interferences increases. Therefore, increasing the selectivity of modern instruments has become as important as improvements in sensitivity. Consequently, scientists are turning increasingly toward integrating techniques with complementary capabilities such as liquid chromatography/mass spectrometry and mass spectrometry/mass spectrometry (MS/MS).

Huge volumes of data are being produced by integrated technique instruments. The chemist has turned to the computer for help in storing and interpreting all this data.

The development of artificial intelligence guided instrumentation is a long range goal of the Enke research group. This has included development of an expert system to determine structures from low-resolution mass spectral data and an intelligent instrument control system for the triple quadrupole mass spectrometer. The

integration of these two systems will allow data acquisition decisions to be made by the expert system and then be automatically performed on the MS/MS instrument. This thesis will focus on my part in the ongoing design and development of an expert system to perform structure determination using MS/MS data, and in particluar, the development of spectral matching algorithms for MS and MS/MS spectra for the determination of molecular substructures.

Before describing the structure determination system, it is useful to digress to illustrate how and why the proposed method was developed, to examine previous work performed in this area and to note how the current project compares with previous accomplishments. The discussion starts by defining artificial intelligence and expert systems, and proceeds to discuss applications of expert systems to structure elucidation.

Artificial Intelligence

Artificial intelligence (AI) has many definitions; one of the broadest and probably most accurate was stated by Patrick Winston. "Artificial intelligence is the study of ideas which enable computers to do the things that make people seem intelligent" (1). This definition seems all encompassing since intelligence appears to be a conglomeration of many different information storage and processing abilities. The definition of artificial intelligence programs is more restrictive. "AI programs are those programs designed to emulate human performance in problem-solving activities through inductive reasoning and semantic

information processing" (1). In general, those programs which use inductive reasoning are those most commonly thought of as AI programs. Inductive reasoning is the process of reasoning from some observed cases to a universal conclusion regarding similar cases, some of which are unobserved.

The first conceptualization of AI as a usable tool was in the 1930s by E. Post, a logistics mathematician (2). He developed a set of production rules for manipulating groups of symbols. These rules served as a foundation for building several levels of rules into a knowledge base. In 1971, the first natural language interaction with a computer was demonstrated (2). In the early 1980s, Carnegie-Mellon University and Digital Equipment Corporation (DEC) collaborated on an AI project to configure computer systems (3,4). The development of AI concepts and tools for general use has been slow to emerge until very recently, when the development of powerful, inexpensive hardware has renewed interest in exploring what the computer can do for both the scientist and the businessman.

Artificial intelligence research and development can be divided into four areas:

1) **Expert systems** - The computer reasons from knowledge in a particular domain with expert ability.

2) **Natural language processing** - The computer interprets oral or written commands, acts upon them and then reports the results to the operator.

3) **Cognitive research** - The exploration of the human mind by using a computer to emulate the thinking process.

4) **Robotics** - The development of computers with special appendages to perform hazardous or mundane tasks.

The application of artificial intelligence to automated structure elucidation involves the development of an expert system to help the scientist analyze acquired experimental data and to postulate plausible structures. For MS/MS in particular, the goal is to let the expert system identify substructures represented by MS/MS spectra and then to combine these substructural units to elucidate a molecular structure.

Expert Systems

An expert system is a program that uses resident knowledge to solve specific problems normally requiring human expertise (4-9). Although types of expert systems vary, each system is made up of three parts: a means of knowledge representation, an inference engine, and a user interface.

(1) **Knowledge representation.** A method must be available to either maintain a knowledge base or to deduce knowledge from acquired data. Those systems which maintain a knowledge base are termed knowledge-based systems (6). Knowledge is stored in them as heuristic rules, or rules-of-thumb, which are entered by an expert in the specific

field. Those systems which deduce knowledge from experimental data are termed power-based systems (6). Instead of operating with an established set of rules, conclusions are derived solely using the data present. These systems are termed data-driven systems and are usually accompanied by large numeric data bases (8). The system we developed for identifying organic compounds using MS/MS data is best described as a data-driven, power-based expert system (although some heuristics are present). Large spectral and structure data bases are used to help elucidate the compound of interest.

(2) **Inference Engine.** Each expert system must have a method for operating on the knowledge. In knowledge-based systems, an inference engine sorts through the available heuristic rules and applies those that are relevant to the situation. In power-based systems, the inference engine is a computer algorithm that uses numerical techniques to compare experimental and reference data and thereby reach a conclusion.

(3) **User Interface.** Each expert system must identify the task to be solved. For specific problems this interface is very simple. As the range of problems increases, the user interface becomes more complicated. The interface reports the results of applying the system's expertise and demonstrates the validity of its conclusions by illustrating its reasoning. This is extremely important in not only providing the operator with confidence in its conclusions, but in determining the shortcomings of the system by pointing out flaws in the knowledge base or inference engine.

The most important characteristic of the expert system is the ability to comprehensively evaluate all possibilities for a given situation. In mathematical terms, the expert system systematically reduces the search space of possibilities until only one remains (5). It does this by subdividing the problem into many pieces and solving each piece by applying the appropriate algorithm. Solutions from subproblems are then combined to generate the final conclusion. Although the use of building blocks in solving complex problems is not new, it is well applied to expert systems (2). The inference engine methodically evaluates every possible situation. This feature alerts the operator to possibilities that he did not consider or have time to evaluate.

Those inference engines that use experimental data to reason toward a conclusion are data-driven systems. They use forward-chaining rules (5) to progress from the problem's beginning to the finish. There are also inference engines that reason backwards from the goals to the data. Engines that run in reverse are used in goal-driven systems. They use backward-chaining logic to progress through the knowledge base (5). Some expert systems use both forms of logic. The expert system being developed for structure determination uses MS/MS data and forward reasoning to reach its conclusions. The results from the expert system, however, are only as good as the information in the knowledge base. The saying, "garbage in, garbage out", is nowhere more applicable than in describing the misuse of expert systems.

Expert systems model certain aspects of human behavior better than others (10). Reasoning from well established rules and from a large data base are the strong points of expert systems. There are, however, several areas of weakness. An expert system cannot reason by analogy or from "first principles". In addition, an expert system has no common sense and therefore may produce ludicrous results. Advice and consultation of experts is needed when building a knowledge base, even for power-based systems. Most expert systems do not have certainty values associated with heuristic rules. While rules-of-thumb hold in the majority of the cases, many exceptions cause conflicts in logic by the inference engine. Demon rules (1) and fuzzy logic (11) are two methods currently being explored to resolve these situations.

The Focus of Expert Systems

The growth of expert system development in society stems from human demands and from developments in technology (5). Human demands for expert systems are fostered by the scarcity of human expertise and its perishable nature. The slow and error-prone diffusion of knowledge among coworkers is unsatisfactory in many situations. Lastly, the automation of laboratories has generated volumes of data that need interpretation (12,13).

Recent developments in both computer hardware and software have helped spur development of expert systems. Inexpensive processors, memory, and disk drives have made AI techniques available to laboratories that could not afford a personal computer five years ago.

Maturing of the AI field has produced new commercial software and hardware. Symbolic processors (Xerox, Symbolics), AI operating systems (PROLOG), knowledge environments (KEE, M1), and AI languages (LISP) are now available for the expert system developer (5). The application of expert systems will continue to increase as fifth generation computers, using large numbers of parallel processors, become available (14). In 1984 there were over 100 expert systems under development by over 2500 knowledge engineers and AI programmers. These figures are expected to double in 1985 and to exponentially increase over the next ten years (5).

Recent developments of expert systems have focused on interpretation of data from intelligent instruments and on the development of high-value specialized systems (5). Both of these areas represent limited knowledge domains, where a closed system of expertise can be developed. One example of a high-value expert system already developed is R1, a system developed by Carnegie-Mellon and DEC to configure VAX computer systems based on customer needs (3). DEC estimates that this system saves them 10 million dollars annually.

Several established companies currently involved in developing expert systems include: Digital Equipment Corporation, Ford Aerospace, General Motors, Hewlett-Packard, International Business Machines, National Aeronautics and Space Administration, Rand Corporation, and Xerox (5,15). Several new companies exclusively marketing expert systems have been born. These include: Intellicorp, IntelliGenetics, Corporation Systems, Teknowledge, and Visual Intelligence (5,8).

Applications of Expert Systems to Structure Elucidation

The development of computer methods for analyzing spectral data to elucidate structures has matured from the numerical reduction of data from a single instrument to expert systems capable of interpreting data from several different sources. Expert systems have commonly been applied to spectroscopic instruments including mass spectrometry (16-22), H^1NMR (16,22,23), C^{13}NMR (16,19,22-27), IR (19,23,26,28), and UV spectroscopy (29). Even x-ray powder diffraction and x-ray crystallography data have been incorporated into one structure elucidation system (16).

All expert systems performing structure elucidation include one or more of the techniques described below. Although a variety of data sources have been used, the discussion in this work focuses on those systems using mass spectra.

(1) **Numerical reduction of spectral data.** This category includes spectral identification through library searching and data grouping. The comparison of a sample spectrum to a stored library of spectra has been implemented for both low- and high-resolution mass spectra (17). Library search methods range from simple linear regression (18,30-35), to multiple regression techniques (36), and statistical methods (37). Identification of the sample spectrum often depends on the size and nature of the accompanying spectral library. Data grouping methods include factor analysis (38), pattern recognition

(26,36,39,40), cluster analysis (41,42), and applied information theory (43,44). In many cases the complete molecular structure cannot be solely determined by numerical methods and the identification of a substructure or functional group is considered satisfactory (45).

(2) **Substructure determination.** Often the entire molecular structure is not elucidated by using a single spectroscopic technique or numerical method. The operator may, however, be able to identify relevant substructures. Substructures are commonly determined by comparing the molecular structures of top matching compounds for structural similarities (42,46-53). Substructures representing the skeletal backbone of a molecule or its prominent functional groups can often be determined. Several structure determination methods involve the interpretation of mass spectra.

(3) **Spectral interpretation.** Applications of interpretive methods to analyze mass spectra have had mixed success. Techniques have included the use of high-resolution mass spectra (17), self-training algorithms (54-57), and learning machines (58-62). The Dendral project, which started in 1965 and concluded in 1980, was the most successful and ambitious project to date (63-65). It represented the first attempt to develop heuristic rules which represent chemical knowledge concerning structure stability and the fragmentation of molecules. It was also the first to use substructures as building blocks to elucidate structures. Various substructures were determined through spectral interpretation and then presented to a structure generator (GENOA) to generate possible molecular structures.

(4) **Structure generation.** Once all relevant substructures of an unknown compound are determined, they must be combined to produce the molecular structure. This is a straightforward yet constrained process. The structure generator must contain heuristics to comprehensively generate all plausible molecules yet prevent a combinatorial explosion by elimination of structures that are chemically impossible. Several structure generators have been developed (26,66). The most celebrated is the GENOA program (63,64) which allows overlapping and alternative substructural fragments to be identified as substructural constraints.

(5) **Assessment of determined structures.** Once molecular structures have been postulated, they are validated by predicting the spectral properties they should exhibit. Structure checking programs generate theoretical spectra from heuristics for the candidate compound and compare it with the measured experimental spectrum (64). If the two spectra are not similar, the candidate structure may be suspect.

MS/MS in Automated Structure Elucidation

The application of conventional mass spectrometry to structure elucidation has several drawbacks. The mass spectrum represents a superposition of products resulting from all reaction pathways a fragmentated molecular ion may follow. In addition, ion rearrangements and consecutive neutral losses complicate the mass spectrum making interpretation difficult. The abundance of some spectral peaks is small due to multiple consecutive neutral losses while others appear high if

that peak represents products from several fragmentation pathways. When electron-impact (EI) sources are used to fragment large molecules, the molecular weight information disappears as the intensity of the molecular ion peak decreases. Chemical-ionization (CI) sources obtain molecular weight information but leave little structural information.

Mass spectrometry/mass spectrometry (MS/MS) instruments provide additional capabilities over conventional mass spectrometry for structure elucidation. By selecting the parent ion and fragmenting it, information concerning an isolated portion of the molecule is obtained. Hence the moleuclar substructure associated with the parent ion in the normal mass spectrum may be determined. In this manner, structures that differ by only one spectral feature may be identified. Secondary fragmentation conditions are carefully controlled so that only first-order fragmentations occur. Daughter peaks, therefore, result from single neutral losses representing simple bond cleavages of the parent ion. A complete description of MS/MS spectrometry and the triple quadrupole mass spectrometer is readily available (67).

The substructure-property relationship determined by MS/MS is unique in that unambiguous substructural information is determined (Figure 1.1). In contrast to the interpretation of a mass spectrum to elucidate a molecule's structure, MS/MS allows substructures of the molecule to be deduced from MS/MS spectra. The molecular structure may then be generated by using a structure generator. Structure elucidation using MS/MS becomes an empirical determination as opposed to the interpretative methods used with conventional mass spectrometry.

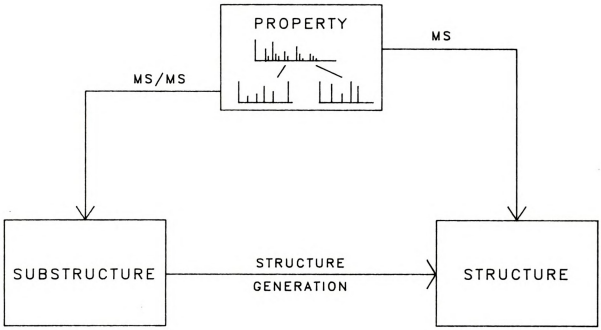


FIGURE 1.1. STRUCTURE–PROPERTY RELATIONSHIPS IN MASS SPECTROMETRY

The identification of the substructural components of a molecule eliminates the need for a large conventional mass spectra data base. To identify a conventional mass spectrum through spectral matching, the spectral library must contain the unknown compound (or a similar analog). The use of MS/MS data in spectral matching requires only a data base of spectra representing known substructures. This data base is relatively small, while the number of compound spectra needed in conventional mass spectrometry is impossibly large.

Another capability of MS/MS is the ability to determine the molecular formula by fragmenting parent ions of isotopic species (68). This eliminates the need for high resolution mass spectrometry and determines a piece of information essential to the structure generation process.

The use of MS/MS for structure elucidation was first proposed by Beynon in 1978 (68). The capabilities of MS/MS at that time (MIKES) were still immature. Poor resolution of daughter spectra (> 1 amu) and complex instrumentation hindered progress. In addition, an MS/MS instrument's ion path is complicated with several electrostatic lenses and varying longitudinal energies. Variation in peak intensities with instrumental parameters has traditionally caused problems for determining standard MS/MS operating conditions and has slowed the development of MS/MS libraries (69).

Increases in resolution were made as improved mass filter hardware became available (70). The triple quadrupole mass spectrometer and FT-MS provide unit mass resolution or better. Recently developed hybrid instruments (EBEB, BEQQ, BEQ) provide high mass resolution for at least one mass filter and atleast unit resolution for the other (71,72).

The technological advances in computer hardware have helped increase the reproducibility of MS/MS spectra. Those instrumental parameters along the ion path which affect the intensity of MS/MS peaks have been identified, and standard operating conditions determined. In addition, expert systems have been developed to control the instrument and to "tune" the ion path to obtain reproducible library quality spectra without human intervention (73,74).

The focus of this thesis is the design and development of several elements in an expert system for elucidating organic compounds using MS/MS data. The system proposed herein employs numerical reduction, substructure determination, and structure generation techniques to optimize the elucidation process using MS/MS data.

Since the daughter spectrum/substructure relationship is central to the structure determination process, the majority of this work describes the determination of substructures through spectral matching. The overall structure determination scheme using MS/MS data is described in chapter 2. An MS/MS information management system to support this process in presented in chapter 3. The development of a spectral

matching program to determine substructures from MS/MS data is presented in chapter 4. The effects of instrumental parameters on MS/MS spectra and the substructure determination process is described in chapter 5. Lastly, chapter 6 details the design and development of a data base to maintain structures and substructures, and chapter 7 suggests future work on the project.

References

1. Winston, P. H., **Artificial Intelligence**, Addison-Wesley Publishing Co., Don Mills, Ontario (1977).

2. Fleig, C. P., **Hardcopy**, August, p. 56 (1983).

3. McDermott, J., **Proc. First Ann. Nat. Conf. Art. Intell.**, 269 (1980).

4. Duda, R. O., Shortliffe, E. H., **Science**, 220 (1983).

5. Hayes-Roth, F., **Computer**, October, p. 263 (1984).

6. Hayes-Roth, F., **Computer**, September, p. 11 (1984).

7. Stefik, M., Aikins, J., Balzer, R., Benoit, J., Birnbaum, L., Hayes-Roth, F., and Sacerdoti, E., **Artificial Intelligence**, **18**, 135 (1982).

8. Gevarter, W. B., **IEEE Spectrum**, August, 39, 261 (1983).

9. Hayes-Roth, F., Waterman, D. A., Lenat, D. B., **Building Expert Systems**, Addison-Wesley Publishing Co., Don Mills, Ontario (1983).

10. Interview with Natalie Dehn, **Personal Computing**, June, p. 49 (1983).

11. Zadeh, L. A., **Fuzzy Sets Syst.**, 11, 3 (1978).

12. Dagani, R., **Chemical and Engineering News**, August 12, 7 (1985).

13. Zurer, P. S., **Chemical and Engineering News**, August 19, 21 (1985).

14. Cohen, C., **Electronics**, July 28, 101 (1983).

15. Elmer-DeWitt, P., **Time**, September 2, 44 (1985).

16. Milne, G. W., Heller, S. R., **Computer Assisted Structure Elucidation**, ACS Symposium Series, **54**, 26 (1977).

17. Hilmer, R. M., Taylor, J. W., **Anal. Chem.**, 51, 1361 (1979).

18. Venkataraghavan, R., Dayringer, H. E., Pesyna, G. M., Atwater, B. L., Mun, I. K., Cone, M. M., McLafferty, F. W., **Computer Assisted Structure Elucidation**, ACS Symposium Series, **54**, 1 (1977).

19. Heller, S. R., **J. Chem. Inf. Comput. Sci.**, **25**, 224 (1985).

20. McLafferty, F. W., Stauffer, D. B., **J. Chem. Inf. Comput. Sci.**, **25**, 245 (1985).

21. Sasaki, S., Kudo, Y., **J. Chem. Inf. Comput. Sci.**, **25**, 252 (1985).

22. Zupan, J., Denca, M., Hadzi, D., Marsel, J., **Anal. Chem.**, **49**, 2141 (1977).

23. Yamasaki, T., Abe, H., Kudo, Y., Sasaki, S., **Computer Assisted Structure Elucidation**, ACS Symposium Series, **54**, 108 (1977).

24. Suprenant, H. L., Reilley, C. N., **Computer Assisted Structure Elucidation**, ACS Sym. Series, **54**, 77 (1977).

25. Swenzer, G. M., Mitchell, T. M., **Computer Assisted Structure Elucidation**, ACS Sym. Series, **54**, 58 (1977).

26. Shelley, C. A., Woodruff, H. B., Snelling, C. R., Munk, M. E., **Computer Assisted Structure Elucidation**, ACS Sym. Series, **54**, 92 (1977).

27. Abe, H., Yamasaki, T., Fujiwara, I., Sasaki, S., **Anal. Chim. Acta**, **133**, 499 (1981).

28. Lowery, S. R., Huppler, D. A., Anderson, C. R., **J. Chem. Inf. Comput. Sci.**, **25**, 235 (1985).

29. Sasaki, S., Kudo, Y., **J. Chem. Inf. Comput. Sci.**, **25**, 252 (1985).

30. Heller S. R., **Anal. Chem.**, **44**, 1951 (1972).

31. Damen, H., Henneberg, D., Wiemmann, B., **Anal. Chim. Acta**, **103**, 289 (1978).

32. Lebedev, K. S., Tormyshev, V. M., Derendyaev, B. G., Koptyug, V. A., **Anal. Chim. Acta**, **133**, 517 (1981).

33. Lefkovitz, D., **J. Chem. Inf. Comput. Sci.**, **15**, 14 (1975).

34. Pesyna, G. M., McLafferty, F. W., in **Determination of Organic Structural Physical Methods**, **6**, 91, (1976).

35. Knock, B. A., Smith, I. C., Wright, D. E., Ridley, R. G., Kelly, W., **Anal. Chem.**, **42**, 1516 (1970).

36. Clark, H. A., Jurs, P. C., **Anal. Chim. Acta**, **132**, 75 (1981).

37. McLafferty, F. W., Stauffer, D. B., **Int. J. of Mass Spectrom Ion Phys.**, **58**, 139 (1984).

38. Malinowski, E. R., Howery, D. G., **Factor Analysis in Chemistry**, Wiley and Sons, New York, NY, 165 (1980).

39. McGill, J. R., and Kolwalski, B. R., **J. Chem. Inf. Comput. Sci.**, **18**, 52, (1978).

40. Isenhour, T. L., Kolwalski, B. R., Jurs, P. C., **CRC Crit. Rev. Anal. Chem.**, **3**, 1 (1974).

41. **Massart, D. L., Kaufman, L., The Interpretation of Analytical Chemistry by the use of Cluster Analysis,** Vol 65, Anal. Chem. and Its Applications, Wiley and Sons, New York, NY, (1983).

42. Willet, P., **J. Chem. Inf. Comput. Sci., 24,** 29 (1984).

43. Fetteralf, D. D., Yost, R. A., **Int. J. Mass Spectrom. Ion Proc., 62,** 33 (1984).

44. deHaseth, J. A., Insenhour, T. L., **Computer Assisted Structure Elucidation,** ACS Sym. Series, **54,** 46 (1977).

45. Rasmussen, G. T., Isenhour, T. L., **J. Chem. Inf. Comput. Sci., 19,** 179 (1979).

46. Dromey, R. G., **J. Chem. Inf. Comput. Sci., 18,** 222 (1978).

47. Adamsom, G. W., Cowell, J., Lynch, M. F., **J. Chem. Doc., 13,** 153 (1972).

48. Bawden, D., **J. Chem. Inf. Comput. Sci., 23,** 14 (1983).

49. Varkony, T. H., Siloach, Y., Smith, D. M., **J. Chem. Inf. Comput. Sci., 19,** 104 (1979).

50. Cone, M. M., Venkataraghavan, R., McLafferty, F. W., **J. Am. Chem. Soc., 99,** 7688 (1977).

51. Willet, P., **J. Chem. Inf. Comput. Sci., 25,** 114, (1985).

52. Synge, R. L. M., **J. Chem. Inf. Comput. Sci., 25,** 50 (1985).

53. Kudo, Y., Chihara, H., **J. Chem. Inf. Comput. Sci., 23,** 109 (1983).

54. Dayringer, H. E., McLafferty, F. W., Venkataraghavan, R., **Org. Mass Spectrom., 11,** 895 (1976).

55. Haraki, K. S., Venkataraghavan, R., McLafferty, F. W., **Anal. Chem., 53,** 386 (1981).

56. Kwok, K., Venkataraghavan, R., McLafferty, F. W., **J. Am. Chem. Soc., 95,** 4185 (1973).

57. Hippe, Z., **J. Chem. Inf. Comput. Sci., 25,** 344 (1985).

58. Bender, C. F., Shepherd, H. D., Kolwalski, B. R., **Anal. Chem., 45,** 617 (1973).

59. Isenhour, T. L., Jurs, P. C., **Anal. Chem., 43,** 20A (1971).

60. Jurs, P. C., **Anal. Chem., 43,** 1812 (1971).

61. Smith, D. M., **Anal. Chem., 44,** 536 (1972).

62. Martinson, D. P., **Applied Spectroscopy, 35,** 255 (1981).

63. Smith, D. H., **Anal. Chim. Acta., 133,** 471, (1981).

64. Carhart, R. E., Smith, D. H., Gray, N. B., Nourse, J. G., Djerassi, C., **J. Org. Chem.,** 1708 (1981).

65. Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., Lederberg, J., **Applications of Artificial Intelligence to Organic Chemistry: The Dendral Project,** McGraw Hill, New York, N. Y. (1980).

