# An Enlarged Data Base of Electron-Ionization Mass Spectra

Fred W. McLafferty, Douglas B. Stauffer*, Andrea B. Twiss-Brooks, and Stanton Y. Loh*
Department of Chemistry, Baker Laboratory, Cornell University, Ithaca, New York, USA

The computer-searchable data base of reference mass spectra described earlier has been increased in size by 76%, so that it now contains 139,859 different spectra of 118,144 different compounds. The average number of peaks per spectrum is 53. All spectra were examined for errors by the Probability Based Matching (PBM) and the Quality Index (QI) algorithms and by human inspection. An improvement to the QI algorithm is based on the Terwilliger suggestion concerning saturated spectra. The number of different elemental compositions of compounds has increased by 64%. By using unknowns from the original data base with PBM, the probability that these incorrectly match a new spectrum is only 33% of that of incorrectly matching a spectrum in the original data base, further demonstrating that the variety of data in the library has been substantially expanded. Including additional reference spectra (measured under different conditions) of the same compound in the data base reduced the proportion of incorrect best-matching spectra by 42%. *(J Am Soc Mass Spectrom 1991, 2, 432–437)*

Nearly two decades have passed since the untimely deaths of Professors Einar and Stina Stenhagen, two pioneers in the characterization of complex organic molecules by using mass spectrometry. Twenty-five years ago Einar Stenhagen and the senior author began collaborating in the collection and evaluation of the first computer-searchable data base of reference mass spectra. With wonderful cooperation from colleagues around the world in supplying spectra, and careful coordination and checking by many of our students, the forerunner of our current data base was published by Wiley-Interscience [1]. The computer expertise of Professor Sixten Abrahamsson, now also deceased, was a key to the compilation of these 6,800 different spectra. Their availability on magnetic tape was a great impetus to the development of computer programs for the automatic identification of unknown mass spectra, with the first program by Abrahamsson [2] showing impressive advantages over earlier systems, such as that using Hollerith cards [3]. This article is dedicated to these pioneers whose foresight and initiative were critical to the present utility and comprehensiveness of this data base.

## Introduction

Literally thousands of mass spectrometers are now sold annually for routine analytical problems, and these can generate millions of unknown mass spectra per year. It was long ago recognized that computer identification was an absolute necessity to attack this problem, and impressive progress has been made in the development of such algorithms [2–15]. However, it is axiomatic that the algorithm performance can be no better than the quality and diversity of the reference spectra. If the reference spectrum corresponding to the unknown is not in the data base, or is incomplete or contains errors, even the cleverest algorithm will be compromised.

By far the most widely used reference files have been our previous update, the *1982 Registry of Mass Spectral Data* [16], the collection distributed by the National Bureau of Standards, now the National Institute of Science and Technology (NIST) [17, 18], and the *Eight Peak Index* [19] extracted from 66,720 mass spectra. In a 1989 collaboration the spectra of 112,272 compounds were published in seven bound volumes [20]; of these, 73,978 are from the Wiley collection, 6,117 from the NIST collection, and the balance from both. This article reports a more detailed study of an expanded version of this collection.

## Description and Discussion

The *1982 Registry of Mass Spectral Data* contained 79,560 different spectra of 67,128 different compounds representing 24,290 different elemental compositions. The *1989 Registry* discussed here now contains 139,859 different spectra of 118,144 compounds (3,893 of which are isotopically labeled) representing 39,854 different elemental compositions. The average molecular

*Present address: Palisade Corporation, 31 Decker Road, Newfield, NY 14867.
Address correspondence to Fred W. McLafferty, Baker Chemistry Laboratory, Cornell University, Ithaca, NY 14853-1301.


1044-0305/91/$3.50

Received May, 21, 1990
Accepted February 13, 1991

weight has increased to 271 from 262 in 1982. This contains all spectra of the 1989 NIST collection, which uniquely contributed 17,537 different spectra of 14,271 different compounds. Other large individual contributions of spectra have been acknowledged in the printed edition [20]. The distribution of these spectra as a function of the number of peaks contained in each is shown in Figure 1.

*Registry Numbers of the Chemical Abstracts Service* (Chemical Abstracts Services, 2540 Olentangy River Road, Columbus, OH 43210) have been assigned for the structures of 118,637 spectra and 97,510 compounds, while the corresponding Chemical Abstracts Services (CAS) structural images are available for 112,410 spectra of 91,529 compounds. The file contains 279,457 compound names, including trivial and trade names; for each spectrum of a compound all names submitted on all spectra are listed, plus all names from the CAS file. For the first name listed in many cases the standard CAS nomenclature has been replaced by a more common name selected by the authors, such as "styrene" and "cocaine" for "benzene, ethenyl" and "8-azabicyclo[3.2.1]octane-2-carboxylic acid, 3-(benzoyloxy)-8-methyl-, methyl ester, [1-(exo,exo)]-", respectively.

*Identifying spectra of the same compound.* Location of other spectra of the same compound in the file was first done by comparison of CAS Registry Numbers, when available, and then by human inspection of all compound names in the collection of the same elemental composition (molecular formula). A modified (see below) Quality Index (QI) algorithm [21] was used to select the