66. Carhart, R. E., Varkony, T. H., Smith, D. H., **Computer Assisted Structure Elucidation,** ACS Sym. Series, **54,** 126, (1977).

67. Yost, R. A., Enke, C. G., **J. Am. Chem. Soc., 100,** 2274 (1978).

68. Borogzadeh, M. H., Morgan R. P., Beynon, J. H., **Analyst, 103,** 613, (1978).

69. Dawson, P. H., Sun, W. F., **Int. J. Mass Spectrom. Ion Proc., 55,** 155 (1983).

70. Cook, R. G., Bush, K. L., Glish, G. L., **Science, 222,** 273 (1983).

71. McLafferty, F. W., **Accounts of Chem. Res., 13,** 33 (1980).

72. Chapman, J. R., **J. Phys. Ed.,** 365 (1978).

73. Wong, C. M., Lanning, S., **Energy and Technology Review,** Lawrence Livermore National Laboratory, February, p. 8 (1984).

74. Wong, C. M., Crawford, R. W., Barton, V. C., Brand, H. R., Neufield, K. W., Bowman J. E., **Rev. Sci. Inst., 54,** 996 (1983).

CHAPTER II

AN AUTOMATED STRUCTURE ELUCIDATION SYSTEM FOR MS/MS DATA*

Abstract

An automated system was developed to evaluate the structural information contained in mass spectrometry/mass spectrometry (MS/MS) spectra. The system employs several software tools to assist in the determination of unknown organic structures. These include tools to: 1) match conventional and MS/MS spectra, 2) assist in the determination of correlations between spectral characteristics and substructures, and 3) assist in the determination of structures from identified substructures. Correlations of MS/MS spectra with substructures are determined by matching MS/MS spectra with common mass parent ions and identifying substructures leading to ions with common spectral characteristics. Identified substructures are then combined using a constrained structure generator to postulate molecular structures. The scheme is totally empirical and does not assume that structural integrity is maintained in the ionization or fragmentation process; it does not require the ion structures to be identified.

*Note: This chapter is adapted from a preliminary draft of a manuscript written by the author of this thesis, to be published in the ACS Symposium Series entitled, "Artificial Intelligence Applications in Chemistry", with P. T. Palmer, C. F. Beckner, A. B. Giordani, H. R. Gregg, P. A. Hoffman, and C. G. Enke as coauthors.

Introduction

The development of mass spectrometry/mass spectrometry (MS/MS) has given the analyst a powerful tool for structure elucidation. The primary goal of this project has been to further develop triple quadrupole mass spectrometry (TQMS) as a tool for structure determination by developing a software system to organize MS/MS data, to aid in the discovery of MS/MS spectrum/substructure relationships, and to aid in the determination of a compound's structure from identified substructures.

The information and data presented in this chapter represents the culmination of several years of work by several different individuals. The structure elucidation project goals and methods were conceived by Chris Enke and Anne Giordani, and expounded upon by other project members. The multi-dimensional data base was designed and developed by Hugh Gregg (See Figure 2.1). The reference spectrum data base was designed and developed by Phil Hoffman. The data format for storing structural information was developed by Carl Beckner. Pete Palmer has been active in acquiring data for a MS/MS spectra data base. The phthalate data presented in this chapter were contributed by him. Kevin Hart has been active in implementing the molecular structure generator.

My contributions include the design and development of MS/MS spectral matching routines and the structure/substructure data base. In addition, I have combined the developed software tools into a cohesive,

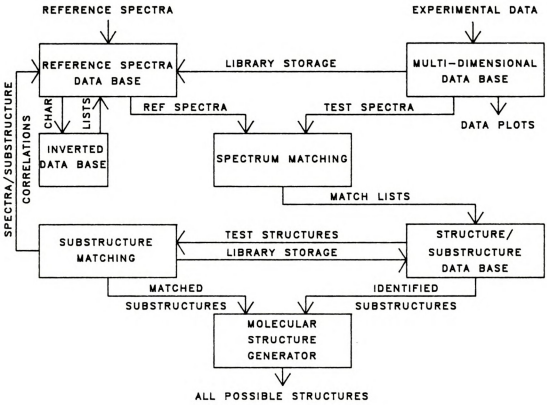


Figure 2.1 Software Tools for Structure Determination by MS/MS

interactive information management system system capable of aiding in the structure determination process.

The use of two-dimensional analytical instruments allows the experimenter the unique opportunity to categorize molecular substructures by their physical properties. In the case of MS/MS, the mass spectrum of a chosen parent ion, called a daughter spectrum, acts as a fingerprint in identifying a particular substructure (or substructures). MS/MS data are very clear: 1) daughter spectra reveal structural characteristics of isolated sections of the molecule, and 2) all masses in a daughter spectrum are simple cleavages or rearrangements from the parent ion. Hence MS/MS provides clear substructure-property relationships. The philosophy of the structure elucidation process is to employ a MS/MS instrument along with computer automated spectral interpretation to exploit this substructure-property relationship.

Data from this instrument were used in two different ways: 1) to develop substructure/spectrum correlations from the spectra of known compounds, and 2) to use the developed correlations to determine the substructures and overall structure of unknown compounds. We are in effect substituting a reference library of the substructures for a library of spectra of known compounds. If successful, this approach should allow determination of unknown compounds not previously studied by mass spectrometry.

The development of a reference library of MS/MS spectra and the substructures associated with each spectrum is a difficult preliminary

step to the determination of an unknown compound. The instrumental operating conditions must be carefully controlled to obtain substructures representative of all data in the daughter spectrum. In addition, all numerical and structural information for each compound must be correlated and stored in a data base where it can be easily and quickly retrieved for spectral matching.

In automating the structure elucidation process, several software tools were developed and integrated into a comprehensive system to acquire, store, match, and correlate the MS/MS data (Figure 2.1). The experimental data are placed into a user's data base, termed a multi-dimensional data base (MDDB). This data base was designed and developed by Hugh Gregg (1). The MDDB is local to each experiment and allows data inspection and massaging before inclusion into a MS/MS reference library. Once MS/MS data are introduced into the information management system, they may be plotted, compared with existing data, or stored in a reference data base for later referral.

Substructure determination is initiated by the spectral matching program which compares the experimental daughter spectra against those in a reference library and retrieves similar MS/MS reference spectra from the reference data base (2,3). By comparing an experimental daughter spectrum against a library of daughter spectra from the same mass parent, the unknown daughter spectrum is identified. The best matching reference daughter spectra are used to extract corresponding molecular structures or substructures from a structure related data base (4). By acquiring a daughter spectrum for every significant ion in the

conventional mass spectrum, the user can identify many of the substructures making up the complete unknown molecule. The process yields redundant and overlapping substructures while identifing the major substructures.

If relevant substructural data representing the unknown spectrum are not available, the complete molecular structures of the top matching daughter spectra are compared for the largest common substructure. This substructure matching process is currently performed manually using a program for plotting both molecular structures and substructures (5).

Neutral fragments (lost when the major substructures are formed) are identified by acquiring parent spectra of the highest m/z parent ion representing each identified substructure. These fragments are identified by matching daughter spectra (whose parent mass corresponds to the mass of the neutral lost) against a library of daughter spectra. This method ensures acquisition of substructural information from all parts of the unknown compound that can be represented by MS/MS spectra.

When all substructural information has been determined, the overlapping substructures are transferred to a constrained molecular structure generator called GENOA (6). GENOA postulates the number and identity of all possible, yet plausible, molecular structures of the unknown compound. If the number of possible molecular structures is too large, additional substructural information must be provided to limit the number of structural possibilities. This act may require obtaining additional MS/MS data or the addition of other spectral or non-spectral

information. This scheme ensures completeness in evaluating the structural possibilities of an unknown compound. The experimenter's chemical knowledge and intuition, however, still remains central to the elucidation process and crucial to its success. An example demonstrating the elucidation process will be presented later.

Development of an MS/MS Data Base

The concern over instrumental operating conditions stems from the need to reliably identify daughter spectra of molecular substructures and to distinguish them from all other daughter spectra. The latitude in the operating conditions meeting these criteria was tested by exhaustively collecting data under different experimental operating conditions for certain classes of compounds. (See chapter 5.)

The instrumental parameters experimentally determined as crucial in obtaining reproducible MS/MS spectra are collision energy and collision cell pressure. For this application it was critical that daughter spectra correspond only to substructures represented by single collisions. Therefore, the collision cell pressure used for acquiring reference spectra must ensure first-order fragmentations of the compounds. Brief kinetic studies were carried out for each class of compounds to determine the acceptable range of collision cell pressures for acquiring reference daughter spectra.

The optimum collision energy differs for each parent, and also for each daughter of that parent. The collision energy setting used was

based on the parent ion m/z value.

The procedure for acquiring MS/MS reference spectra parallels that for identifying an unknown compound. A daughter spectrum is obtained at every significant m/z value in the conventional spectrum and matched against a library of reference daughter spectra. The substructures associated with the top matching spectra are examined to determine their relevance to the known compound. Acquired daughter spectra having significant associated substructures are stored in the spectrum data base and linked to their respective substructures. Substructures not identified through spectra matching are identified by comparing the complete molecular structures corresponding to top matching spectra for the largest common substructural fragment. If the fragment represents a significant substructure in the known molecule, it is stored in the structure data base and the associated daughter spectrum is stored in the spectrum data base. Only daughter spectra representing significant substructures are saved for later referral.

MS/MS Spectra Data Base Format

There are two data bases present in our MS/MS information management system. One data base manages the MS/MS spectra, while the other manages the structures and substructures. The two data bases are logically linked together so that all information concerning a particular molecule or substructure is associated.

The MS/MS reference spectrum data base is capable of storing and correlating all types of MS/MS spectral data including parent, daughter, neutral-loss, and conventional mass spectra. All spectra for each compound are logically associated with that compound. This data base was designed and developed by Phil Hoffman (2). Redundant spectra, such as those taken under different operating conditions, are all associated with a single compound registry number thereby simplifying both the retrieval and maintenance of the data base information.

The most important design feature of the reference spectrum data base is the provision to generate and store inverted data. The data in the spectrum data base may be inverted upon a specified characteristic, such as m/z value, and then be retrieved using that characteristic. For instance, a data file inverted about the daughter m/z value will contain, for each m/z value, a list of reference daughter spectra. Hence all reference daughter spectra containing m/z 43.0 may be retrieved. When boolean algebra operations are performed on inverted data lists, the power of the design becomes apparent. When all reference daughter spectra containing peaks at 43.0 and 57.0 but not 119.0 are retrieved, the list of suitable reference spectra rapidly shrinks to a manageable size.

In addition to a daughter m/z value, spectral data may be inverted about molecular weight, empirical formula, and parent ion m/z value. This feature allows the matching program to prefilter candidates before retrieving candidate spectra thereby significantly reducing the overall spectral matching time. Over 30,000 conventional mass spectra

are currently stored in the spectrum data base as well as MS/MS spectra corresponding to specific classes of compounds.

Structure/Substructure Data Base Format

The structure data base was designed to contain both molecular structures and substructures (4). The MS/MS instrument specifically provides a substructure-property relationship where many daughter spectra may correspond to a single substructure. Hence, there is no logical link between the molecular structure and its associated substructures; unlike many existing structure data bases. There is, however, a logical link between the MS/MS spectra in the spectrum data base and the respective molecular structure and substructures in the structure data base. This link allows retrieval of structural information from the reference daughter spectra best matching the unknown spectrum. Structures present in the structure data base may be retrieved and drawn via substructure number, Chemical Abstracts Service number, or spectrum data base number.

The structures and substructures are stored unambiguously using the Morgan algorithm for encoding molecular structures via connectivity tables. The version of the algorithm implemented was that described by Wipke and Dyott (7) and includes representation for stereochemical isomers. The notation of the elements was expanded from the organic elements included in the original version to include all known elements. This notation was developed by Carl Beckner. Any molecule up to 128 atoms in size (excluding hydrogens) may be included in the data base.

The structure data base contains over 30,000 structures corresponding to the spectra in the MS/MS reference library as well as substructures corresponding to various reference daughter spectra.

Matching MS/MS Spectra

The MS/MS spectra matching program allows the chemist to match any MS/MS spectrum against either MS or MS/MS spectra in the reference spectrum data base (3). The program uses inverted data organized by m/z value to eliminate inappropriate reference spectra. The program determines the data base frequency of each significant peak in the experimental daughter spectrum and ranks the peaks in ascending order of frequency. Inverted lists of reference spectra containing each spectral peak are then retrieved and logically ANDed together to reduce the number of candidate reference spectra. Additional prefiltering of candidate spectra using molecular weight, parent ion m/z value, and empirical formula may be invoked to further reduce the number of candidate spectra. When matching daughter spectra, the parent ion m/z value usually serves as an adequate prefilter. The exception is the case where no similar daughter spectra of that parent ion are available. In this case, daughter spectra of higher parent ion m/z values may help deduce any substructures present. Until intensity-based matching is performed, the reference data base is not accessed and abundance values are not considered. This design considerably reduces the overall matching time and makes it practical to work with unabridged spectra.

Once the number of candidate reference spectra has been reduced

to reasonable size (25-100), intensity-based matching is performed to characterize the correspondence between the experimental and remaining candidate spectra. Several match factors describing the quality of the match are used to quantitatively characterize the match and to infer whether any substructures representing those in the experimental spectrum are present. The various match factors returned to the user are listed is Table 2.1.

The overall match factor (PT) is a combination of forward and reverse searching techniques. It takes into account the deviations in intensity of the sample spectrum peaks with respect to the candidate spectrum peaks and vice versa for all peaks in both spectra. The pattern correspondence match factor (PC) is a forward searching match factor which takes into account the intensity deviations of sample spectrum peaks with respect to the candidate spectrum peaks for peaks common to both spectra. This factor detects structural similarities, such as substructures, based on common spectral patterns.

Table 2.1. Match Factor Definitions

PT - An overall match factor that indicates how well the intensities of all the peaks in the two spectra match.

$$PT = (\Sigma Ys + Yr - 2*| Yr - Ys|) / (\Sigma Ys + \Sigma Yr) * 100$$

where $Yi = \log_2$ (Intensity/Total Ion Current) + SENS

PC - A pattern correspondence factor that indicates how well the intensity of the peaks in common match.

$$PC = (\Sigma Ys - | Yr - Ys |) / (\Sigma Ys) * 100$$

NC - The number of peaks common to both the candidate and unknown sample spectrum.

NS - The number of peaks remaining unmatched in the unknown sample spectrum.

NR - The number of peaks remaining unmatched in the reference spectrum.

IS - The percent total ion current of the sample 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.

NP - The number of peaks in common between the two spectra that are used in parabolic fitting of the quotient spectrum.

Chi^2 - The reduced chi-square value obtained from parabolic fitting of the quotient spectrum.

Substructure Identification

After the spectral matching process has been completed, substructures associated with the top matching daughter spectra are identified and retrieved. If relevant substructures are unavailable, the molecular structures of the top matching candidates are drawn and compared for common substructures. An heuristic program written by Dr. Craig Shelley (5) has been adapted for our computer system to display molecular structures and substructures from connectivity tables. Since molecular structures and substructures are stored in a unique form, the structure drawings facilitate visual comparison for commonalities. Although the process of substructure determination is currently performed manually, it will soon be automated using an atom-by-atom substructural search program.

Generation of Molecular Structures

The GENOA program is a constrained molecular structure generator resulting from the Stanford Dendral project (6) and is marketed by Molecular Design Ltd. (8). This program generates molecular structures using the overlapping substructural information obtained from the daughter spectrum/substructure relationship and the empirical formula of the compound. Additional spectral and non-spectral information from other sources may also be included. Heuristic rules determine whether a particular generated structure is chemically plausible, and whether or not it is retained. The advantage of the GENOA program is its ability to exhaustively produce all the plausible compounds given the generation constraints. This capability eliminates the possibility that the chemist might overlook any chemically possible compounds. In many cases the number and types of different structures produced will suggest missing pieces of structural data.

An essential piece of information required by GENOA is the empirical formula of the unknown compound. M+1 daughter spectrum data have been used instead of the high-resolution molecular-ion mass spectrum to assist in determining the empirical formula. The daughter spectrum of the M+1 isotope ion contains peak pairs at adjacent masses representing the C^{13} isotope mixture of the M+1 isotope fragment ion. The relative peak areas of these daughter pairs depends on the ratio of carbon atoms lost to carbon atoms retained by the M+1 ion. Hence the peak area ratios determine the number of carbon atoms present in the

compound. An existing program then calculates all possible empirical formulas from the molecular weight and number of carbons. The resulting reasonable empirical formulas are given to GENOA. This method was developed by Pete Palmer.

An Example: The Elucidation of Di-n-octylphthalate

To demonstrate the automated structure determination process, di-n-octylphthalate MS/MS spectra were acquired and treated as unknown spectra (Figure 2.2). Di-n-octylphthalate daughter spectra of m/z 149^+ and m/z 105^+ served to identify the phthalate substructure while daughter spectra of m/z 113^+ served to identify the alkyl groups. The results are presented in Table 2.2.

The m/z 149^+ daughter spectrum of di-n-octylphthalate was matched against a reference library of similar m/z 149^+ daughter spectra (Table 2.3). The top three matching spectra all correspond to the same molecular substructure, namely the phthalate substructure (Structure #1 in Figure 2.3).

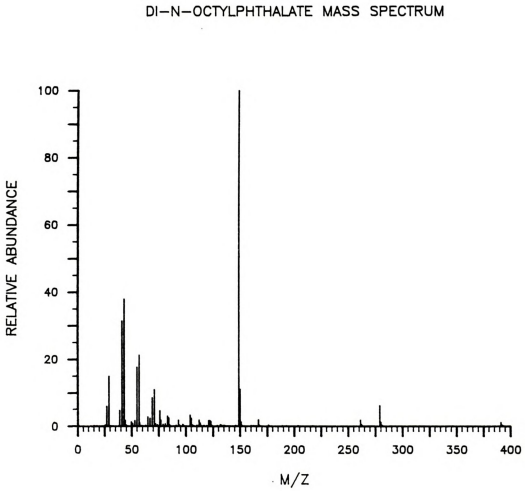


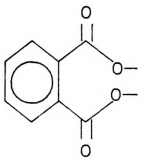
Figure 2.2 Di-n-octylphthalate Mass Spectrum

Table 2.2. Daughter Spectra of Di-n-octylphthalate

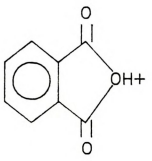
Parent Ion	Daughter Ion/Relative Abundance
149	65/4, 93/11, 121/9, 149/100
113	43/16, 57/79, 71/100, 84/17, 113/46
105	77/62, 105/100

Table 2.3. Match of 149* Di-n-octylphthalate Daughter Spectrum

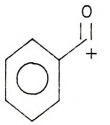
PT	PC	NC	NS	NR	IS	IR	Reg #	Name
100	100	6	0	0	0	0	15	Di-n-octylphthate
96	100	4	2	0	3	0	11	Dibutylphthalate
91	99	5	1	0	2	0	13	Dipentylphthalate
72	95	3	3	7	5	2	7	2-t-butyl-4-methylphenol
67	92	1	5	1	8	29	9	P-t-amylphenol
66	96	4	2	9	3	14	5	P-t-butylbenzyl alcohol
50	88	1	5	3	8	40	3	Benzyl-t-butanol
35	75	3	3	10	1	26	1	2-t-butyl-6-methylphenol



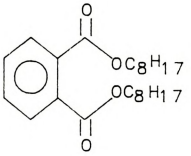
I



II



III



IV

Figure 2.3 Di-n-octylphthalate.

I) Molecular Substructure. II) 149+ Ion Structure.

III) 105+ Ion Structure. IV) Molecular Structure.

The ion structure represented by these daughter spectra (Structure #2 in Figure 2.3) is not identified during the elucidation process. Instead, the molecular substructure is associated with the daughter spectra and stored in the structure data base. The elucidation process is totally empirical and does not assume that structural integrity of an ion is maintained in the ionization or fragmentation process. As a result, the ion structures need not be identified.

The compounds yielding the top three daughter spectra are di-n-octylphthate, dibutylphthalate, and dipentylphthalate. Although the daughter spectra represent the same substructure, they are not identical. Different NS values amongst the three candidates indicate that the three spectra contain different spectral peaks. It is important that the spectral matching program properly groups these spectra together and ensures a substantial difference between the overall match factors of these spectra and those corresponding to unrelated substructures. The difference between the overall match factor of di-n-octylphthalate and the best matching daughter spectra corresponding to a different substructure is 19. Since the overall match factor range is 0-100 and the variance among similar daughter spectra is 9, a value of 19 represents good separation. This daughter spectrum corresponds to a substructure of 2-t-butyl-4-methylphenol. The need for multiple reference daughter spectra representing one substructure still remains since daughter spectra vary for different compounds and under different conditions.

The m/z 105^+ daughter spectrum of di-n-octylphthalate was matched against a reference library of similar m/z 105^+ daughter spectra and the results presented in Table 2.4. Again, the top three matching spectra all correspond to the same phthalate substructure (Structure #1 in Figure 2.3). In this case the daughter spectra are highly similar; all three contain the same spectral peaks, only the intensity patterns are different (NR, NS, IS, and IR for the three are all zero). Better clustering is probably due to the greater stability of the ion structure yielding these daughter spectra (Structure # 3 in Figure 2.3). Note again the large difference in overall match factor values (25) between daughter spectra representing the correct substructure and that of the next best match.

Table 2.4. Match of 105+ Di-n-octylphthalate Daughter Spectrum

PT	PC	NC	NS	NR	IS	IR	Reg #	Name
100	100	2	0	0	0	0	16	Di-n-octylphthalate
94	94	2	0	0	0	0	14	Dipentylphthalate
91	91	2	0	0	0	0	12	Dibutylphthalate
66	89	2	0	2	0	31	10	P-t-amylphenol
60	80	2	0	2	0	20	8	2-t-butyl-4-methylphenol
49	54	1	1	4	30	20	4	Benzyl-t-butanol
47	56	1	1	3	30	29	6	P-t-butylbenzyl alcohol
37	51	1	1	3	30	52	2	2-t-butyl-6-methylphenol

The parent spectrum of m/z 149$^+$ was used determine the alkyl groups attached to the phthalate substructure (Figure 2.4). The largest ion (149$^+$) associated with the phthalate substructure was used since it will yield neutrals corresponding to the groups attached to the complete substructure and not include pieces of the identified substructure. The parent spectrum has 4 major (non-isotopic peaks) at m/z 167, 260, 279, and 390. The neutral corresponding to a loss of 18 (167-149) is water. The neutral corresponding to the loss of 130 (279-149) is $C_8H_{17}OH$ which may represent an alkyl group. The neutral corresponding to the loss of 113 (262-149) is the C_8H_{17} radical which confirms the presence of a C_8H_{17} alkyl group. The neutral corresponding to the loss of 241 (390-149) is the $C_8H_{17}OC_8H_{17}$ radical; a rearrangement product.

To determine the branching of the C_8H_{17} group, the daughter spectrum of m/z 113$^+$ was matched against a library of daughter spectra. The alkyl group was found to be unbranched. Hence the alkyl groups n-C_8H_{17}, and n-$C_8H_{17}OH$ were used in conjunction with the phthalate substructure for generating possible molecular structures.

The last piece of information required is the empirical formula. To determine the empirical formula, the daughter spectrum of the M+H+1 peak in the CI di-n-octylphthalate spectrum (m/z 392) was obtained (Figure 2.5). The relative peak areas of adjacent peak pairs at m/z 149 and 150 is 2:1 indicating that the M+1 ion is twice as likely to lose a C^{13} atom as retain it. The ratio of the number of carbon atoms lost to those retained is, therefore, 2:1. Since the identified phthalate substructure contains 8 carbons, the unknown compound

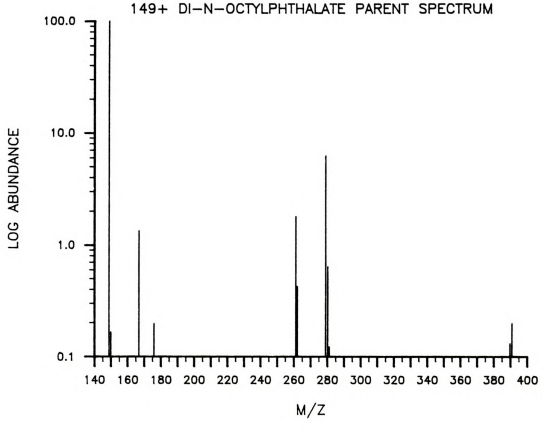


Figure 2.4 149+ Di-n-octylphthalate Parent Spectrum

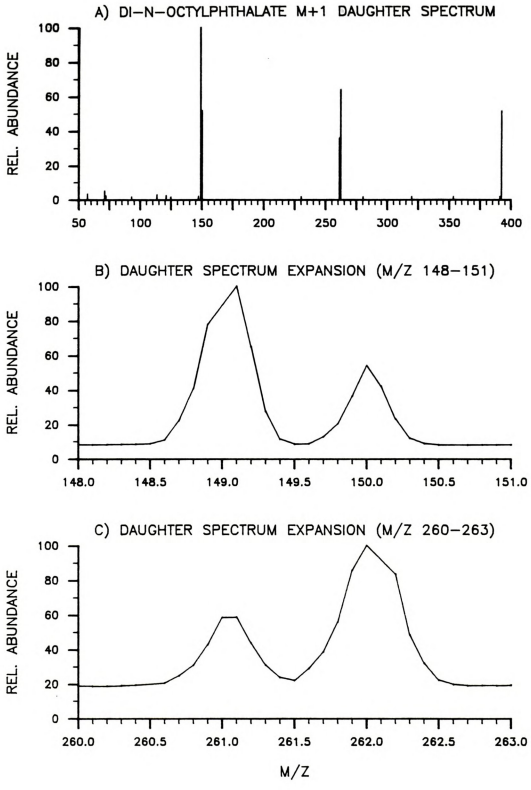


Figure 2.5 149+ Di-n-octylphthalate M+1 Spectra

(di-n-octylphthalate) must contain 24 carbon atoms and the empirical formula must be $C_{24}O_4H_{38}$.

Armed with the phthalate substructure, the two alkyl substructures, and the empirical formula, we are now ready to generate all plausible molecular structures. The oxygen in the $C_8H_{17}OH$ group is allowed to overlap with either terminal phthalate oxygen. With this information, GENOA constructs only one molecular structure (Structure #4 Figure 2.3) and it is the correct structure. The number of generated structures depends on the completeness of the information provided. If the branching of the alkyl group had not been specified, 89 different structures would have been generated. The identities of these generated structures, however, would provide clues as to the needed information. In cases where MS/MS information cannot determine a unique result, additional spectral and non-spectral information may be given to GENOA as structural constraints.

Conclusions

The automated structure determination system of software tools for aiding in the elucidation of organic structures from MS/MS data is now at a stage where the chemist can actively apply it to real elucidation problems. Nearly all of the software tools have been developed and integrated into a comprehensive, interactive system. The system has been successfully used to develop daughter spectra/substructure correlations and to extend MS/MS data bases. Preliminary results from applying the system to structure determination

problems have been very encouraging. An expert system is currently being implemented to oversee the entire structure determination process by examining results from the software tools and suggesting further experimentation.

References

1. Gregg, H. R., Hoffman, P. A., Enke, C. G., Crawford, R. W., Brand, H. R., Wong, C. M., **Anal. Chem.**, **56**, 1121 (1984).

2. Hoffman, P. A., Enke, C. G., presented at 31st Annual Conference on Mass Spectrometry and Allied Topics, Boston, MA (1983); bound p. 556.

3. Cross, K. P., Enke, C. G., **Computers and Chemistry**, in press.

4. Cross, K. P., Beckner, C. F., Enke, C. G., in preparation.

5. Shelley, C. A., **J. Chem. Inf. Comput. Sci.**, **23**, 61 (1978).

6. Carhart, R. E., Smith, D. H., Gray, N. B., Nourse, J. G., Djerassi, C., **J. Org. Chem.**, **46**, 1708 (1981).

7. Wipke, T. W., Dyott, T. M., **J. Am. Chem. Soc.**, **96**, 4834 (1974).

8. Molecular Design Ltd., 1122B Street, Hayward, CA 94541.

CHAPTER III

The Development of a Mass Spectra/Mass Spectra Information Management System

Introduction

To determine unknown molecular structures using MS/MS data a system was needed to handle the additional dimension of data provided by MS/MS instruments. The large amount of data generated using computer-controlled instrumentation has inspired the development of commercial information management systems for organizing and storing relevant information. Many spectroscopic instruments commonly contain data acquisition and reduction systems. Laboratory information management systems (LIMS) capable of integrating data acquired from several different laboratory instruments have also become quite popular (1).

Since our research group has access to three MS/MS instruments, an information management system was needed to acquire, view, and compare MS/MS data obtained from different instruments and different laboratories. The MS/MS information management system was developed to move information onto a central minicomputer and to transform spectral data into a common format. Various software tools allow the user to massage experimental data, to view MS/MS data in different two-dimensional planes, to store MS/MS spectra in a reference data base, and to match MS/MS data against a reference library (Figure 3.1).

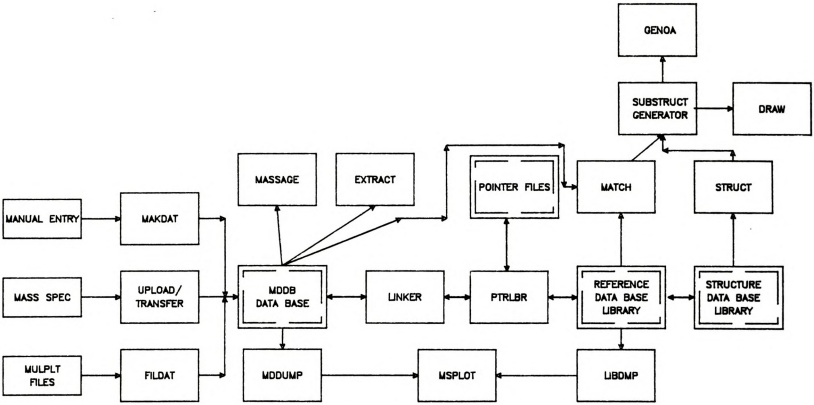


Figure 3.1 MS/MS Information Management System

Dedicated microprocessors are used for acquisition of data from MS/MS instruments and their control. The information management system was designed to manage MS/MS data after it had been acquired via the microcomputer attached to each instrument. The next four sections of this chapter focus on a description of the various tools developed to manage MS/MS data and to aid in structure elucidation.

Data Base Designs

At the heart of the information management system are two data bases that handle experimental and reference MS/MS data. All other components of the system use these data bases to retrieve or operate on the MS/MS data. The philosophy of the system is to separate those data used for general reference from raw experimental data which have recently been acquired but not yet reduced. The division of MS/MS data based on its functionality allows each data base to be separately optimized for its own purpose.

The first data base, termed the multi-dimensional data base, accepts data from an MS/MS instrument microcomputer and allows the user to view and massage the raw data. This data base was designed by Hugh Gregg in conjunction with a group at Lawrence Livermore National Laboratories to contain the results of multivariant MS/MS experiments performed by individual users (2). It was designed to contain data for a set of related experiments as opposed to containing large amounts of reference information. This design enables the experimenter to perform complex, multivariant experiments without having the order of data taken

affect the convenience of correlating any two variables after the experiment. A single experiment can be used for a variety of interpretations. A complete description of the multi-dimensional data base may be found in the doctoral dissertation of Hugh Gregg (3).

The second data base, termed the reference data base library, was designed by Phil Hoffman to efficiently store large amounts of MS/MS data and to enable rapid retrieval of library spectra (4,5). The data base management program, termed PTRLBR, was tailored to the DEC RSX-11M operating system to obtain maximum speed when retrieving spectra through memory mapping, indexed spectrum retrieval, and linked list data storage (6,7).

In the library, all MS/MS spectra associated with each compound are correlated. Daughter spectra of different parent ions are correlated with their primary spectrum in a hierarchical format. Different primary spectra of the same compound are also correlated, enabling spectra taken under different operating conditions, (such as with different ion sources), to be easily retrieved once the compound of interest is identified. Such a data base was essential for representing MS/MS spectra and enabling fast, interactive searching of such spectra.

Inverted lists of spectral data are generated to allow retrieval of reference spectra based on the specific characteristic about which the file was inverted (mass, empirical formula, etc.) (8). Combinations of these lists are termed pointer files. Pointer files enable prefiltering of the data by applying Boolean algebra to obtain a list of

compounds with specific characteristics. For example, one may select and retrieve all the compounds in the data base that have a particular molecular weight, empirical formula, or specific combination of mass spectral peaks. The structure of this data base is described in the doctoral dissertation of Phil Hoffman (9). The format of the data in the reference library and the spectral characteristics of the library are detailed later in this chapter.

Data are entered into the MS/MS information management system by placing the data in a multi-dimensional data base. Data may be entered into a multi-dimensional data base in several ways. The program MAKDAT allows manual data entry or entry using a MULPLT format data file (10). The program FILDAT enters data into multi-dimensional data base from a floppy disk written by a FORTH microcomputer. The most common mode of data entry is by uploading MS/MS data directly onto a minicomputer from the microcomputer controlling the MS/MS instrument. This process requires little user intervention. Movement of the data to the appropriate minicomputer, however, may be required to perform structure determination.

Computer Architecture

The computer network present in our laboratory is detailed in Figure 3.2. Two triple quadrupole instrument computers (connected to the ENKE TQMS and EL-400TQ-3) and a magnetic time-of-flight instrument (BTOF) computer are networked such that information from all three instruments may be easily transferred to a central minicomputer for

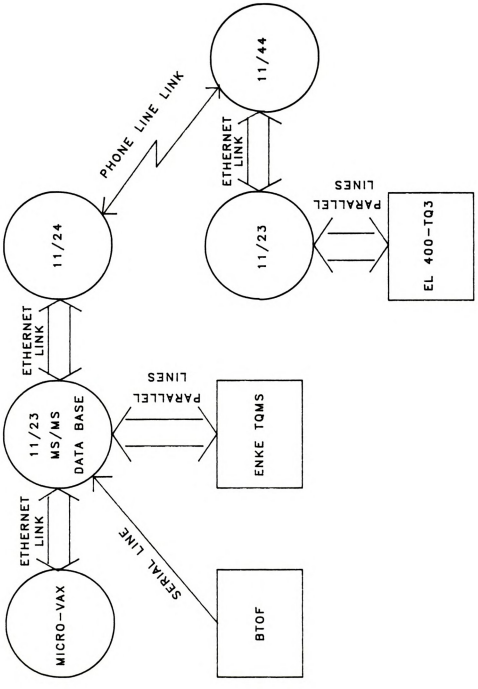


Figure 3.2 MS/MS Computer Network

reduction and comparison. The PDP-11/23 containing the reference data base library, termed the MS/MS 11/23, is the central computer in the data network. Although the other PDP-11 computers contain specific information management software, any use of the reference data base library must be performed on the MS/MS 11/23. The reference data base library and all information management software except GENOA are present on the MS/MS 11/23. The GENOA software is present only on the micro-VAX computer.

The microcomputers dedicated to the ENKE triple quadrupole mass spectrometer transfer spectral data on parallel lines directly onto the MS/MS 11/23. The microcomputer controlling the BTOF instrument uses a serial line to more slowly transfer spectral data onto the MS/MS 11/23. In both cases, the data are transformed and placed into a multi-dimensional data base as the spectra enter the MS/MS 11/23.

Data from the Extranuclear EL-400-TQ3 instrument are uploaded to a PDP-11/23 physically adjacent to the instrument (in the Biochemistry building). Data present in a multi-dimensional data base are then transferred to the Biochemistry PDP-11/44 computer through a DECNET Ethernet link. Spectral data are passed via modems through phone lines from the Biochemistry 11/44 to the PDP-11/24 in the Chemistry building. Finally the data are moved to the MS/MS 11/23 via a DECNET Ethernet link.

Although this process seems complex, once standard procedures were established, the indirect routing of MS/MS data through four

minicomputers merely became inconvenient. The rate-limiting step is the transfer of data at 120 bytes per second over a phone line between the Biochemistry and Chemistry buildings. As the campus wide network (MSUNET) becomes available, the phone link will be replaced with a high speed link between the two buildings.

Software Tools for Data Manipulation

The transfer of data from the microcomputer controlling the MS/MS instrument to the minicomputer is performed by a program called UPLOAD (11). UPLOAD was written by Phil Hoffman in MACRO-11 to quickly move data between the two computers using either parallel lines or a serial line. The data are transferred as data packets to allow for error checking and thereby ensure the integrity of the received data. The form of the data, however, remains unchanged.

When UPLOAD has successfully moved the raw data onto a minicomputer, the program TRANSFER is automatically activated to enter the spectral data in a multi-dimensional data base. For each spectrum, the data points are scaled and the instrumental operating parameters and their values are identified. TRANSFER looks at the headers of all the spectra to determine which instrumental parameters varied during the course of the experiment and which did not. Those parameters that did not change are identified as static parameters and their values are stored once in the multi-dimensional data base. Those parameters that varied throughout the course of the experiment are termed dynamic parameters. Their values are stored with each spectrum in the

multi-dimensional data base. In addition, TRANSFER selects the names for each parameter using the multi-dimensional data base dictionary. These parameter names vary with the particular MS/MS instrument and are defined on each instrument's dedicated microcomputer. Bruce Wilson developed the software to create and manage the multi-dimensional data base dictionary (12).

Once data are present in a multi-dimensional data base it may be massaged or plotted. The MASSAGE program includes scan averaging, adding, background subtraction, normalization, and noise elimination routines. In addition, scans may be commented, deleted, merged into other data bases, or viewed in a tabular format. For massive data reduction, a macro function is available which allows an individual function, such as scan averaging, to be repeated many times. For instance, a macro may be invoked to average data from every five consecutive scans until a total of 100 averages for 500 spectra have been performed. This flexible function provides a powerful tool for reducing large amounts of data into a concise, meaningful form. Normalization routines available include normalization to the base peak of the spectrum, the total ion current, a peak of the user's choice, or the second largest spectral peak. The last case is useful for normalization of MS/MS spectra where a large parent peak may dominate the daughter spectrum.

The EXTRACT program, developed by Hugh Gregg, allows the user to graphically observe how a spectrum changes as the parameters of the instrument vary (13). For example, if an experiment varying the

collision cell pressure was performed, the user may graphically display the abundance of each mass spectral peak as a function of pressure. Other types of graphical output include ion chromatograms, collision energy breakdown curves, and ion transmission curves. Graphical results demonstrating this program are presented in chapter 5.

Two other programs were written to display data present in a multi-dimensional data base. The program MDDUMP creates tabular listings of data. Data base comments and static parameters are listed followed by the dynamic parameters and spectrum for each selected spectrum. MDDUMP is also used by the program MSPLOT to retrieve MS/MS data for graphical output.

MSPLOT allows the user to plot multiple MS/MS spectra on a single page without learning to run both MULPLT and RASTER programs (See Figure 3.3.) (10,14). The program queries the user for the spectra in a multi-dimensional data base and specific plot characteristics, and then outputs the spectra to a selected graphics device. It supervises the fundamental tasks MDDUMP, LIBDUMP, MULPLT, and RASTER which actually acquire and plot the data thereby providing an easy method for quickly displaying related spectra.

The LINKER program serves as the datapath between multi-dimensional data bases and the reference data base library. It allows new experimental data from a multi-dimensional data base to be archived or existing data in the reference data base library to be updated. It functions independently of the software managing the

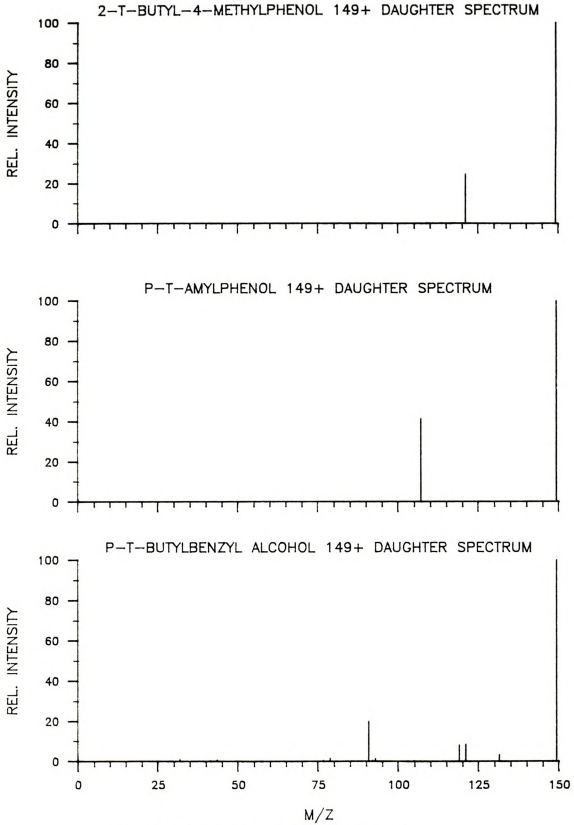


Figure 3.3 Example of MSPLOT Output

reference data base library (PTRLBR), thereby allowing noninteractive modification of the library to occur at a later time. Hence the library is not "down" for long periods of time while being updated. The LINKER does this by creating data packets containing the static and dynamic instrument parameters and data for each spectrum. The update or addition command is also included in the packet so that when PTRLBR picks up the data packet it knows what to do with it. In addition, the LINKER also allows data to be extracted from the library and placed into a multi-dimensional data base. This capability allows the creation of multi-dimensional data bases that are subsets of the reference data base library.

Two programs, LIBDMP and LIBIDX, allow quick and easy access to the data in the reference data base library. The program LIBDMP extracts MS/MS data from the library for tabular listings or spectral plotting using MSPLOT. The program LIBIDX extracts data based on selected spectral characteristics using an inverted data file (5). For example, the molecular weight pointer file may be used to select those spectra of compounds with molecular weights between 100 g/mole and 130 g/mole and then to place that spectral information into an output file.

Software Tools for Structure Determination

The major reason for developing the MS/MS information management system was to facilitate the structure determination process. Central to the structure determination process is the ability to identify substructures (often represented by daughter spectra) by comparing

measured MS/MS spectra against a reference spectrum library. The MATCH program allows comparison of experimental spectra with reference spectra contained in the reference data base libary or in another multi-dimensional data base. MATCH is used to identify experimental MS/MS spectra and to help determine any substructures that are associated with that spectrum. Maintenance of the substructure-property relationship determined by MS/MS is the primary function of the information management system. Experiments demonstrating the substructure determination process were described in chapter 2.

MATCH also allows spectra contained in a multi-dimensional data base representing a single experiment to be compared with each other. This operation allows comparison of spectra taken under different experimental conditions. The concern over instrumental operating conditions stems from the need to reliably classify daughter spectra of identical parent ions and to distinguish them from daughter spectra of all other parent ions. Experiments determining the dependence of the spectral matching program on instrumental operating conditions are presented in chapter 5.

The storage and maintenance of molecular structures and substructures is handled by the STRUCT program. This program maintains the substructure-property relationship determined through MS/MS by correlating the MS/MS spectra in the reference data base library with their associated structures and substructures in the structure data base. The progression from spectra to substructure is the first conclusive step in the structure elucidation process. Only through the

empirical identification of molecular substructures can the process of structure determination using spectral data become automated. A detailed description of the STRUCT program and structure data base is presented in chapter 6.

When there are no substructures associated with the daughter spectra matching an unknown daughter spectrum, the operator must determine the substructure(s) the daughter spectrum represents. This is done by comparing the complete molecules of compounds yielding matching daughter spectra. This substructure determination process is most prevalent when the substructure data base is initially being developed.

In the future, an atom-by-atom substructure searching program will compare the complete molecular structures for common substructures. Until then, substructures must be determined by manually comparing molecular structures. An heuristic structure drawing program was acquired from Craig Shelley at Eastman Kodak (15) and adapted to our system. The heuristics of the program displays molecules in a conventional fashion and minimizes overlapping atoms and atom crowding. By representing structures in a consistent manner, commonalities of the structures are easily perceived. These commonalities are candidates for the substructure(s) giving rise to the matching MS/MS spectra. In either case the operator must decide whether the determined substructural information is significant, relevant, and representative of the unknown daughter spectrum. Examples of output using the DRAW program are presented in chapter 6.

Once all the substructures from the MS/MS data have been determined, they are given to the structure generator (GENOA) (16) residing on the micro-VAX computer. GENOA accepts overlapping and redundant substructures regardless of origin and combines them with the empirical formula of the compound to produce plausible molecular structures. When a large number of structures are generated, additional structural information must be supplied. The nature of the generated molecules elucidates the ambiguous portions of the unknown molecule. This information can be used to suggest additional spectroscopic and non-spectroscopic experiments.

The Storage of MS/MS Data

In the past, the immense volume of data contained in mass spectral data bases has forced the compression of spectra to obtain efficiency in both the storage and retrieval processes. This has resulted in the use of compression algorithms ranging from simple abundance based compressions, where peaks below a certain threshold are eliminated (17-19); to mass-windowed abundance schemes, where only a few abundant peaks are retained in each 14 amu window (20-23); to statistical uniqueness schemes, where assigned probability values determine if the peak is retained (24-28).

In MS/MS data bases even more data are present as several related MS/MS spectra are associated with each compound. Secondary mass spectra contain many small peaks whose significance cannot be empirically or statistically determined. Hence any data compression

becomes difficult since the amount of structural information lost cannot be readily ascertained (29,30). Additional storage problem arise since intensity values must represent the large dynamic range (eight orders of magnitude) of abundances possible in MS/MS spectra.

As a result, the format of the data in the reference data base library becomes important for both storage considerations and matching procedures. The MS/MS spectrum is first normalized to the total ion current. The m/z value of each peak is retained to 1/8 amu resolution, which is near the limit of our MS/MS instruments. The m/z value is stored as a 16-bit integer to conserve storage space. The $\log_2$ of the normalized individual ion abundance is added to 30 and then multiplied by 1024 as shown below.

$$I = 1024*((\log_2 \text{Intensity/TIC}) + 30) \qquad (3.1)$$

McLafferty earlier demonstrated that the log normal distribution of peak abundance values is linearly related to the frequency of those abundance values (31). Hence abundance values (which have a large dynamic range in MS/MS) were efficiently packed in the spectral library using the $\log_2$ function. Intensity values are then suitable for matching when retrieved since log weighting is already performed. Adding a value of 30 to the normalized log intensity value ensures retention of the entire dynamic range of the initial MS/MS spectrum. Lastly, a multiplication factor of 1024 allows the intensity value to be stored as a 16-bit binary integer.

Mass Spectrometry Data Base Characteristics

The mass spectra incorporated into our reference data base library were acquired from a 1974 Wiley Interscience Registry of Mass Spectral Data magnetic tape (32). A considerable amount of effort was spent "cleaning up" these spectra and converting them into a usable format. In addition, molecular structures for the majority of the spectra were entered into a structure data base by a group of undergraduates. This structure entry process required over 500 man hours. The result was a good, clean library of over 30,000 mass spectra and their associated molecular structures.

With a large number of compounds it is difficult to comprehensively characterize the spectra in the reference data base library. Several generalities, however, can be made. The 30,000 plus mass spectra in our library contain over 0.75 million mass spectral peaks greater than 1% total ion current intensity. Molecular weights of the compounds range from 2.0 g/mole to 1674.0 g/mole. All of the spectra represent EI mass spectra acquired on low resolution mass spectrometers. Most spectra are of common organic compounds that range from long chain hydrocarbons to complex ring structures. There are only a few spectra of polymers and pharmaceutical drugs present.

Frequency distributions of the data provide a closer look at the library characteristics. The frequency distribution of the molecular weights of compounds in the library is shown in Figure 3.4. Although the

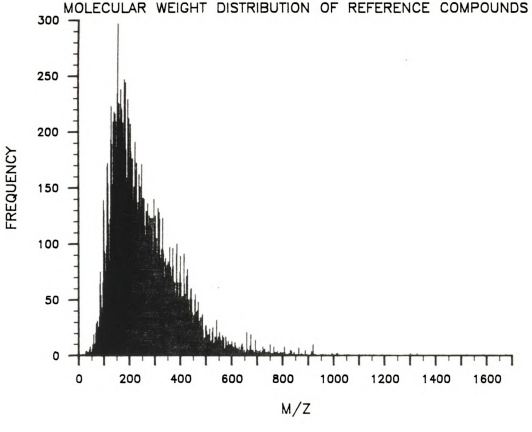


Figure 3.4 Molecular Weight Distribution of Compounds in the Reference Data Base

molecular weight range is large, the majority of compounds weigh between 100 and 400 g/mole. The most frequent molecular weight is 154 g/mole. Compounds at this weight represent functional groups of alkanes, alkenes, alkynes, acids, esters, ketones, alcohols, ethers, phenols, substituted benzenes, silanes, phosphates, chlorides, thiols, nitriles, azines, furans, bicyclic compounds, and a few drugs.

Statistical Occurrence of Mass Values in Mass Spectra

The frequency distribution of individual mass spectral peaks with abundances greater than 1% total ion current is illustrated in Figure 3.5. The distribution of peaks is similar in shape to the molecular weight distribution. The elimination of small peaks (< 1% total ion current) provides an uninhibited look at the significant peak statistics in the data base. Though there is atleast one peak present at nearly every mass value, it is clear that certain mass values are more common than others: m/z 39 is far more common than m/z 48. The presence of several homologous ion series causes frequency increases every 14 amu and results in "bumps" present throughout the distribution.

Several spectral matching programs have used probabilities of mass spectral peak occurrences for spectral compressions (24,25,33). The most notable being the Probability Based Matching program developed by McLafferty at Cornell University (25). These methods use the log function to linearize the frequency distribution of mass values and then assign probability values to individual spectral peaks based on their statistical occurrence. Such a frequency distribution for our library

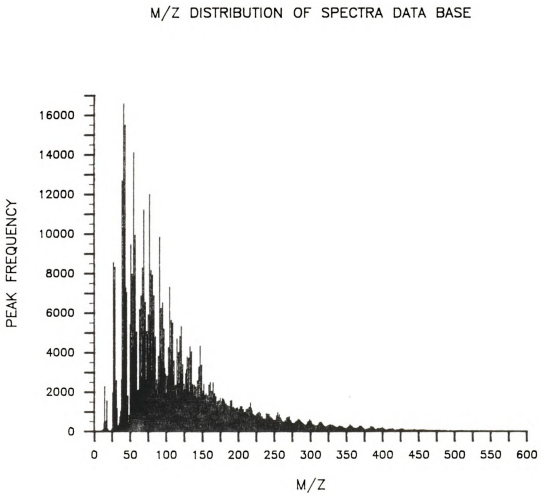


Figure 3.5 Frequency Distribution of Spectral Peaks in the Reference Data Base

is presented in Figure 3.6.

While the log distribution for our library of mass spectral peaks is approximately linear to m/z 800, beyond that the linear relationship no longer holds as fewer compounds are present at higher molecular weights. Therefore, spectral matching programs that use probabilities become dependent on the statistics of the mass spectral library. In addition, the probability values of spectral peaks in the reference library are difficult to update as the nature of the library changes. Since statistics for the occurrence of MS/MS spectral peaks have yet to be compiled, applying statistical methods to MS/MS spectral libraries still remains undeveloped.

For these reasons the spectral matching program for MS/MS data does not use assigned probability values. Instead, the actual empirical frequency of the peak in the reference data base library is used to help reduce the number of candidate reference spectra. A complete description of the MS/MS matching program is presented in chapter 4.

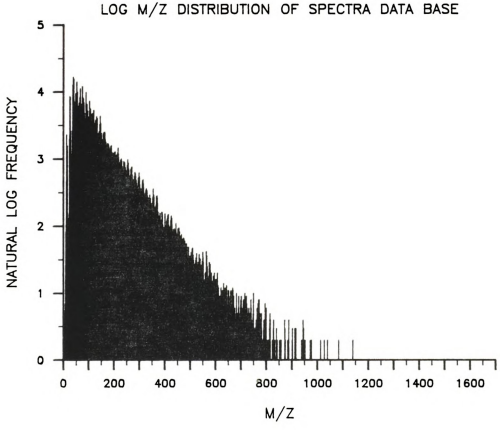


Figure 3.6 Log Frequency Distribution of Spectral Peaks in the Reference Data Base

Statistical Occurrence of Abundance Values in Mass Spectra

When matching mass spectra, it is common practice to weight mass spectral peaks in some fashion (22,25,34). Hence it is instructive to examine the distribution of abundance values for mass spectra in the reference data base library. Since it is impractical to record the frequency of each intensity value in each mass spectrum, abundance bins representing the different intensity ranges were identified. These bins accumulate the frequency total for each intensity range as the data base is examined.

Previous work indicated a log-log relationship between an abundance value and its frequency of occurrence (31). Therefore, the abundance bins limits were increased exponentially as the abundance values increased. The thirty abundance bins defined are presented in Table 3.1.

Table 3.1. Abundance Bins (Percent Total Ion current)

Bin Number	Abundance Range
1	0 to 1.9E-7%,
2	1.9E-7% to 3.7E-7%,
3	3.7E-7% to 7.5E-7%,
4	7.5E-7% to 1.5E-6%,
5	1.5E-6% to 3.0E-6%,
6	3.0E-6% to 6.0E-6%,
7	6.0E-6%. to 1.2E-5%,
8	1.2E-5% to 2.4E-5%,
9	2.4E-5% to 4.8E-5%,
10	4.8E-5% to 9.5E-5%,
11	9.5E-5% to 1.9E-4%,
12	1.9E-4% to 3.8E-4%,
13	3.8E-4% to 7.6E-4%,
14	7.6E-4% to 0.0015%,
15	0.0015% to 0.003%,
16	0.003% to .0061%,
17	0.0061% to 0.012%,
18	0.012% to 0.024%,
19	0.024% to 0.049%,
20	0.049% to 0.10%,
21	0.10% to 0.20%,
22	0.20% to 0.39%,
23	0.39% to 0.78%,
24	0.78% to 1.6%,
25	1.6% to 3.1%,
26	3.1% to 6.2%,
27	6.2% to 12.5%,
28	12.5% to 25%,
29	25% to 50%,
30	50% to 100%.

The frequency distribution of abundance values in the reference data base libary is presented in Figure 3.7. When the library was examined, no spectral peaks were found below an intensity of 1.9 X 10^{-4}% total ion current. Abundance bin numbers one through eleven were empty. The shape of the distribution indicates that the most common peak abundance value is 0.2% total ion current. If a peak's abundance varies from this value, its frequency of occurrence quickly drops.

The log frequency distribution of abundance values illustrates the details of the lower frequency values (Figure 3.8). These results parallel the earlier findings of McLafferty who illustratd linear log-log behavior down to intensity values of 0.1% base peak (31). That study, however, did not present a distribution of intensity values lower than 0.1% base peak.

In Probability Based Matching, McLafferty assigns a probability value between one and five to a peak based on its abundance (25). In a similar effort, Lebedev uses the $\log_2$ of the peak abundance to weight peaks during spectral matching (34). It was from these studies that it was decided to weight the peaks in the MATCH program by $\log_2$ of the abundance. This weighting, which de-emphasizes the importance of large peaks in the spectrum, has proven to work quite well (see chapter 4). Very small peaks, less than 0.1% total ion current, have little significance in normal mass spectra.

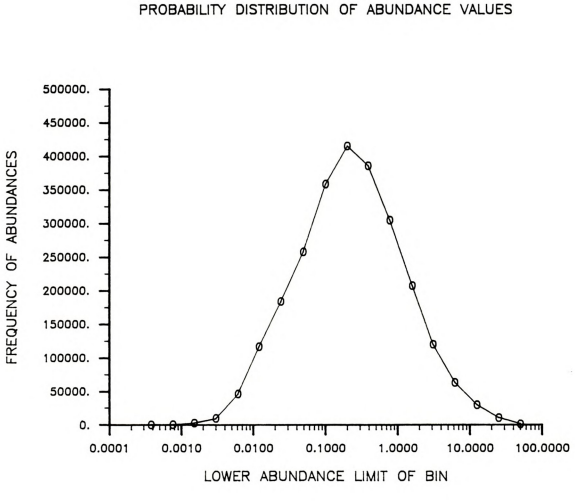


Figure 3.7 Frequency Distribution of Abundance Values in the Reference Data Base

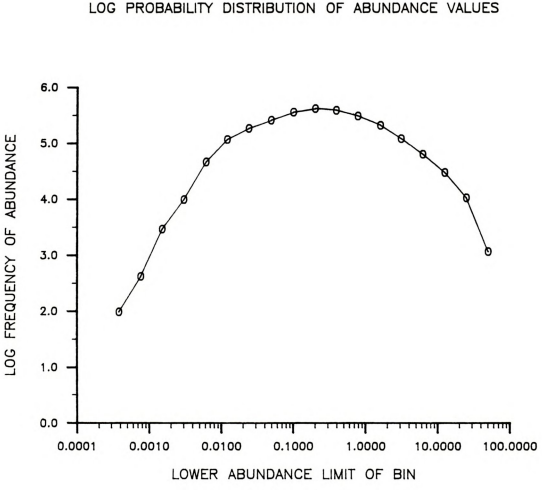


Figure 3.8 Log Frequency Distribution of Abundance Values in the Reference Data Base

Due to the large dynamic range in MS/MS spectra, very small peaks (less than 0.1 % total ion current) may still be relevant to the identification of substructures represented by that spectrum. The data base of MS/MS spectra is not yet large enough to obtain any statistical abundance information. However, one study indicated that the presence or absence of a peak in a MS/MS spectrum is more important than its actual abundance value (3).

Conclusions

The implementation of an effective information management system is essential for any group handling significant amounts of data. The additional dimension of data provided by MS/MS instruments required special considerations when designing both the storage and comparison features of the MS/MS information management system.

Characteristics of statistical data bases can provide useful clues for designing spectral matching programs and information management systems. Care must be taken, however, not to develop a system that is library dependent or to assign characteristics to individual spectra that must be updated as new spectra are acquired.

References

1. Varian Associates, Perkin-Elmer, Hewlett Packard.

2. Gregg, H. R., Hoffman, P. A., Enke, C. G., Crawford, R. W., Brand, H. R., Wong, C. M., **Anal. Chem.**, **56**, 1121 (1984).

3. Gregg, H. R., Ph. D. Dissertation, Michigan State University (1985).

4. Hoffman, P. A., Presented at Am. Soc. for Mass Spectrom. and Allied Topics., Boston, MA, (1983); bound p. 556.

5. Hoffman, P. A., Beckner, C. F., Enke, C. G., in preparation.

6. Lefkovitz, D., **J. Chem. Inf. Comput. Sci.**, **15**, 14 (1975).

7. Ullman, J. D., **Principles of Data Base Systems**, Computer Science Press, Rockville, MD (1980).

8. Heller, S. R., **Anal. Chem.**, **44**, 1951 (1972).

9. Hoffman, P. A., Ph. D. Dissertation, Michigan State University, in preparation.

10. Written by T. V. Atkinson, MSU Department of Chemistry, East Lansing, Michigan.

11. Written by P. A. Hoffman, MSU Department of Chemistry, East Lansing, Michigan.

12. Written by B. E. Wilson, MSU Department of Chemistry, East Lansing, Michigan.

13. Written by H. R. Gregg, MSU Department of Chemistry, East Lansing, Michigan.

14. Written by H. R. Gregg, MSU Department of Chemistry, East Lansing, Michigan.

15. Shelley, C. A., **J. Chem. Inf. Comput. Sci.**, **23**, 61 (1978).

16. Molecular Design Limited, Hayward, CA.

17. Smith, J. C., Kelly, W., Bickstock, A., Ridley, G. E., **Proc. 15th Annu. Conf. Mass Spectrom. Allied Top.**, 102 (1967).

18. Knock, B. A., Smith, I. C., Wright, D. E., Ridley, G. E., **Anal. Chem.**, **42**, 1516 (1970).

19. Crawford, L. R., Morrison, J. D., **Anal. Chem.**, **40**, 1464 (1970).

20. Hertz, M. S., Hites, R. A., Biemann, K., **Anal. Chem.**, **42**, 855 (1970).

21. Hertz, M. S., **Anal. Chem.**, **43**, 681 (1971).

22. Damen, H., Henneberg, D., Weimann, B., **Anal. Chim. Acta**, **103**, 289 (1978).

23. Milne, G. W. A., Heller, S. R., Martinson, D. D., **Adv. Mass Spectrom.**, **8B**, 1578 (1979).

24. McLafferty, F. W., Herten, R. H., Villwock, R. D., **Org. Mass Spectrom.**, **9**, 690 (1974).

25. Pesyna, G. M., Venkataraghavan, R., Dayringer, H., McLafferty, F. W., **Anal. Chem.**, **48**, 1369 (1976).

26. Kwok, K., Venkataraghavan, R., McLafferty, F. W., **J. Am. Chem. Soc.**, **95**, 4185 (1973).

27. Haraki, K. S., Venkataraghavan, R., McLafferty, F. W., **Anal. Chem.**, 53, 386 (1981).

28. McLafferty, F. W., Mun, K. I., **Int. J. Mass Spectrom. Ion Phys.**, **47**, 317 (1983).

29. Pesyna, G. M., McLafferty, F. W., in **Determination of Structural Physical Methods**, Academic Press, New York, NY, **6**, 91 (1976).

30. Henneberg, D., **Adv. Mass Spectrom.**, **8B**, Heyden and Sons, London England, 1511, (1979).

31. Pesyna, G. M., McLafferty, F. W., Venkataraghavan, R., Dayringer, H. E., **Anal. Chem.**, **47**, 1161 (1975).

32. 1974 Registry of Mass Spectral Data on Magnetic Tape, Wiley and Sons, New York, NY (1974).

33. Derendyaev, B. C., Koptyug, V. A., Lebedev, K. S., Sharapova, O. N., **Avtome**, **4**, 3 (1973).

34. Lebedev, K. S., Tormyshev, V. M., Derendyaev, B. G., Koptyug, V. A., **Anal. Chim. Acta.**, **133**, 517 (1981).

CHAPTER IV

A SPECTRAL MATCHING SYSTEM FOR MS/MS DATA*

Abstract

An automated mass spectrometry/mass spectrometry (MS/MS) search program has been developed which allows the user to match an unknown MS/MS spectrum against either primary or secondary spectra in a reference data base. The program employs several matching techniques for flexibility and avoids data compression or dependence on theoretical spectral properties. The strategy of the program is to eliminate the majority of candidate MS/MS spectra by prefiltering the candidates through inverted data files. An intensity-based matching algorithm then determines seven match factors to completely characterize the correspondence between the unknown and each remaining candidate spectrum. Parabolic fits to quotient spectra are also used, with limited success, to mask some deviations in spectra taken under different conditions. An experiment to characterize the program used 500 mass spectra from an old data base as unknowns for matching against the current MS/MS data base. The results indicated that the program retrieved an identical or structurally closely related reference compound (when no identical compound was present), 93% of the time.

*Note: This chapter is adapted from a manuscript written by the author of this thesis to be published in Computers and Chemistry, with C. G. Enke as a coauthor.

Introduction

The development of mass spectrometry/mass spectrometry (MS/MS) instruments has provided mass spectroscopists with several new types of spectral data (1). With the growth of MS/MS data libraries, the need for a system to organize, store, and search standard (primary) spectra, or parent, daughter, and neutral-loss (secondary) spectra has also developed. In response to this need, an MS/MS automated search program and the associated data bases were developed to allow the user to match the different types of MS/MS spectra against either primary or secondary reference mass spectra in a library. This paper describes the major features of our search program.

Reducing the Number of Candidate MS/MS Spectra

The strategy of the search program is to eliminate inappropriate candidate spectra without using intensity information in the data base. When matching a daughter spectrum against a small library of daughter spectra, the parent ion mass serves as an adequate filter for reducing the number of candidates. When primary mass spectra are matched against large data bases, the program eliminates the vast majority of candidate spectra by using inverted pointer files to prefilter the candidate spectra.

An inverted pointer file arranges mass spectral data according to some specified characteristic (2,3) such as m/z value. For each m/z value, the pointer file will contain a list of all daughter spectra that

include ions of that mass. Each of these lists is called a pointer stream. A program which logically ANDs several pointer streams together allows the user to quickly obtain a small subset of all reference spectra that contain the same combination of characteristics as the unknown. An intensity-based matching algorithm is then used to compare the unknown spectrum with each spectrum in this subset of the reference library. In addition to m/z values, pointer files enable the specification of other constraints, called pointer keys, such as empirical formula, molecular weight, and parent ion m/z values for use in obtaining a library subset. Until intensity-based matching is performed, the data base containing the intensity information is not accessed and abundance values are not considered. This design reduces the overall matching time and makes it practical to work with large numbers of unabridged spectra.

The MS/MS automated search program allows the user to eliminate candidates either manually or automatically. In the manual mode, the user chooses the pointer file containing the inverted data and specifies the pointer key value used to retrieve the desired pointer stream of candidate registry numbers into a pointer buffer. This pointer stream may be either logically ANDed or ORed with the current buffer contents. When the user is satisfied with the number of candidates remaining in the buffer, he may perform intensity-based matching on that subset. The user may view the current buffer contents or buffer length, change the current pointer file, or clear the pointer buffer at any time. The status of the pointer file and directory of the pointer keys are available to help the user decide on the applicability of the inverted

data. Similarly, the user may view the unknown MS/MS spectrum in order to select the significant data and characteristics for use as pointer keys. In the manual mode, the program provides the flexibility of allowing the user to specify the inverted data source and the type of buffer manipulation to operate on those data.

In automatic mode, the program reviews the unknown spectrum and determines the pointer stream length for each significant peak in the spectrum. A significant peak is defined as one which exceeds a minimum intensity value specified by the user. The significant peaks are then ranked according to their frequency in the data base; that is, in ascending order of pointer stream length. The pointer streams are logically ANDed together, starting with the shortest streams. This procedure quickly reduces the number of library candidates. Logical reduction of candidate spectra continues until the maximum number of allowed candidates as specified by the user has been reached. Intensity-based matching is then performed on the remaining candidates.

Intensity-Based Matching of MS/MS Spectra

Henneberg has identified four situations which may arise when searching a mass spectra library (4). They are listed below.

(1) The unknown is included in the library with a similar intensity pattern.

(2) The unknown is included in the library but with a different pattern.

(3) The unknown is a mixture whose components are in the library.

(4) The unknown, or one component of a mixture, is not in the library but compounds with similar structures are present.

Given no knowledge of which situation will arise, a single search method cannot adequately handle all four possible situations. In fact no computer algorithm can account for all the discrepancies between unknown and candidate spectra that are produced by different operating conditions. Most existing mass spectra search algorithms employ either "forward" or "reverse" searching techniques and the result of the search is presented as a single number termed a "match factor" or "similarity index" (4,5). More sophisticated procedures use several match factors in determining the best or nearest match (6-9). The solution employed in this program was the development of an algorithm which recognizes different kinds and degrees of similarity by using several matching techniques and calculating several match factors.

The intensity-based matching algorithm developed for this program was based on ideas from the SISCOM algorithm developed by D. Henneberg (4), and the matching algorithm proposed by K. S. Lebedev (10). The abundance values used for matching are placed in the form:

$$Yi = \log_2 (\text{Intensity}/\text{TIC}) + \text{SENS} \qquad (4.1)$$

where TIC is the total ion current and SENS is a sensitivity variable. A sensitivity factor is invoked to specify an intensity range desired for matching. Abundances falling below the minimum yield a negative Yi value and are considered noise. The desired abundance range used in

spectral matching differs according to whether primary or secondary spectra are being matched. In secondary MS/MS spectra, a wide dynamic range with very little noise is seen, and hence a high SENS value is used. In primary spectra where the dynamic range is smaller, the spectra richer, and the noise level higher, a lower SENS value may be used without impairing the results.

Lebedev earlier demonstrated the value of using intensity weighting factors in spectral matching that are based on the frequency of the intensity value (10-12). Hence, in this program the intensity of each spectral peak is weighted during matching by using the $\log_2$ of the normalized intensity (Equation 4.1).

The intensity-based matching algorithm calculates several match factors when comparing an unknown spectrum with a candidate spectrum. These match factors and their definitions are listed in Table 4.1.

The pattern correspondence match factor (PC) takes into account the deviations in intensity of sample peaks with respect to candidate peaks for those peaks common to both spectra. This forward searching match factor does not take into account those peaks in the reference spectrum not in common with peaks in the unknown spectrum. However, it does describe how well the spectral patterns of peaks common to both sample and candidate spectra compare. This factor detects structural similarities, such as substructures, that are difficult to recognize visually.

Table 4.1. Match Factor Definitions

PT - An overall match factor that indicates how well the intensities of all the peaks in the two spectra match.

$$PT = (\Sigma Ys + Yr - 2*| Yr - Ys|) / (\Sigma Ys + \Sigma Yr) * 100$$

where $Yi = \log_2$ (Intensity/Total Ion Current) + SENS

PC - A pattern correspondence factor that indicates how well the intensity of the peaks in common match.

$$PC = (\Sigma Ys - | Yr - Ys |) / (\Sigma Ys) * 100$$

NC - The number of peaks common to both the candidate and unknown sample spectrum.

NS - The number of peaks remaining unmatched in the unknown sample spectrum.

NR - The number of peaks remaining unmatched in the reference spectrum.

IS - The percent total ion current of the sample 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.

NP - The number of peaks in common between the two spectra that are used in parabolic fitting of the quotient spectrum.

Chi^2 - The reduced chi-square value obtained from parabolic fitting of the quotient spectrum.

The overall match factor (PT) takes into account the deviations in intensity of the sample peaks with respect to the candidate peaks and vice versa for all peaks in both spectra. This factor is a combination of both forward and reverse searching techniques. The forward searching technique takes into account deviations in the candidate spectrum from the unknown spectrum while the reverse searching technique takes into account the deviations in the unknown spectrum from the candidate spectrum.

If the unknown spectrum is of "poor" quality as a result of being acquired under experimental operating conditions different from those for library spectra, the reverse searching technique will still contribute toward a large overall match factor when comparing identical compound. Likewise, if the candidate spectrum is of "poor" quality, the forward searching technique will still contribute toward a large overall match factor when comparing identical compounds. The combination of these two techniques enhances contributions due to the commonalities of the two spectra and is, therefore, more versatile when comparing spectra of varying quality than a single method.

The chi-square value employs spectral ratioing in determining the similarity between spectra. Henneberg has demonstrated the use of parabolic fits to quotient spectra in masking deviations in spectra taken under different source conditions (13). It was hoped that the same procedure would help mask some of the deviations in MS/MS spectra arising from different collision cell conditions. The sparsity of peaks in daughter spectra, however, severly limits this application.

Results

The matching of a measured n-butylbenzene spectrum against a library of reference spectra demonstrates how the MS/MS automated search program operates. The user first specifies the data base containing the unknown spectrum. The unknown data base may be the one containing reference spectra, another data base of the same format (14), or a data base of a different, known format (15). The user then specifies the data base to serve as the library of reference spectra.

After viewing the unknown n-butylbenzene spectrum, the user determines the minimum significant intensity present in the spectrum and selects a pointer file to be used for reducing the number of candidate spectra. The program scans the unknown (n-butylbenzene) spectrum, determines the significant peaks present as specified by the user, and then obtains the pointer stream length for each peak. The significant peaks in the unknown spectrum are then ranked according to their increasing frequency in the data base. The frequency results for those peaks in the mass spectrum of n-butylbenzene which have intensities above 1% total ion current are tabulated in Table 4.2. The m/z 134^{+} peak is the least frequent significant unknown peak present in the reference library and is contained in only 10,042 of the over 32,000 reference mass spectra. The peak at m/z 92^{+} has the next lowest frequency and is contained in 14,566 of the reference mass spectra.

Table 4.2. Frequency of Mass Spectra Peaks of N-butylbenzene in the Data Base.

Ion	Data Base Frequency
134^{+}	10,422
92^{+}	14,566
93^{+}	14,793
105^{+}	16,040
78^{+}	17,633
79^{+}	18,333
65^{+}	18,411
91^{+}	18,789
77^{+}	21,670

When the first two pointer streams are ANDed together, only those reference spectra containing peaks at both m/z 134^{+} and m/z 92^{+} are retained. The Venn diagram in Figure 4.1 illustrates the logical ANDing of several pointer streams and the resulting subset of reference spectra. The graph in Figure 4.2 demonstrates how the number of reference compounds is reduced as pointer streams corresponding to each of the significant peaks in the unknown spectrum are ANDed together. The greatest reduction occurs during the first few ANDing passes since these use peaks with the lowest frequency in the reference data base. Results involving a number of different unknowns are shown in Figure 4.2 to illustrate the varying results of the logical reduction process.

After the automatic logical ANDing process has finished, almost 3,000 candidate spectra still remain. To reduce this number further, the molecular weight pointer file could be used to eliminate candidates with different molecular weights. Although this act would reduce the number of candidate spectra, it would eliminate those similar spectra of other molecular weights. It is desirable to retain similar higher molecular weight candidates in cases where the unknown spectrum is not present in the library and a list of structurally similar compounds is desired. After a reasonable subset of candidate spectra is obtained (25-100), the user instructs the program to perform intensity-based matching on the remaining set of candidate spectra. A sensitivity factor of 5 is specified so that very small peaks (< 1.0E-5 total ion current) in the reference spectra are ignored.

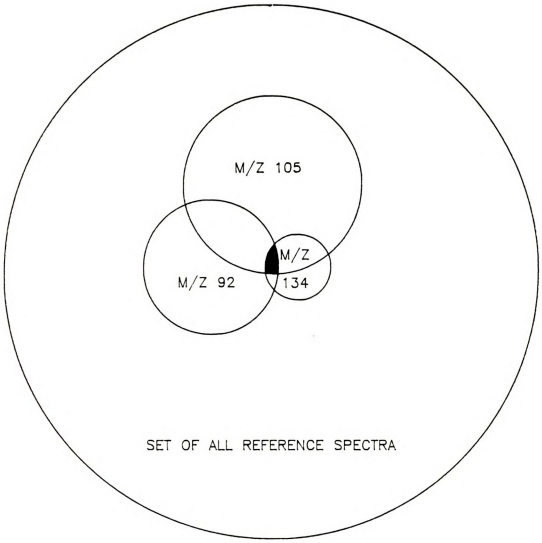


Figure 4.1 Logical Reduction of Candidate Spectra (Venn Diagram).

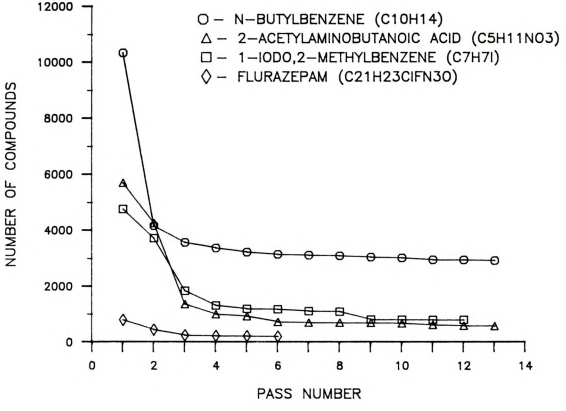


Figure 4.2 Logical Reduction of Candidate Spectra

The results of the intensity-based match are presented in Table 4.3. The best matching candidate spectrum was a reference spectrum of n-butylbenzene. The overall match factor of 86 indicates that the best matching spectra were not identical. The value of NC indicates that there were 23 peaks in common between the unknown and best candidate spectrum while NS indicates that there were no peaks in the unknown spectrum that weren't present in the best candidate spectrum. The value of NR indicates the there were 60 peaks in the reference spectrum of n-butylbenzene that were not in the unknown spectrum. The value of IR indicates, however, that these 60 discrepant peaks constituted only 10% of the total ion current in the reference spectrum and were noise peaks.

The value of PC indicates that the spectral patterns for the peaks in common between the two spectra matched very well. The overall match factor (PT) value is less than that of the PC match factor since PT takes into account those 60 small, discrepant peaks present in the best reference spectrum.

The next best matching reference spectrum is an isobutylbenzene spectrum with an overall match factor value of 71. The isobutylbenzene spectrum has 23 peaks in common with the unknown spectrum (as does the n-butylbenzene), but the pattern correspondence (PC) value between the mass spectral pattern of isobutylbenzene and the unknown spectrum is only 75.

The results in Table 4.3 illustrate the discriminating power of this matching algorithm and its ability to pick out the correct unknown from a large volume of data. Note that many of the 15 candidates listed have functional groups in common with the unknown, and that the skeletons of many of the candidates are similar to that of the unknown.

Table 4.3. Match of N-butylbenzene Mass Spectrum

PT	PC	NC	NS	NR	IS	IR	Name
86	89	23	0	60	0	10	N-BUTYLBENZENE
71	75	23	0	69	0	11	ISOBUTYLBENZENE
65	72	22	1	100	0	20	1-PHENYLHEXANE
62	65	22	1	86	0	20	1-PHENYL-3-METHYLBUTANE
55	62	21	2	29	1	16	3-PHENYLPROPANAL
55	62	21	2	83	1	21	1-PHENYLHEPTANE
54	64	23	0	65	0	34	1-BENZYLOXY-2-BUTANOL
52	54	21	2	87	1	26	1-PHENYL-2-METHYLBUTANE
51	62	22	1	28	1	37	3-PHENYLPROPANOL
49	59	22	1	38	1	29	(2-CHLOROPROPYL)BENZENE
48	56	23	0	91	0	23	1,3-DIMETHYL-2-ETHYLBENZENE
48	60	22	1	67	0	30	CINNAMYLALCOHOL
48	51	21	2	89	1	26	4-PHENYLBUTENE-1
48	60	20	3	46	2	35	2-(2-(BENZYLOXYETHOXY)ETHANOL
47	62	23	0	39	0	40	3-PHENYLNITROPROPANE

Matching N-butylbenzene Daughter Spectra Against Similar Compounds

In order to characterize the ability of the matching program to correctly identify compounds using MS/MS spectra, daughter spectra of several similar compounds (n-butylbenzene, t-butylbenzene, diethylbenzene, triethylbenzene) were taken under different collision cell conditions. The collision energy was varied from 10 to 40 volts, and collision cell argon pressure was varied from P/P_o of 1/3 to 1/10 (where P_o is the intensity of the parent peak with no collision cell gas present, and P is the intensity of the parent peak when the collision gas is present). N-butylbenzene was chosen since its daughter spectrum varies considerably with experimental conditions. The m/z 92^+ daughter ion intensity in the m/z 134^+ daughter spectrum is dependent on collision energy, while m/z 65^+ is dependent on collision pressure, and other ions (m/z 105^+, m/z 78^+) seem independent of collision cell conditions (16). N-butylbenzene (m/z 134^+) daughter spectra taken under various conditions are compared with m/z 134^+ daughter spectra of structurally related compounds in Table 4.4.

The matching results demonstrates the ability of the program to correctly identify the sample when reference spectra are obtained under various instrumental conditions. The ability of the search program to discern the spectra of the identical compound from those of other

Table 4.4 M/Z 134^{+} DAUGHTER MATCH FACTORS
SAMPLE SPECTRUM (N-BUTYLBENZENE 0.33 P/P_0, 28 eV CE)

PT	PC	NC	NS	NR	IS	IR	NP	Chi-Square	Name	P/P_0	CE(eV)
100	100	9	0	0	0	0	6	0.1526D-03	N-BUTYLBENZENE	.33	28
96	96	8	1	0	0	0	6	0.7615D+01	N-BUTYLBENZENE	.33	33
93	93	7	2	0	0	0	6	0.2337D+02	N-BUTYLBENZENE	.33	20
81	82	8	1	2	0	0	6	0.3911D+03	N-BUTYLBENZENE	.10	22
69	79	9	0	0	0	0	6	0.1622D+03	N-BUTYLBENZENE	.10	27
58	55	8	1	0	24	0	5	0.9602D+02	N-BUTYLBENZENE	.10	17
50	53	8	1	1	43	18	5	0.9690D+02	T-BUTYLBENZENE	.10	30
42	45	3	6	0	32	0	2	0.0000D+00	T-BUTYLBENZENE	.33	30
42	44	3	6	0	32	0	2	0.0000D+00	TRIETHYLBENZENE	.33	10
39	42	3	6	0	32	0	1	0.0000D+00	DIETHYLBENZENE	.33	20

compounds is visually illustrated in Figure 4.3. The distance between the collision pressure lines of the n-butylbenzene and other spectra quantifies the discriminating ability of the matching program in terms of overall match factor values.

A plot of pattern correspondence factor (PC) versus collision energy and pressure for the results in Table 4.4 is shown in Figure 4.4. The separation of the collision pressure lines for n-butylbenzene spectra from other spectra is much less than in Figure 4.3. This smaller separation demonstrates the benefit of using a combination of forward and reverse searching techniques such as with the overall match factor (PT) over using a single forward matching technique such as the pattern correspondence match factor (PC).

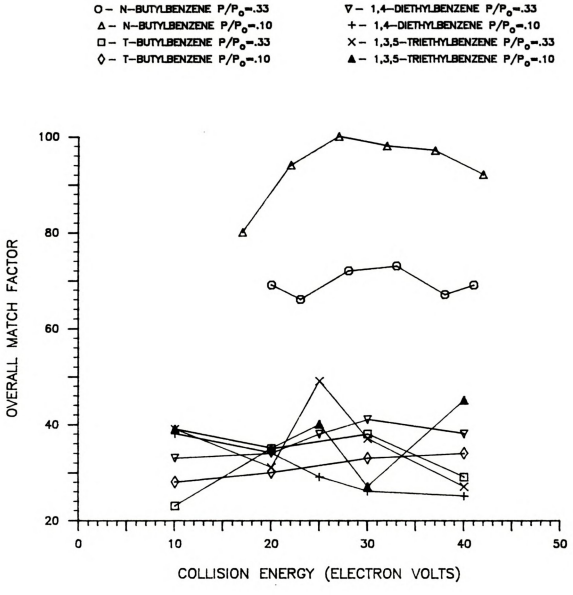


Figure 4.3 Substituted Benzene Matching Results
Sample: n–Butylbenzene, P/P_o = .10, CE = 27 eV

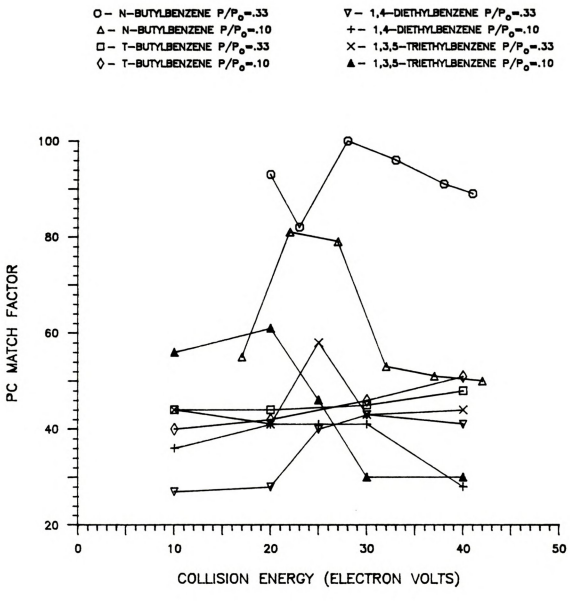


Figure 4.4 Substituted Benzene Matching Results
Sample: n–Butylbenzene, P/P_o = .33, CE = 27 eV
PC Match Factor Results

Performance Characteristics of the MS/MS Automated Search Program

To test the reliability of the MS/MS automated search program, an older spectrum data base served as a source of unknown mass spectra to be matched against the current library. Five hundred random primary spectra were chosen from the old reference data base and matched against the current mass spectrum library. The results indicate that 86% of the time the search program retrieved the identical compound or an isomer as the best match. The mean overall match factor (PT) was 96 and indicates the high confidence in the identity of the retrieved compounds. In 7% of the cases a related compound was retrieved as the match. These reference compounds had functional groups or a common structural backbone identical to the unknown. The mean overall match factor for these cases was 65 and indicates the presence of some structural or functional similarity, but that identical compounds or isomers were not found. An unrelated compound was retrieved as the best match in 7% of the cases. However, the low mean overall match factor was only 35 and indicates that there was little resemblence between the unknown compound and any reference compound.

The time spent reducing the reference spectra candidates to a reasonable subset depends on the number significant peaks in the unknown spectrum and their frequency in the data base. Those unknown spectra containing significant peaks of low frequency in the data base will cause the logical reduction procedure to quickly converge to a small subset. Unknown spectra with many high frequency peaks in the library

will slowly converge, while unknown spectra containing only a few high frequency peaks may not converge at all. In the latter case, intensity-based matching must be performed on a larger than desired subset of candidate spectra. The average time required to complete candidate reduction in the matching of the 500 unknown spectra was 9 minutes. Variation in response times reflects the different frequencies of the unknown mass spectral peaks in the spectrum data base. An additional experiment was performed involving 50 unknown spectra where the candidate spectra were initially prefiltered by molecular weight. This process decreased the average candidate reduction time to 40 seconds.

The time spent performing intensity-based matching is linearly related to the number of remaining candidates and by the number of peaks per candidate spectrum. The timing results for intensity-based matching of some of the 500 unknown spectra are presented in Figure 4.5. The average number of points per candidate spectrum in the distribution was found to be 50. Deviations from linearity in Figure 4.5 result from variations in the number of points per candidate spectrum. The MS/MS automated search program was written in FORTRAN-77 and is implemented on a DEC PDP-11/23 minicomputer running the RSX11M multi-tasking operating system in a multi-user environment. A 474 megabyte disk drive with an 18 millisecond average access time is used to hold the mass spectra library.

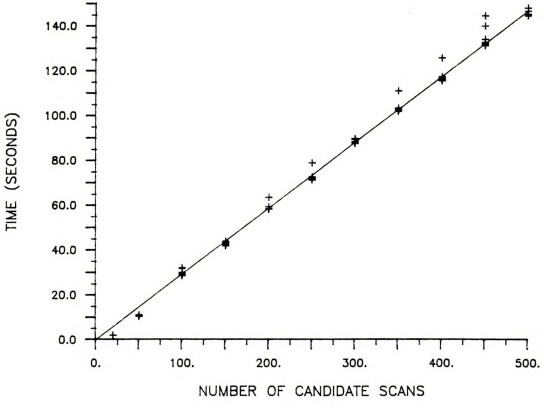


Figure 4.5 Intensity–Based Matching Speeds

Conclusions

The development of the MS/MS automated search program permits the identification of unknown MS/MS spectra taken under a variety of conditions. Characteristics of this program include the ability to match MS/MS spectra using unabridged library spectra and the ability to mask some deviations in spectra taken under different instrumental conditions. Transparent match factors allow the user to see how the program obtained its results.

The ability of the program to correctly identify identical compound MS/MS spectra taken under different conditions from similar spectra of other compounds has proven valuable in determining how instrumental conditions affect MS/MS spectra. The limitations of the program may provide the basis for determining the appropriate range of standard operating conditions used in the development of a large MS/MS library.

References

1. Yost, R. A., Enke, C. G., **J. Am. Chem. Soc.**, **100**, 2274 (1978).

2. Lefkowitz, D., **J. Chem. Inf. Comput. Sci.**, **15**, 14 (1975).

3. Heller, S. R., **Anal. Chem.**, **44**, 1951 (1972).

4. Damen, H., Henneberg, D., Weimann, B., **Anal. Chim. Acta**, **103**, 289 (1978).

5. Hertz, M. S., Hites, R. A., Biemann, K., **Anal. Chem.**, **42**, 855 (1970).

6. Pesyna, G. M., McLafferty, F. W., in **Determination of Structural Physical Methods**, Academic Press, New York, NY, **6**, 91 (1976).

7. Henneberg, D., **Adv. Mass Spectrom.**, **8B**, Heyden and Sons, London England, 1511 (1979).

8. Haraki, K. S., Venkataraghavan, R., McLafferty, F. W., **Anal. Chem.**, **53**, 386 (1981).

9. Clark, H. A., Jurs, P. C., **Anal. Chim. Acta**, **130**, 75 (1982).

10. Lebedev, K. S., Tormyshev, V. M., Derendyaev, B. G., Koptyug, V. A., **Anal. Chim. Acta**, **133**, 517 (1981).

11. Pesyna, G. M., Venkataraghavan, R., Dayringer, H., McLafferty, F. W., **Anal. Chem.**, **48**, 1369 (1976).

12. McLafferty, F. W., Mun, K. I., **Int. J. Mass Spectrom. Ion Phys.**, **47**, 317 (1983).

13. Henneberg, D., presented at 31st Annual Conference on Mass Spectrometry and Allied Topics, Boston, MA (1983); bound p. 185.

14. Hoffman, P. A., Beckner, C. F., Enke, C. G., in preparation.

15. Gregg, H. R., Hoffman, P. A., Enke, C. G., Crawford, R. W., Brand, H. R., Wong, C. M., **Anal. Chem.**, **56**, 1121 (1984).

16. Dawson, R. W., presented at 31st Annual Conference on Mass Spectrometry and Allied Topics, Boston, MA, (1983); bound p. 203.

CHAPTER V

INSTRUMENTAL PARAMETER EFFECTS ON MATCHING DAUGHTER SPECTRA

Introduction

Many instrumental parameters influence ion transmission in MS/MS instruments. These include voltages on physical devices such as the ion source, mass filters, electrostatic lenses, and those parameters which effect the framentation process such as collision cell pressure and collision energy. All parameters effect the ion transmission and thereby the sensitivity of the MS/MS instrument.

Collision cell pressure and collision energy directly affect the fragmentation in the collision cell. As collision cell pressure increases, we expect more fragmentation to occur due to the increased number of collisions. At high collision cell pressures, we expect the sensitivity of the instrument to decrease due to ion scattering. As collision energy increases, we expect more fragmentation due to a greater efficiency of fragmentation per collision.

Different combinations of instrumental parameters effect MS/MS spectral patterns. This fact has hindered the determination of standard operating conditions for acquiring MS/MS spectra and the development of MS/MS libraries. Available techniques for comparing conventional mass spectra do not account for the effects of different experimental conditions on spectral features. This chapter focuses on how instrumental parameters affect MS/MS spectral patterns and how the MS/MS

spectral matching program is used to compare MS/MS spectra taken under different conditions.

Previous attempts to develop standard operating conditions for acquiring MS/MS spectra have included an interlaboratory round-robin study by Dawson (1,2). Each laboratory followed a prescribed procedure to obtain daughter spectra of n-butylbenzene. Poor correlation among the spectra resulted. Various laboratories, using different instruments, produced spectra with substantial deviations; especially regarding collision cell pressure effects. The conclusion of the study suggested that MS/MS spectroscopists can not expect to accurately reproduce spectra (both mass and intensity values) coming from different instruments or laboratories.

With this study in mind, we sought to develop a spectral matching program to recognize those spectral features dependent on the compound over those features arising from instrumental parameters. The goal was to correctly identify spectra taken under "similar" conditions on identically configured instruments. This study excluded variations in spectra due to source type, varying decomposition mechanisms, or instrument configuration. If this goal were to be achieved for a limited range of operating conditions, a great step toward developing MS/MS libraries would be made.

Instrumental Parameter Effects on CID Efficiency

MS/MS spectra differ from conventional mass spectra in several fundamental ways. The spectra are simpler and represent only a few decomposition pathways. In addition, there are frequently no isotope peaks and little noise in the MS/MS spectrum. In comparison to the fragmentation process in an EI source, the decomposition process occurring in a triple quadrupole mass spectrometer collision cell is a "soft" or low energy process.

MS/MS spectra are more sensitive to instrumental conditions than conventional spectra (3). Fourteen devices are present in a triple quadrupole MS/MS instrument ion path (the repeller, CI volume, EI volume, extractor lens, lens 1, lens 2, lens 3, quadrupole 1, lens 4, quadrupole 2, lens 5, quadrupole 3, the conversion dynode, and the electron multiplier) (Figure 5.1). This large number of parameters complicates the development of standard operating conditions and causes observed daughter spectra to exhibit fragmentation characteristics dependent on instrumental parameter settings.

The operator can dramatically change a spectrum's appearance by varying device potentials along the ion path of the MS/MS instrument. Collision energy (commonly defined as the voltage difference between the source and quadrupole 2 and expressed in eV) and collision cell pressure variations can eliminate or produce daughter ions. Changes in lens voltages affect peak shape and intensity (4). To improve

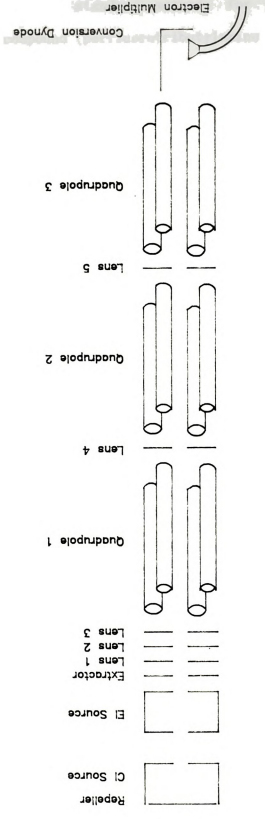


Figure 5.1 ExtraNuclear EL 400-TQ3 Triple Quadrupole Mass Spectrometer

reproducibility, standard methods for "tuning" the ion path using a standard compound (perfluorotributylamine) have been developed (3,5).

During the course of an experiment, the conditions for collisionally induced dissociation (CID) fragmentation are defined by the collision energy and collision cell pressure settings. All other devices, except drawout potential, are independent of these two parameters. Collision energy is measured as the voltage difference between the EI or CI source and the second quadrupole dc level (Q2). Drawout potential is measured as the voltage difference between the second and third quadrupole dc levels (Q2 - Q3).

To study the ion transmission effects of collision energy, collision cell pressure, and drawout potential, m/z 136^+ and 105^+ daughter spectra of methylbenzoate were acquired under a variety of conditions using an EI source. Collision cell pressure of Argon was varied from 5.2 X 10^{-5} torr to 4.3 X 10^{-2} torr. Collision energy was varied from 0 eV to 40 eV for each collision cell pressure. The drawout potential varied from Q2+20 V to Q2-30 V for each combination of collision cell pressure and collision energy. This led to the acquisition of 210 m/z 136^+ daughter spectra and 210 m/z 105^+ daughter spectra.

The data in Figure 5.2 demonstrate the effects of varying all of these parameters on the total ion current (TIC) (including the parent ion abundance) for m/z 136^+ methylbenzoate daughter spectra. Each multiplet of 5 peaks represents the total ion current for a different

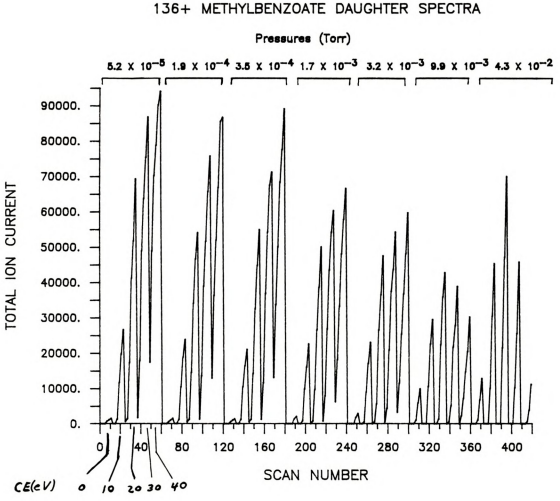


Figure 5.2 Instrumental Parameter Effects on Total Ion Current
Each Peak is a Scan of Drawout Potential From Q2+20 V to Q2−30 V

collision cell pressure. The seven pressures investigated were 5.2 X 10^{-5}, 1.9 X 10^{-4}, 3.5 X 10^{-4}, 1.7 X 10^{-3}, 3.2 X 10^{-3}, 9.9 X 10^{-3}, and 4.3 X 10^{-2} torr. Each peak in a multiplet represents the total ion current as the collision energy is scanned at a constant pressure. The five collision energies investigated were 0, 10, 20, 30, and 40 eV. Each point on a peak then represents the drawout potential effect on the total ion current for that collision energy and collision pressure. The collision cell pressure increases with scan number, and the collision energy increases with each peak in a multiplet. The drawout potential becomes more negative over the course of a peak.

At low pressures (scans 1 - 180), total ion current increases with collision energy and drawout potential. At higher pressures (scans 181 - 420), the total ion current drops off the due to ion scattering. The effect of collision energy on total ion current also changes at higher pressures where significant fragmentation occurs in the collision cell.

Total ion current does not adequately describe the extent of MS/MS fragmentation or the ion collection capabilities of the MS/MS instrument. Collision induced dissociation (CID) efficiency takes into account both the fragmentation efficiency ($\sum f_i/TIC$), and collection efficiency (TIC/P_o). CID efficiency is defined as fragmentation efficiency multiplied by the collection efficiency ($\sum f_i/P_o$).

The combined effects of all parameters upon CID efficiency of MS/MS spectra are illustrated in Figures 5.3 and 5.4. The intensity of

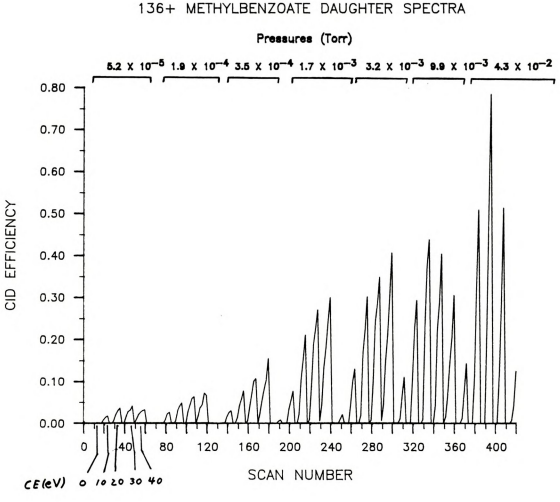


Figure 5.3 Instrumental Parameter Effects on Collision Induced Dissociation Efficiency for 136+ Methylbenzoate Daughter Spectra Each Peak is a Scan of Drawout Potential From Q2+20 V to Q2−30 V

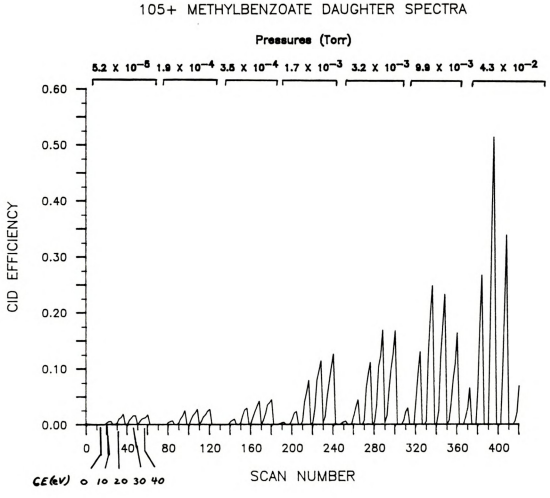


Figure 5.4 Instrumental Parameter Effects on Collision Induced Dissociation Efficiency for 105+ Methylbenzoate Daughter Spectra Each Peak is a Scan of Drawout Potential From Q2+20 V to Q2−30 V

each 5 peak multiplet increases with scan number indicating that fragmentation efficiency increases as collision cell pressure increases. This effect was expected as a greater number of Argon atoms become collision targets. The effect of collision energy at higher pressures also becomes apparent (scans 300 - 420) since the impact energy determines the extent of fragmentation. The maximum CID efficiency for both m/z 136^{+} and m/z 105^{+} daughter spectra is obtained at 4.3×10^{-2} torr, 20 electron Volts collision energy and -40 Volts drawout potential.

The effects of drawout potential on CID efficiency are illustrated by plotting CID efficiency versus collision cell pressure (at a constant collision energy) (Figure 5.5). The results indicate that CID efficiency increases as the drawout potential becomes more negative. Peak shape, however, suffers as the drawout potential reaches large negative values. This causes peak splitting and causes the CID efficiency to become artificially large. Therefore, I found it best to apply a constant drawout potential by tracking Q3 behind Q2 by only -10 or -20 volts.

The effect of collision energy and collision cell pressure upon CID efficiency for m/z 136^{+} and m/z 105^{+} methylbenzoate daughter spectra are illustrated in Figures 5.6 and 5.7. Fragmentation initially increases with collision cell pressure as fragmentation efficiency increases. As collision cell pressure becomes higher, ion scattering becomes dominant. This causes the collection efficiency to decrease and the CID efficiency to subsequently decline. By judiciously selecting

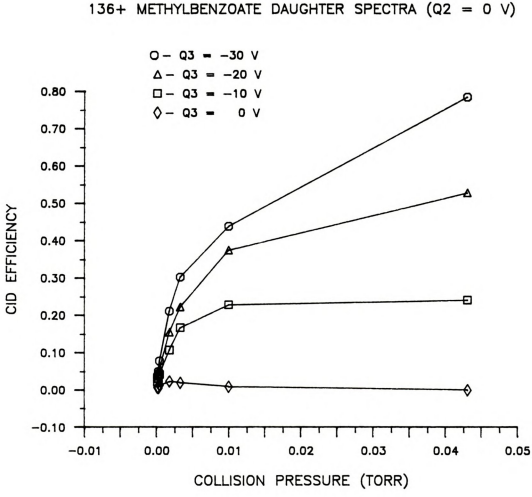


Figure 5.5 Drawout Potential Effects on Collision Induced Dissociation Efficiency for 136+ Methylbenzoate Daughter Spectra

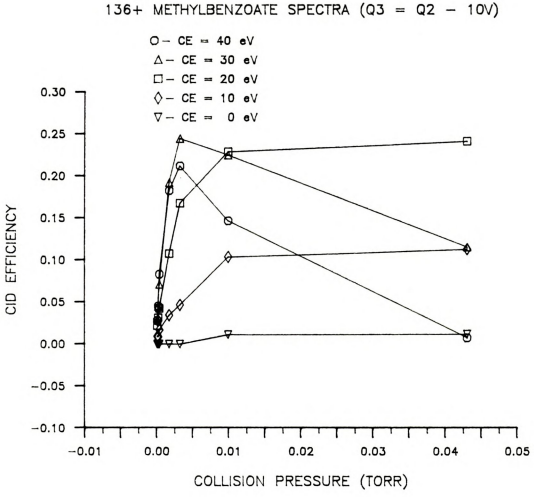


Figure 5.6 Collision Energy Effects on Collision Induced Dissociation Efficiency for 136+ Methylbenzoate Daughter Spectra

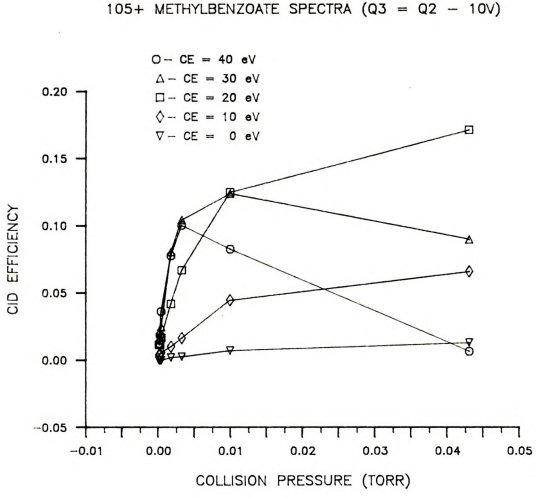


Figure 5.7 Collision Energy Effects on Collision Induced Dissociation Efficiency for 105+ Methylbenzoate Daughter Spectra

the collision energy, ion scattering can be minimized. For methylbenzoate, 20 eV collision energy produced a large, constant CID efficiency value in the higher collision cell pressure region for both m/z 136$^+$ and m/z 105$^+$ daughter spectra.

To determine collision energy and collision cell pressure effects on individual daughter ions, it is necessary to examine the collision energy breakdown curves for m/z 136$^+$ and m/z 105$^+$ daughter spectra at different collision cell pressures (Figures 5.8, 5.9). At very low collision cell gas pressures (5.2 X 10^{-5} to 3.5 X 10^{-4} torr), little fragmentation occurs and the parent ion dominates the daughter spectrum. At medium collision cell pressures (1.7 X 10^{-3} to 3.2 X 10^{-3} torr), fragmentation increases and the relative abundance of the parent ion decreases. At high collision cell pressures (9.9 X 10^{-3} to 4.3 X 10^{-2} torr), the decomposition order changes causing different ions to appear in the spectrum.

The collision cell pressure required for a particular experiment depends on the type of information desired. If consistent fragmentation resulting from single collisions of the parent ion and collision gas is desired, the collision cell pressure should remain in a region low enough such that only first-order fragmentation occurs. Higher collision cell pressures may cause multiple collisions in the collision cell resulting in additional fragmentation that complicates the spectrum. The effect of collision cell pressure on matching MS/MS spectra was investigated in the next section.

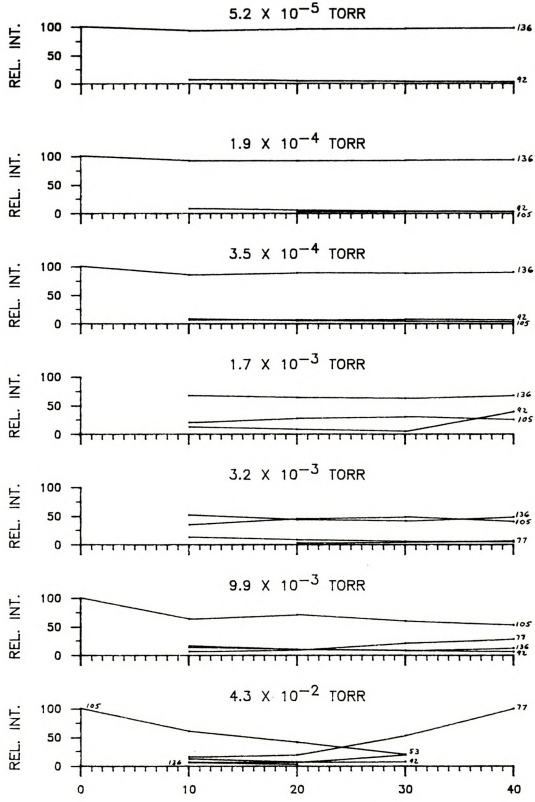


Figure 5.8 CE Breakdown Curves for 136+ Methylbenzoate Daughters

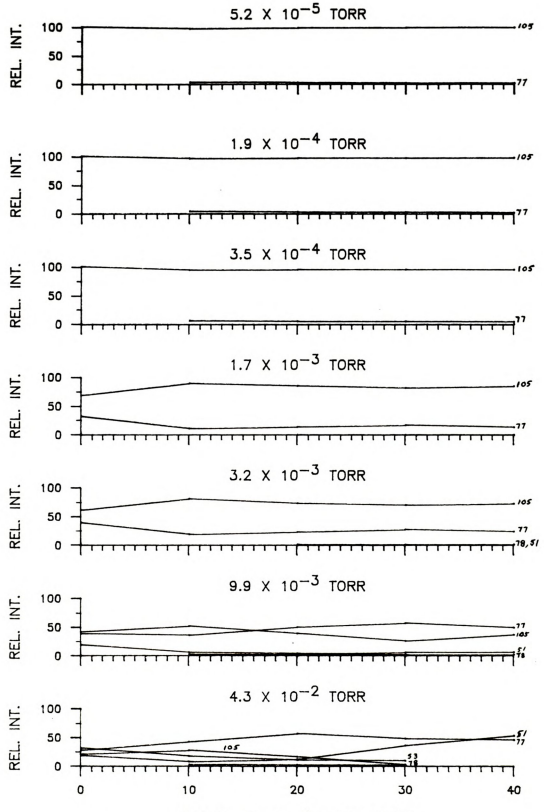


Figure 5.9 CE Breakdown Curves for 105+ Methylbenzoate Daughters

Instrumental Parameter Effects on Spectral Matching

The MS/MS spectral matching program was designed to be a flexible, interactive tool for comparing MS/MS spectra. Flexibility is necessary to account for deviations due to a variety of experimental situations. The search program uses several match factors to characterize the correspondence between the sample and each candidate spectrum (5). Similarity and difference factors are weighted and combined to form an overall match factor. A complete description of the MS/MS spectral matching program was presented in chapter 4. The resulting match factors returned to the user are described in Table 4.1.

To test the effectiveness of the MS/MS spectral matching program on varying MS/MS spectra, methylbenzoate daughter spectra were acquired under a variety of instrumental conditions. When different instrument conditions were required, only the parameter of interest was changed. This parameter, however, was varied over a range sufficient to cause an appreciable change in the spectrum's appearance. The resulting spectra were matched against each other to determine which combinations of instrumental parameters produced similar spectra.

The results of matching a methylbenzoate (m/z 136^+) daughter spectrum against identical compound spectra acquired under various experimental conditions are displayed in Table 5.1. Each table entry represents a m/z 136^+ daughter spectrum of methylbenzoate taken under different instrumental conditions. Results of the top matching spectra are displayed in the table.

Table 5.1 $m/z\ 136^{+}$ Methylbenzoate Daughter Spectra Match Factors
Sample Spectrum (Coll Press: 9.9 X 10^{-3} Torr, CE: 20 eV, Drawout: −10V)

PT	PC	NC	NS	NR	IS	IR	Coll Press(Torr)	CE(eV)	Drawout(V)
100	100	6	0	0	0	0	9.9 X 10^{-3}	20	−10
85	79	5	1	0	5	0	9.9 X 10^{-3}	10	−20
84	80	5	1	0	4	0	9.9 X 10^{-3}	10	−10
80	83	5	1	2	5	10	9.9 X 10^{-3}	20	−20
77	68	4	2	0	9	0	9.9 X 10^{-3}	10	−30
71	84	6	0	3	0	16	4.3 X 10^{-2}	10	−20
70	66	5	1	0	5	0	3.2 X 10^{-3}	30	−20
69	73	5	1	2	5	10	9.9 X 10^{-3}	30	−10
67	52	3	3	0	18	0	9.9 X 10^{-3}	30	0
66	68	5	1	4	5	8	3.2 X 10^{-3}	40	−30
66	65	5	1	1	5	3	3.2 X 10^{-3}	30	−10
66	65	5	1	1	5	4	3.2 X 10^{-3}	40	−10
66	67	4	2	2	13	19	9.9 X 10^{-3}	0	−30
65	64	5	1	1	5	3	3.2 X 10^{-3}	20	−10
65	69	4	2	5	13	18	4.3 X 10^{-2}	10	−30

The spectral matching data (Table 5.1, 5.2) demonstrate that similar spectra may be obtained by using different combinations of instrumental parameters. By grouping spectra with the best match factors, an acceptable operating range for each instrumental parameter required to obtain similar matching daughter spectra was determined.

Candidate spectra acquired at the same collision cell pressure matched the best, indicating that collision cell pressure affected the spectra the most. Collision energy also showed a significant effect on the spectral pattern of the daughter spectra. Drawout potential had a lesser effect on the appearance of the mass spectrum since different values for these parameters appear throughout the top matching spectra.

Collision energy and pressure effects on the overall match factor for the tabular data are illustrated in Figures 5.10 and 5.11. To eliminate the effects of other parameters, only spectra taken at Q2-10 volts were considered. The spectrum treated as the sample spectrum in the matching process was acquired at 9.9×10^{-3} torr collision cell pressure and 20 eV collision energy.

The effects of collision energy vary with collision pressure. At low collision cell pressures (5.2×10^{-5} to 3.5×10^{-4} torr), the collision energy makes little difference due to the sparse fragmentation. At higher collision cell pressures, the probability of an ion/atom collision increases. The collision energy determines if fragmentation will occur and, therefore, has a greater effect on the overall match factor.

Table 5.2 $m/z\ 105^{+}$ Methylbenzoate Daughter Spectra Match Factors
Sample Spectrum (Coll Press: 9.9 X 10^{-3} Torr, CE: 20 eV, Drawout: −10V)

PT	PC	NC	NS	NR	IS	IR	Coll Press(Torr)	CE(eV)	Drawout(V)
100	100	6	0	0	0	0	9.9 X 10^{-3}	20	−10
98	98	6	0	0	0	0	9.9 X 10^{-3}	20	−20
93	89	5	1	2	2	1	9.9 X 10^{-3}	20	−30
90	90	6	0	0	0	0	9.9 X 10^{-3}	10	−20
89	89	6	0	0	0	0	9.9 X 10^{-3}	10	−10
87	87	6	0	1	0	1	9.9 X 10^{-3}	10	−30
83	83	6	0	0	0	0	9.9 X 10^{-3}	30	−20
82	84	6	0	5	0	5	4.3 X 10^{-2}	10	−30
80	82	6	0	4	0	4	4.3 X 10^{-2}	10	−20
80	80	5	1	1	2	3	9.9 X 10^{-3}	30	−10
78	69	4	2	0	7	0	9.9 X 10^{-3}	30	0
77	75	6	0	0	0	0	4.3 X 10^{-2}	10	−10
77	73	5	1	0	2	0	9.9 X 10^{-3}	0	−30
77	73	4	2	2	4	2	9.9 X 10^{-3}	30	−30
77	70	4	2	0	4	0	9.9 X 10^{-3}	0	−20

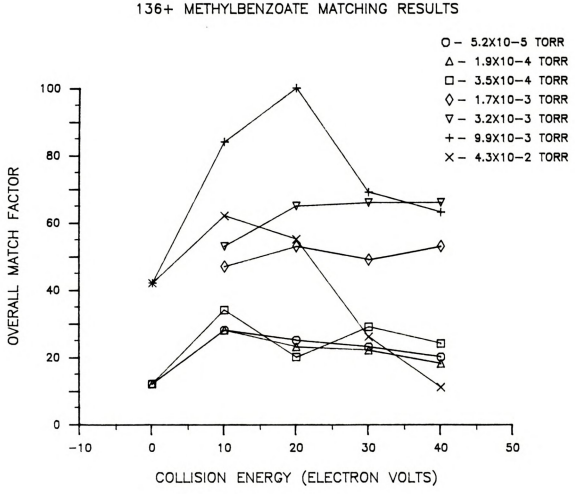


Figure 5.10 Instrumental Parameter Effects on the Overall Match Factor for 136+ Methylbenzoate Daughter Spectra

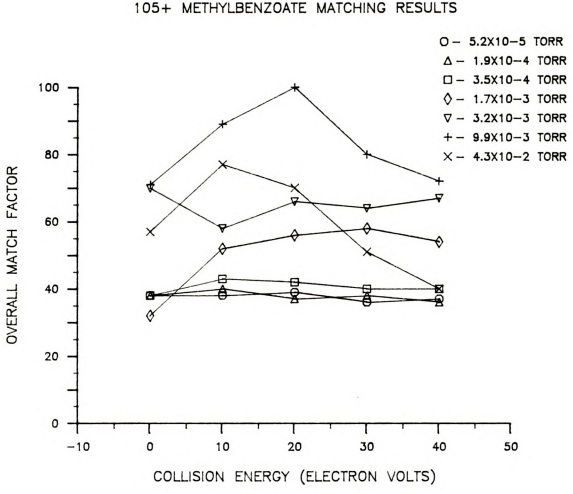


Figure 5.11 Instrumental Parameter Effects on the Overall Match Factor for 105+ Methylbenzoate Daughter Spectra

The effect of collision pressure on overall match factor is more pronounced than collision energy. Low overall match factors are obtained at low collision cell pressures (5.2×10^{-5} to 3.5×10^{-4} torr). Poor results are also obtained at high pressures (4.3×10^{-2} torr) and collision energies (30 - 40 eV) where non-first-order decomposition occurs in the collision cell.

The search program cannot compensate for spectral deviations due to decomposition mechanism changes in either the source or collision cell. Acceptable overall match factors are obtained, however, at all first-order collision cell pressures. Hence, MS/MS library spectra should be acquired in a collision cell pressure region where only first-order decomposition occurs.

Automated Resolution of MS/MS Mixtures

Many compounds analyzed are impure. If the chemist has not identified a compound, he is usually unaware of its purity. Therefore, the development of a method to resolve spectra from mixtures is helpful to the structure determination process.

Many algorithms have been invoked to resolve mass spectra arising from mixtures. Common techniques include spectrum stripping (6-8), multiple linear regressions (9-11), graphical rotation (12,13), pattern recognition (14,16), reverse searching (7,17), factor analysis (18,19), block-cutpoint tree methods (20) and parabolic fits to quotient

spectra (21). Most methods are limited to resolving components of a mixture whose spectral peaks do not overlap. This has serious drawbacks when analyzing mass spectra of mixtures of high molecular weight compounds.

By selecting a parent ion with the initial mass filter, fragmenting it, and recording its mass spectrum using a second mass filter, the triple quadrupole mass spectrometer provides good selectivity. This selectivity reduces the probability that a given spectrum will represent more than one compound. Many MS/MS spectra resulting from a single collision arise from single neutral losses (22). This results in a cleaner, simpler spectrum. Hence, there is less chance that spectral peaks of several components in the mixtures will overlap. These characteristics have enabled MS/MS to be successfully used for direct analysis of complex mixtures (23).

Given these capabilities, is it necessary to develop a means to resolve MS/MS spectral mixtures? In triple quadrupole MS/MS, a daughter spectrum is obtained by selecting a parent ion by m/z value in the first mass filter. Ions at this m/z are passed into a collision region and undergo collisionally induced dissociation. Intensities of the resulting fragment ions are recorded by a second mass filter as a daughter spectrum. If two mixture components fragment in the ion source to give ions of equal m/z values, both ions will enter the collision cell. The resulting daughter spectrum will be a mixture of fragments from these isobaric parent ions. Hence MS/MS mixture spectra do exist and must be resolved. A means to determine the presence of a mixture in a

conventional mass spectrum is also very useful. If the conventional mass spectrum is proven to be pure, no MS/MS mixture spectra resulting from more than one compound will exist. Isobaric parent ions must come from the same molecule.

The MS/MS spectral matching program was extended to resolve spectra due to mixtures of isomeric and isobaric parent ions, poor resolution MS/MS spectra (such as MIKES spectra (22)), and the presence of reagent ions in CI spectra. In particular, isomers of synthesized compounds tend to produce isobaric MS/MS mixture spectra. An example resolving this type of mixture will be demonstrated later.

Iterative parabolic fits to quotient spectra (unknown spectrum/candidate spectrum) were used to identify mixture components and to reduce the dependence of component identification upon instrumental parameters. This algorithm was designed to integrate with existing software tools. It takes advantage of the inverted data in the reference data base library and the several match factors already present in the spectral matching program.

The strategy of the algorithm is to quickly reduce the number of candidate spectra by eliminating those compounds whose mass spectra contain peaks <u>not</u> found in the mixture spectrum. An intensity-based matching algorithm then uses several match factors to identify the major component. Multiple parabolic fits to the quotient spectrum determine which peaks in the mixture spectrum are not completely accounted for by the major component. The portion of the peak intensity belonging to the

minor components is calculated and placed with the unmatched peaks in a residual spectrum. The residual spectrum is then matched against the library to identify the second component. This process is repeated until all components in the mixture have been resolved or until the residual spectrum cannot effectively be matched.

Two examples demonstrate the algorithm developed to resolve mixture spectra. The first example resolves a conventional mass spectrum arising from an ether mixture. The second example resolves a MS/MS mixture spectrum arising from isobaric parent ions.

A 60:40 spectral mixture of β,β-dibromodiethyl ether and diisopropyl ether was created by adding the two experimental mass spectra. This mixture spectrum was used to test the mixture resolution algorithm. Data from the inverted list of mass spectral peaks were used to logically reduce the number of possible component spectra. A list representing all spectra containing a m/z value not in the mixture spectrum were placed into a buffer. The list of spectra containing another m/z value (not in the mixture spectrum) were logically ANDed with the original list (Figure 5.12). This process continued until all spectra containing spectral peaks not present in the mixture spectrum were eliminated. The most frequent spectral peaks in the data base are present at lower mass values. Hence, the logical reduction process started at m/z zero and proceeded to m/z 500. Since, above mass 500 the frequency of the spectral peaks in the data base rapidly drops off, few unique spectra would be eliminated.

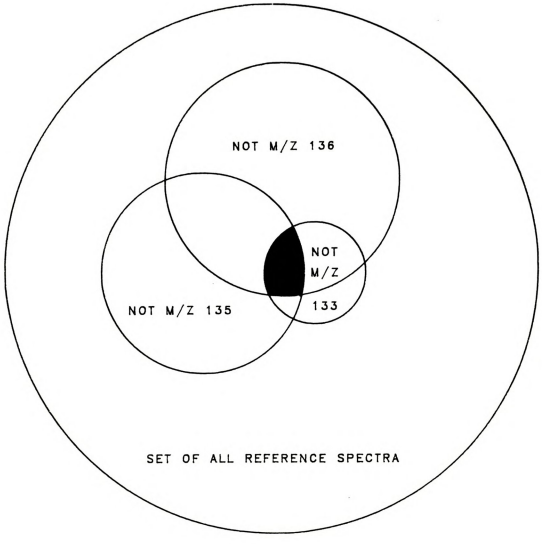


Figure 5.12 Logical Reduction of Candidate Spectra During Mixture Analysis (Venn Diagram)

As successive sets of these spectra are combined, the intersecting portions of all sets becomes very small. The data in Figure 5.13 illustrate how the number of candidate spectra decreased as successive inverted lists were ANDed together.

When the subset of candidate spectra was obtained, intensity-based matching determined the major component of the mixture. The reverse-search match factors, NR and IR, helped determine the major component, since they place no emphasis on peaks in the mixture spectrum that are not present in the reference spectrum. Table 5.3 lists the results of searching for the major component. The compound with the highest overall match factor (β,β-dibromodiethyl ether) was identified as the major component of the mixture. The match factors NR and IR are both zero, indicating that all peaks in the B,B-dibromodiethyl ether spectrum were present in the mixture spectrum. The minor component, Diisopropyl ether, was listed as the second best matching compound.

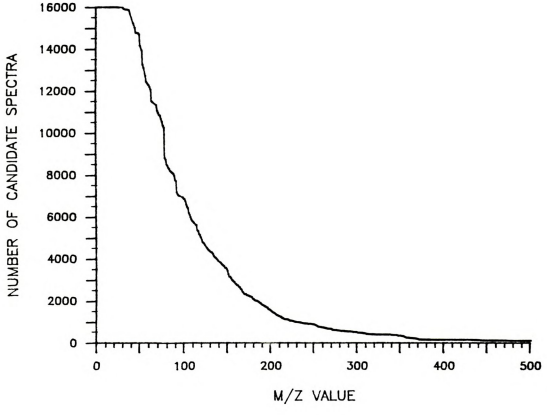


Figure 5.13 Logical Reduction of Candidate Spectra During Mixture Analysis (Stepping Through M/Z Values)

Table 5.3 Match Factors for Determination of the Major Component of the Ether Mixture

PT	PC	NC	NS	NR	IS	IR	Name
81	72	23	9	0	15	0	B,B--DIBROMODIETHYL ETHER
46	31	15	17	0	56	0	DIISOPROPYL ETHER
39	31	15	17	4	56	30	DIPHENYL SULFOXIDE
35	29	9	23	9	59	33	DIVINYL B,B--THIODIPROPIONATE
33	28	12	20	17	56	13	DIBUTYLDIFLUOROMETHYLPHOSPHINE
30	33	11	21	21	53	63	2,3--DIHYDROPYRAN
30	27	11	21	20	58	31	LEVULINIC ACID
29	22	9	23	9	62	47	HELTALDEHYDE
28	31	14	18	25	56	46	ISOEUGENOL
28	29	11	21	16	58	59	2--MERCAPTOETHANOL
27	22	8	24	7	69	64	METHYL N--BUTYRATE
26	27	11	21	21	47	46	P--FLUOROSTYRENE
24	22	9	23	15	68	62	2--DIMETHYLAMINOETHYL ACETATE
22	19	10	22	12	77	47	DIISOAMYL DISULFIDE
19	16	6	26	13	72	67	GAMMA--VALEROLACTONE

All peaks in the mixture spectrum that were not in the major component were then removed and placed into the residual spectrum. A quotient spectrum was created by dividing the intensities of the remaining peaks in the mixture spectrum (now termed the reduced component spectrum) by the peaks in the spectrum of β,β-dibromodiethyl ether.

Several peaks in the mixture result from overlap of the component spectra. A parabolic fit to the quotient spectrum determines if any portions of peaks in the mixture spectrum belong to minor components. Since the chi-square value was above a set threshold, the peak creating the largest discrepancy in the quotient spectrum (m/z 43.0) was removed. The chi-square value was then recalculated. This process was repeated until the chi-square value fell below a specified minimum.

When a suitable parabola was determined, the parabolic equation calculated the probable intensity of the removed peak (m/z 43.0) due to the major component placed it in the reduced component spectrum (Figure 5.14). The difference between the actual intensity and this value was placed into the residual spectrum. The remaining mixture spectrum was termed the reduced component spectrum. Intensity-based matching performed on the residual spectrum correctly found the minor component to be diisopropyl ether (Table 5.4).

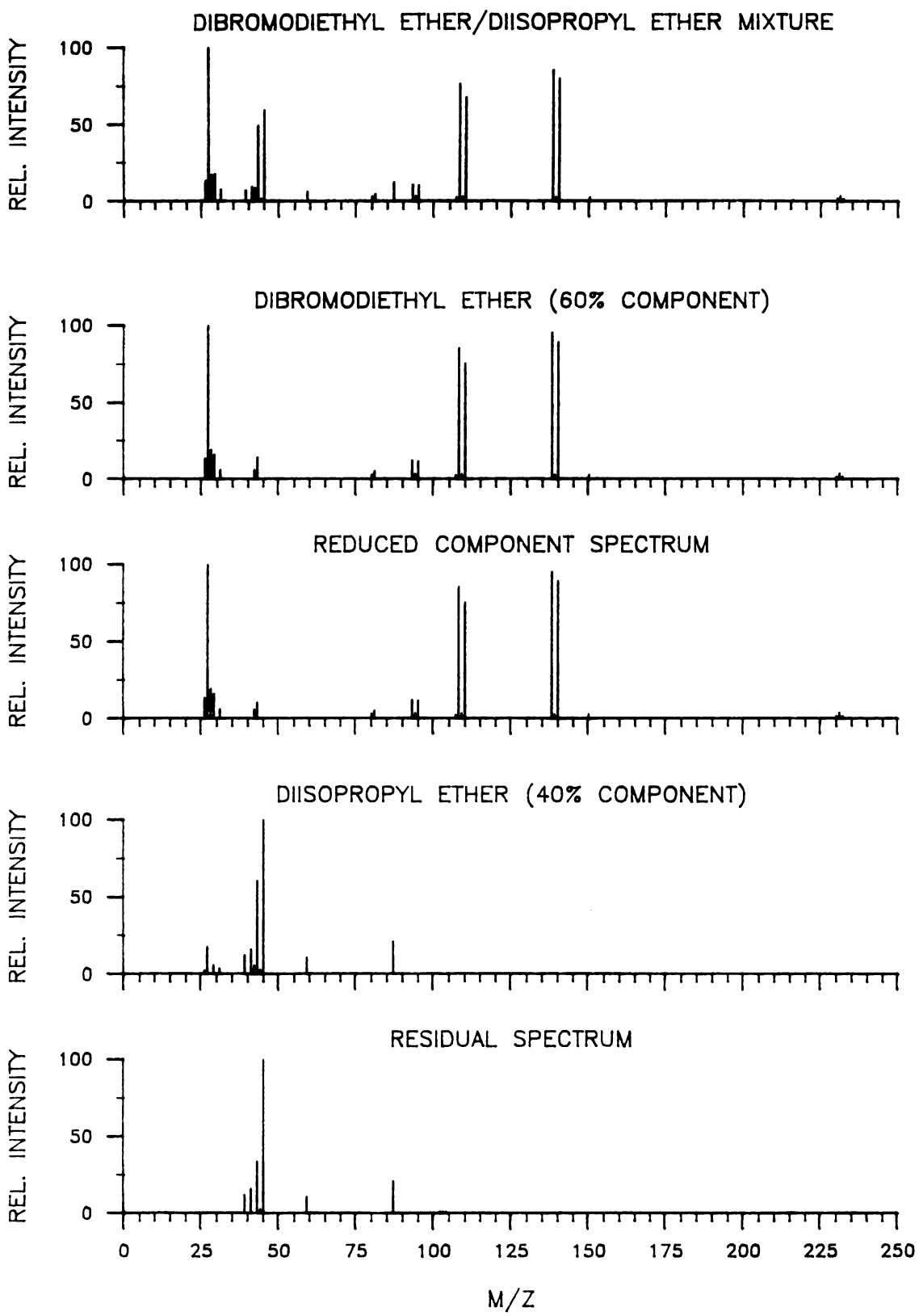


Figure 5.14 MS Ether Mixture Resolution

Table 5.4 Match Factors for Determination of the Minor Component of the Ether Mixture

PT	PC	NC	NS	NR	IS	IR	Name
71	83	10	0	5	0	14	DIISOPROPYL ETHER
50	75	10	0	9	0	51	DIPHENYL SULFOXIDE
40	66	9	1	13	1	48	DIISOAMYL DISULFIDE
27	46	5	5	26	18	61	LEVULINIC ACID
26	46	6	4	23	12	83	DIBUTYLDIFLUOROMETHYLPHOSPHINE
26	37	4	6	14	26	56	DIVINYL B,B–THIOPROPIONATE
25	33	4	6	11	68	85	METHYL N–BUTYRATE
23	38	4	6	20	67	83	2–DIMETHYLAMINOETHYL ACETATE
21	35	5	5	13	25	85	HELTALDEHYDE
21	38	4	6	23	27	80	2–MERCAPOETHANOL
18	44	6	4	26	20	88	2,3–DIHYDROPYRAN
17	40	6	4	33	15	80	ISOEUGANOL
16	34	4	6	28	19	78	P–FLUOROSTYRENE
16	27	3	7	16	69	87	GAMMA–VALEROLACTONE
15	28	3	7	24	69	81	BENZYL ACETATE

A 60:40 mixture of two pesitcide MS/MS spectra (Pirimiphos and the oxygen analog of chlorpyrifos) was created to test the mixture resolution algorithm on MS/MS spectral mixtures resulting from isobaric parent ions (m/z 332). The resulting component spectra determined by the mixture resolution algorithm are presented in Figure 5.15. Although MS/MS spectra of the mixture contains few peaks, the algorithm correctly deconvoluted the two spectra into the two components. No portion of parent ion intensity was placed into the residual spectrum. In the future, the algorithm will be modified to ignore the parent ion intensity in the spectrum stripping process.

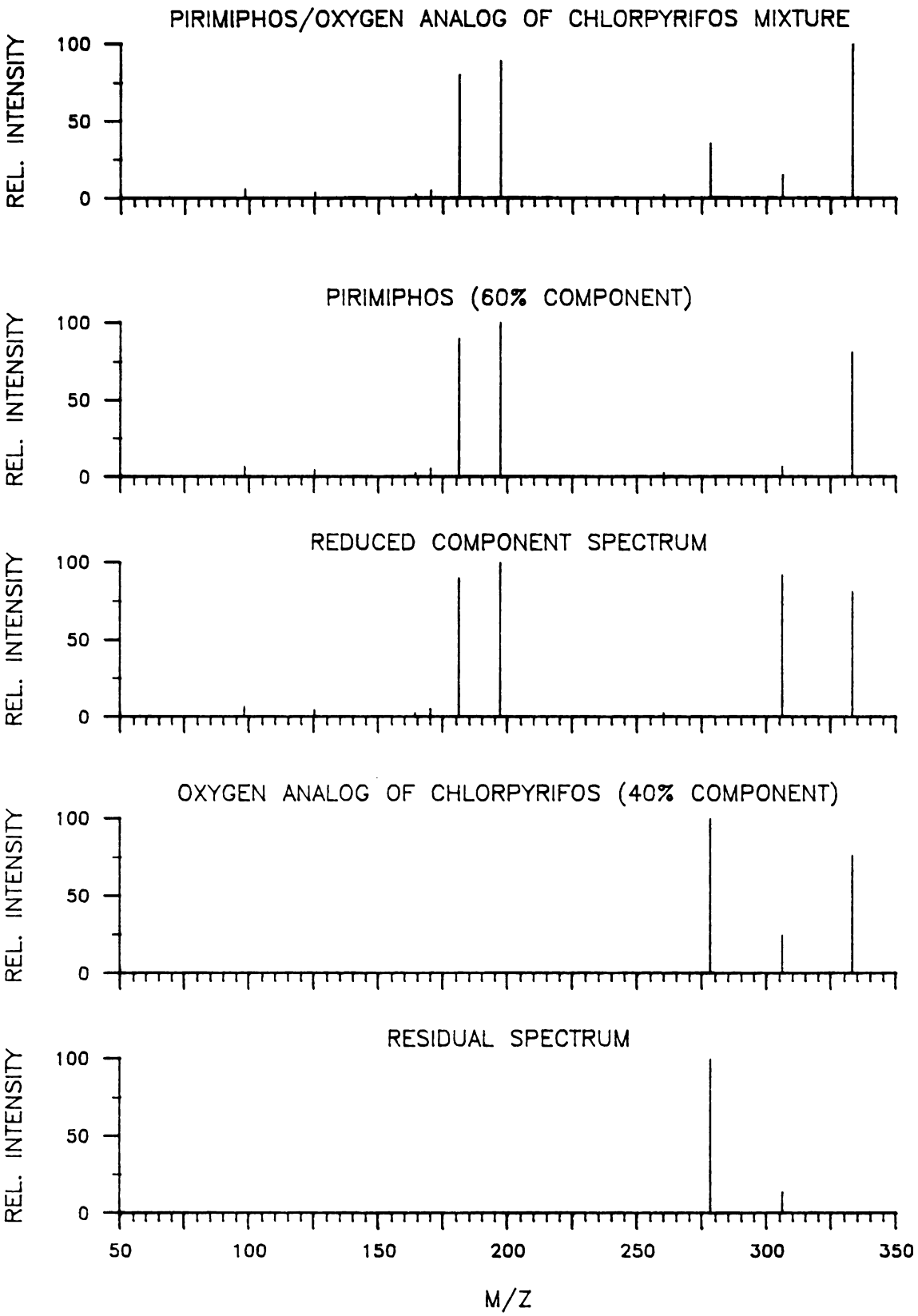


Figure 5.15 MS/MS Pesticide Mixture Resolution

Conclusions

The MS/MS search program has helped illustrate instrumental parameter effects on MS/MS spectral patterns. Different combinations of instrumental conditions may yield similar spectra. The ability of the program to correctly identify MS/MS spectra taken under different conditions has helped determine standard operating condition limits for acquiring MS/MS reference spectra. The most important instrumental parameter is collision cell pressure. MS/MS library should spectra be acquired in a collision cell pressure region where only first-order fragmentation occurs. Collision energy should be adjusted for a maximum CID efficiency (approximately 20 eV). Regarding drawout potential, Q2 should trail the Q2 voltage setting by a fixed negative amount (-10 V).

Results of the mixture resolution algorithm depend on the number of overlapping peaks from component spectra and the number of peaks in each component spectrum. As the number of overlapping peaks decreases, the process of mixture resolution becomes easier. However, if the number of peaks in the quotient spectrum decreases to less than 4, parabolic fits to the data are no longer profitable. If one component spectrum is a subset of another component spectrum, mixture resolution by this method becomes unattainable.

References

1. Dawson, P. H., presented at 31st Annual Conference on Mass Spectrometry and Allied Topics, Boston, MA, (1983); bound p. 203.

2. Dawson, P. H., Sun, W. F., **Int. J. Mass Spectrom. Ion Poc.**, 55, 155 (1983).

3. Wong, C. M., Crawford, R. W., Barton, V. C., Brand, H. R., Neufield, K. W., Bowman, J. E., **Rev. Sci. Inst.**, 54, 996 (1983).

4. Meg Osterby, unpublished work, MSU Chemistry Dept., East Lansing, Michigan.

5. Wong, C. M., Lanning, S., **Energy and Technology Review**, Lawrence Livermore National Laboratory, February, p. 8 (1984).

6. Dromey, R. G., Stefik, M. J., Ridfleish, T. C., Duffield, A. M., **Anal. Chem.**, 48, 1368 (1976).

7. McLafferty, F. W., **Acc. Chem. Res.**, 13, 33 (1980).

8. Atwater, B. L., Venkataraghavan, R., McLafferty, F. W., **Anal. Chem.**, 51, 1945 (1978).

9. Biller, J. E., Biemann, K., **Anal. Letters**, 7, 515 (1974).

10. Biller, J. E., Herlihy, W. C., Biemann, K., **Computer Assisted Structure Elucidation**, ACS Sym. Series, 54, 18 (1977).

11. Giblin, D. E., Peake, D. A., Lapp, R. L., presented at 32 Annual Conference on Mass Spectrometry and Allied Topics, San Antonio, TX, May (1984); bound p. 644

12. Windig, W., Meuzelaar, H. L. C., presented at 32 Annual Conference on Mass Spectrometry and Allied Topics, San Antonio, TX, May (1984); bound p. 665.

13. Ioup, G. E., Thomas, B. S., **J. Chem. Phys.**, 46, 3959 (1967).

14. Chien, M., **Anal. Chem.**, 57, 348 (1985).

15. Van der Greff, J., Tas, A. C., Bouwman, J., Tennuever de Brauw, M. C., Schrueurs, W. H. P., **Anal. Chim. Acta**, 150, 45, (1983).

16. Soltzberg, S. L., Kaberline, S. L., Lam, T. L., Brunner, T. R., Wilkens, C. L., **J. Am. Chem. Soc.**, 98, 7139 (1976).

17. Abramson, F. P., **Anal. Chem.**, 47, 45 (1975).

18. Clemens, J., Kowalski, B. R., **Anal. Chim. Acta**, 133, 538 (1981).

19. Ritter, G. L., Lowery, S. R., Isenhour, T. L., Wilkens, C. L., **Anal. Chem.**, **48**, 591 (1978).

20. Nakayama, T., Fujiwara, Y., **J. Chem. Inf. Comput. Sci.**, **21**, 142 (1981).

21. Henneberg, D., Wiemann, B., presented at 32nd Annual Conference on Mass Spectrometry and Allied Topics, San Antonio, TX, May (1983); bound p. 185.

22. Borogzadeh, M. H., Morgan, R. P., Beynon, J. H., **Analyst**, **103**, 1613 (1978).

23. Kondrat, R. W., Cooks, R. G., **Anal. Chem.**, **50**, 81A (1978).

CHAPTER VI

A STRUCTURE/SUBSTRUCTURE DATA BASE ASSOCIATED WITH MS/MS SPECTRA*

Abstract

A structure/substructure data base was developed to store substructure-property relationships determined using mass spectrometry/mass spectrometry (MS/MS) instruments. The structures and substructures are correlated with the MS/MS spectra in a separate spectrum data base which represent them. The substructures of each compound are not associated with any other substructure of the molecule or the structure of the compound. The structures are stored in connectivity matrices using an extended Morgan algorithm to generate a unique, unambiguous form allowing for representation of stereochemistry, charged species, radicals, and isotopic species. In addition, a unique, invariant linear name describing the molecule is generated. An heuristic drawing program was adapted to allow structures and substructures to be drawn in such a manner that common substructural features are easily recognized.

*Note: This chapter is a draft of a manuscript written by the author of thesis in preparation for submission to Computers and Chemistry with C. F. Beckner and C. G. Enke as coauthors.

Introduction

The development of two-dimensional techniques such as mass spectrometry/mass spectrometry (MS/MS) (1) has inspired development of data bases to handle the different types of related MS/MS spectra as well as the structures and substructures determined from each spectrum. MS/MS instruments allow the determination of a unique substructure-property relationship arising from each spectrum and require an information data base to store and maintain these relationships. To this end, a MS/MS spectrum data base and an associated structure/substructure data base were designed and developed. The spectrum data base is able to store and associate the various types of MS/MS spectra for each compound and is described elsewhere (2). The focus of this paper is the design and implementation of a structure/substructure data base that can manage the structural information generated from MS/MS instruments. Central to the discussion will be the design requirements for this particular application in contrast to the requirements of other structure and substructure data bases.

The structure determination scheme does not require the elucidation of ion structures; thus only molecular substructures need be stored (3). The structure determination scheme uses the substructures as building blocks to produce plausible complete molecular structures. Therefore, the substructures in the data base need not be associated with each other or the parent molecular structure. Instead each structure or substructure in the data base exists independently of each

other and references the MS/MS spectra in the spectral data base that are associated with it. Likewise, each MS/MS spectrum contains references to the structure or substructures that are identified by that spectrum. These links allow the MS/MS spectrum to be correlated with its respective substructure and a system of substructure-property relationships to be developed.

The structure/substructure data base was designed to store both molecular structures and substructures without no distinction in data base structure or data storage format. Hence any comments regarding structures are assumed to refer to substructures as well.

The substructure-property relationships determined by MS/MS are not unique. One daughter spectrum may correspond to several substructures. Likewise, one substructure may be associated with several daughter spectra. The relationship between the spectrum data base and the structure/substructure data base information is complex (n-to-n mapping) and is constantly changing as new information is added to both data bases. New information may include redundant substructures that have yet to be pruned out. Their structural pointers in the spectrum data base must be redirected to the sole remaining representation of that structure when the redundant entries in the structure data base are eliminated. Redundant structures in the data base are found by searching the data base for identical structures. The empirical formula of the structure serves to prefilter nonidentical structures. Due to the dynamic nature of the structural information, the structure data base was designed to be easily maintained and to efficiently reclaim

vacant space. This included easy entry and updating of information in the structure data base.

At the present time the structures are automatically drawn and then compared for structural similarities and commonalities. In the future an atom-by-atom substructural searching program will eliminate redundant structures through automated comparison (4,5).

Data Base Design

The design of the structure/substructure data base is illustrated in Figure 6.1. Flexibility and ease of maintenance are the most important features of the design. To accomplish this, a linked-list architecture was used for storing all data records in the data base (6,7). Variable lengths of information are represented using fixed length records. If the information overflows one record, a continuation pointer specifies another overflow record. The use of fixed length records conserves space while using linked-lists maintains flexibility. If the header records were to be extended at a later date, the data base would not have to be rewritten. A continuation pointer would merely specify the overflow record. Likewise if a structure were deleted and replaced with a larger structure, the space vacated by the deleted record could be reclaimed since a continuation pointer would allow the data to overflow into another record.

This flexible design allows the data base to expand and contract with varying amounts and types of data. In theory all vacated space may

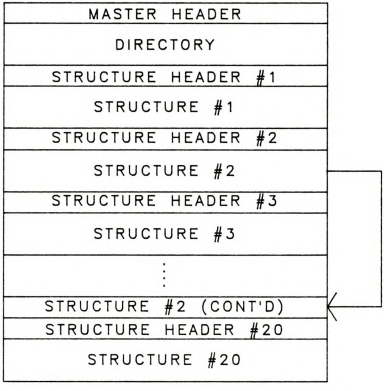


Figure 6.1 Structure Data Base Format

be reclaimed. However, when the number of internal continuation pointers becomes too large, the data associated with each structure becomes badly fragmented and the access speed of the information suffers. At this point the data base can be "merged" with an empty data base and the data restructured such that it resides in contiguous records. Any vacant space in the data base is also compressed out.

A master header contains all static information describing the data base (Figure 6.2). The structure of the data base is denoted by storing the version number of the data base management software, the size of the internal directory, and the master header size. The version number of the software helps maintain consistency during growth and revision of the software. The directory size parameter sets aside space for an internal directory and protects the directory from being written to unless new data are being entered. The size of the master header is fixed but may change if forthcoming versions of the software contain more static information. It is therefore maintained as a variable.

The master header maintains a historical record of the data base by containing the dates of data base creation and latest update. The highest assigned physical record number is kept to determine the next available physical record number. The highest assigned registry number determines the next directory entry (logical record) available for new data. The last registry number defines the limits of the internal directory and also the number of structures in the data base. There is no restriction on the length of the data base, only on the number of structures that may be contained within the data base.

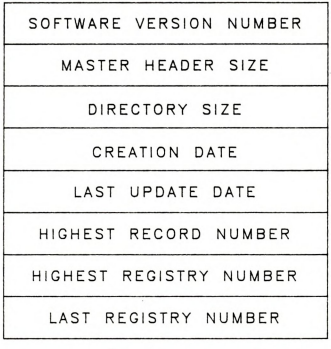


Figure 6.2 Master Header Record Format

An internal directory containing the physical record number for each structure header record follows the master header (Figure 6.1). This design maintains data independence between the logical structure registry numbers and the actual physical record numbers where the data are stored. This feature allows easy maintenance of the data base. When a structure is deleted, the directory entry for that structure is merely forgotten. New structures are sequentially assigned new registry numbers as they are added to the data base. The size of the directory dictates the maximum number of structures that can be stored in the data base. When the last registry number has been assigned, no further additions to the data base may be made. This design conserves overhead space while retaining quick access to the data records by requiring only a single disk read to obtain the location where the desired data resides.

A header record precedes the data record for each structure and identifies the structure (Figure 6.3). An arbitrary registry number is assigned to each structure and uniquely identifies the structure. This number corresponds to the entry in the internal directory for the same structure. The size of the structure header is fixed but an extension variable exists for flexibility. A status variable in the header maintains the current status of the structure and it identifies the data as either a complete structure or a substructure; and specifies whether it is logically deleted or not. This is the only piece of data in the entire data base which distinguishes structures from substructures. All other storage criteria are the same.

REGISTRY NUMBER
HEADER SIZE
STRUCTURE STATUS
STRUCTURE SIZE
INSERTION DATE
UPDATE DATE
CHEMICAL ABSTRACTS NUMBER
WILEY NUMBER
EMPIRICAL FORMULA
NATOMS
NBONDS
NRINGS
NISO
NSCHEM
NMODS
LNKHDR
LNKDAT
REGDAT

Figure 6.3 Structure Header Record Format

The size of the structure is maintained in terms of data records since continuation pointers allow the size to vary. The insertion date and update date are included in the header to help maintain the history of the structure. The Chemical Abstracts Service number is stored to identify complete structures and serves as a key when retrieving molecular structures. In the case of substructures, this variable is an arbitrary but uniquely assigned substructure number which also serves as a search key for substructure retrieval. Hence the structure and substructure retrieval routines are one and the same. The Wiley number of the MS/MS spectrum representing the structure in the spectra data base is stored as a cross-reference. The empirical formula is stored even though the information is redundant with that contained in the structure connectivity tables. In addition to informing the user of the number of hydrogens present in the molecule, the empirical formula serves to double check the integrity of the data and as a prefilter when comparing structures.

The variables NATOMS, NBONDS, NRINGS, NISO, NSCHEM, and NMODS describe the characteristics of the structure by specifying the sizes of the connectivity matrices. NATOMS specifies the number of atoms in the structure; NBONDS specifies the number of bonds; NRINGS specifies the number of ring closure bonds; NISO specifies the number of isomeric atoms; NSCHEM specifies the number of stereochemical isomers; and NMODS specifies the number of modified atoms in the structure. Modified atoms are those containing charges, free valences, or consisting of an isotopic mass.

Lastly, the header maintains continuation pointers for each header (LNKHDR) and data record (LNKDAT). The array REGDAT contains the registry numbers in the spectra data base for spectra representing this structure. Since several spectra may be identified with a single structure this array may contain up to ten pointers. If more spectra pointers are needed, the size of the header may be expanded by using the continuation pointer LNKHDR.

Structure Storage Format

The representation form of structures is central to the performance of the data base (Figure 6.4). The criteria for choosing a storage format includes the ability to uniquely and unambiguously represent the structure as compactly as possible. There are several methods for representing structures - Wiswesser line notation (WLN) and connectivity matrix methods being two of the most popular methods (8,9). While various linear notations (WLN) are unique and compact they have difficulties representing large molecules and stereochemistry. Due to their flexibility in storing molecules in a non-linear fashion, connectivity matrices have become popular for storing large, complex structures. A drawback to using connectivity matrices is that they commonly represent identical structures in a number of different ways. They also require a relatively large amount of space.

An expanded version of the Morgan algorithm (10) was chosen to encode the structures. This algorithm uses a set of connectivity matrices to represent each molecular structure or substructure in a

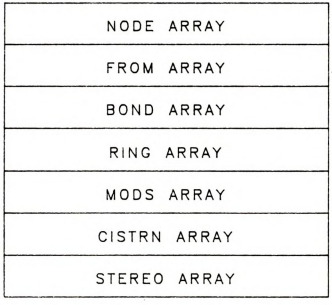


Figure 6.4 Structure Storage Record Format

unique, unambiguous fashion. In addition, the structure can be recalled as an expanded linear name to easily determine identical structures.

The version of the Morgan algorithm implemented includes revisions made by Todd and Wipke and by our own lab. Todd and Wipke used the Morgan naming algorithm to generate a unique linear name for each molecule including representation for cis/trans and stereochemical isomers (11). They expanded the atom numbering scheme developed by Morgan which classifies atoms based on the number of non-hydrogen substituents attached to each atom. Wipke's method calculates the location of each atom relative to the center of the molecule by observing the substituents attached to each atom. These extended connectivity values, along with matrices describing the stereochemistry of the molecule, generate a stereochemically unique name whereby only one form of the connectivity matrices uniquely determines the molecular structure.

The number of elements encoded in Wipke's algorithm was expanded by Carl Beckner from the original organic elements to include all the known elements. Carl Beckner's algorithm transformed the structure from the input matrix into a unique representation in the connectivity matrices. The bond types originally available included aromatic, single, double, and triple bonds. This notation was expanded to include tautomer bonds, ionic bonds, and structural discontinuities and to allow the representation of free valence substructures as well as polymer structural units. The connectivity matrices are large enough to include any molecule up to 128 atoms is size (excluding hydrogens).

Several connectivity matrices are used to describe a structure (Figure 6.4). The NODE array contains the elemental symbols for each atom. The FROM array contains the lowest numbered atom attached to each NODE atom. The BOND array describes the bond between the NODE atom and the FROM atom. The RING array is a two-dimensional array where each pair of elements corresponds to the two atom connectivity numbers closing a ring. One pair of elements exists for each ring present in the structure. The MODS array denotes any modifications for the NODE atom such as charge, free valence, and isotopic masses. The CISTRN array denotes the cis/trans arrangement for a three bond system and is a two-dimensional array whose second dimension contains three elements. Lastly, the STEREO array denotes the stereochemical arrangement of three atoms about a chiral center.

Characteristics and Operation

The structure/substructure data base currently has over 30,000 structures and substructures that correspond to MS/MS spectra in the reference data base library. The majority of the information are structures that correspond to their normal (electron-impact) mass spectra. The remaining substructures are associated with the various MS/MS spectra in the spectral library.

Several functions enable the retrieval and manipulation of the data in the structure data base. A new data base is created by initializing the master header and the internal directory. New

structures are entered manually or by using a designated file format. Structures may be logically deleted and later undeleted if needed. A merge operation physically compress out logically deleted files and also merges structures from two different data bases. An update operation replaces an existing structure with a revised structure by reclaiming the space used by the original structure and overflowing into another record if needed.

The most useful operation of the program is the display of structures. A tabular dump function allows the structure to be retrieved by structure number, Chemical Abstracts number, or spectrum data base registry number. Structural data are presented in a readable tabular form. The output may be directed to either the terminal, the printer, or an output file. In addition, several structures may be displayed with a single command. An example of the tabular output is given in Figure 6.5 for the compound n-butylbenzene. In addition to tabular output, graphical output to one of several graphics devices is available. An heuristic structure drawing program (DRAWC2) that was developed by Shelley (12) has been incorporated to draw the structures. Several structures were drawn using this program are presented in Figure 6.6.

DRAWC2 initially perceives the ring systems in the molecule and assigns spatial coordinates such that they are conventionally displayed; oriented as the chemist would normally draw them. The heuristics of the program eliminates any overlapping bonds or atoms, maintains correct bond lengths, and tries to minimize atom crowding. Identical and

```
Software version:  1
Creation date: 13-JUL-85
Last update date: 13-JUL-85
Last record number:  110160
Highest assigned structure number:    32014
Maximum number of structures:    40000
Structure number:          322
Structure type (Structure=1, Substructure=2):  1
Empirical formula: C10 H14
Insertion date: 13-JUL-85
Last update date: 13-JUL-85
Number of data records:      1
Chemical abstracts number:      104518
Wiley number:        6448
Number of atoms:  10
Number of bonds:  11
Number of rings:   1
Number of associated spectra:     1
Spectra registry #     4043

Atom #  1 type: C
Atom #  2 (C ) is single   bonded to atom #  1
Atom #  3 (C ) is single   bonded to atom #  1
Atom #  4 (C ) is single   bonded to atom #  2
Atom #  5 (C ) is aromatic bonded to atom #  3
Atom #  6 (C ) is aromatic bonded to atom #  3
Atom #  7 (C ) is single   bonded to atom #  4
Atom #  8 (C ) is aromatic bonded to atom #  5
Atom #  9 (C ) is aromatic bonded to atom #  6
Atom # 10 (C ) is aromatic bonded to atom #  8
Ring closure #  1: Atom #  9 attached to atom # 10
```

Figure 6.5 Structure Representing N-butylbenzene

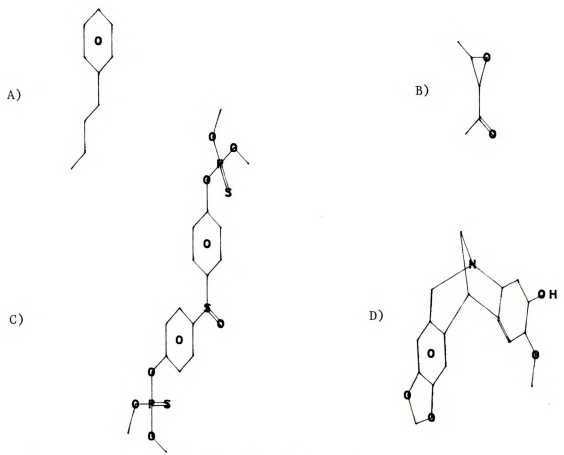


Figure 6.6 Structures output from DRAWC2.
A) n-Butylbenzene, B) 1-(3-methyloxiranyl)- Ethanone
C) O,O,O,O,-(sulfinyldi-4,1-phenylene) O,O,O,O,-tetramethylester Phosphorothoic Acid
D) O2-methyl-Pancracine

similar structures are similarly represented so that the chemist can perceive the commonalities of the structures. DRAWC2 was adapted to take advantage of the graphics devices in our lab. Structures are output in stick fashion with carbons atoms unlabelled. A drawing option allows all atoms, or just carbon atoms, to be tagged with sequence numbers.

To provide flexibility and transportability, the structure data base software was written in FORTRAN-77, and C. The program is currently implemented on a LSI 11/23 minicomputer running the RSX-11M operating system in a multi-user environment. A large capacity, 474 megabyte disk drive with an average access time of 18 milliseconds is used to hold the structure and spectral data bases.

Summary

The ability to acquire an MS/MS spectrum and to retrieve and draw associated substructures is central to the structure determination process. The structure/substructure data base allows the substructure-property relationships determined by MS/MS instruments to be represented and developed. Substructures determined through spectral comparisons form the building blocks for generating possible molecular structures.

References

1. Yost, R. A., Enke C. G., **J. Am. Chem. Soc.**, **100**, 2274 (1978).

2. Hoffman, P. A., Beckner, C. F., Enke, C. G., in preparation.

3. Cross, K. P., Palmer, P. T., Beckner, C. F., Giordani, A. B., Gregg, H. R., Hoffman, P. A., Enke, C. G., Accepted by **ACS Symposium Series**.

4. Varkony, T. H., Shiloach, Y., Smith, D. H., **J. Chem. Inf. Comp. Sci.**, **19**, 104 (1979).

5. Cone, M., Venkataraghavan, R., McLafferty, F. W., **J. Am. Chem. Soc.**, **99**, 7668 (1977).

6. Heller, S. R., **Anal. Chem.**, **44**, 1951 (1974).

7. deHaseth, J. A., Woodruff, H. B., Lowry, S. R., Insenhour, T. L., **Anal. Chim. Acta**, **103**, 109 (1978).

8. Wiswesser, W. J., **Comput. Automat.**, **19**, 2 (1970).

9. Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V., Duffield, A. M., Djerassi, C., **J. Am. Chem. Soc.**, **91**, 2973 (1969).

10. Morgan, H. L., **J. Chem. Doc.**, **5**, 107 (1965).

11. Wipke, W. T., Dyott, T. M., **J. Am. Chem. Soc.**, **96**, 4834 (1974).

12. Shelley, C. A., **J. Chem. Inf. Comp. Sci.**, **23**, 61 (1983).

CHAPTER VII

FUTURE DEVELOPMENTS

The automated structure determination system utilizing MS/MS spectra has advanced to the point where the software tools can be routinely used in elucidating molecular structures. The temptation exists to leave the software tools unchanged and to judiciously proceed to determine structures for many compounds. While this urge may be inviting, further development and revision of the software tools should still continue.

Our software tools have evolved as new and better ideas were suggested. Much of the software has been written from the "ground up" without the benefit of previous examples. In addition, the design of some of the tools has progressed beyond their original goals. The MDDB data base was not designed to handle thousands of spectra. Likewise, the spectral matching program was not originally designed to handle multiple data base formats or to perform automated resolution of MS/MS mixture spectra. The phrase, "hindsight is 20/20", becomes appropriate when evaluating the capabilities and efficiency of the software tools. Hence those whose responsibility it becomes to maintain and upgrade the software should consider the following suggestions.

The spectra data base management software was written by Phil Hoffman in MACRO-11 assembly language for optimization on a PDP 11/40 minicomputer. Since that time our laboratory has upgraded through three generations of DEC minicomputers to the current micro-VAX I workstation.

Since the trend of hardware will continue towards faster computers at cheaper prices, all the software tools developed in our lab should be as transportable and upwardly mobile as possible. This statement is particularly true of the reference spectrum data base management software as it currently cannot be implemented on our micro-VAX. This software should be rewritten as soon as possible to avoid maintenance of archaic, out-of-date software.

In addition, the structure data base should be combined with the spectrum data base under a new architecture combining the best features of both data base management programs. This action will decrease the number of programs and the number of data bases to be maintained. We now have the disk space available to keep all spectra and structural information as a single file.

The matching program has suffered from the evolution process and 32K word memory restrictions. If this program were implemented on the micro-VAX the performance of the spectral matching routines would increase dramatically without sacrificing any of its many features.

Neutral-loss spectra provide valuable information in substructure determination. The approach presented in this work uses neutral-loss information in a very limited manner. The wide spread use of neutral-loss information in spectral matching will greatly enhance the determination of substructures.

A substructural searching program needs to be implemented to determine the largest common fragments of any two structures or substructures. This program will determine the structural fragments that are given to the GENOA program as well as in identify redundant structures or substructures in the structure/substructure data base. I strongly urge the developer of this software to consider the approaches presented in the literature and especially those referenced here (1-9).

The integration of the micro-VAX into the mass spectrometry information management system should be exploited as fully as possible. There are many applications where the computing power of this computer excels. These include GENOA, IONSIM and other compute bound tasks.

The average access time of the micro-VAX disk drives are slow (78 milliseconds) relative to the Fujitsu Eagle disk drive now on the MS/MS 11/23 (18 milliseconds). Hence I/O bound tasks will run slower on the micro-VAX than on the PDP 11/23. These include spectra and structure data base management software. If these tasks are run on the micro-VAX they should use the faster disk on the PDP 11 as a file server over the Ethernet network while taking advantage of the faster micro-VAX processor.

The Xerox 1108 computer should serve as a base to write and develop a knowledge-based expert system for MS/MS. An expert system that evaluates the effect of operating conditions upon MS/MS spectra will represent a milestone in the development of standard operating conditions for MS/MS. Such a system could spur the development of badly

needed community-wide MS/MS libraries. A secondary application of the Xerox computer could be the development of an expert system to replace GENOA. Molecular Design Ltd. is no longer supporting this software and will not maintain it. An expert system would run more efficiently than GENOA and could be updated as we determine new heuristic rules and conditions.

The Xerox computer could play an integral part in the development of AI guided instrumentation. It should be linked to the microcomputers controlling the MS/MS instrument as well as the minicomputers. It should complete a feedback loop to the instrument where it can make decisions regarding the information desired and then instruct the MS/MS instrument to perform experiments acquiring such information.

References

1. Willet, P., J. Chem. Inf. Comput. Sci., 24, 29 (1984).

2. Dromey, R. G., J. Chem. Inf. Comput. Sci., 18, 222 (1978).

3. Adamson, G. W., Cowell, J., Lynch, M. F., J. Chem. Doc., 13, 153 (1972).

4. Bawden, D., J. Chem. Inf. Comput. Sci., 23, 14 (1983).

5. Varkony, T. H., Shiloach, Y., Smith, D. M., J. Chem. Inf. Comput. Sci., 19, 104 (1979).

6. Cone, M. M., Venkataraghavan, R., McLafferty, F. W., J. Am. Chem. Soc., 99, 7688 (1977).

7. Willet, P., J. Chem. Inf. Comput. Sci., 25, 114 (1985).

8. Synge, R. L. M., J. Chem. Inf. Comput. Sci., 25, 50 (1983).

9. Kudo, Y., Chihara, H., J. Chem. Inf. Comput. Sci., 23, 109 (1983).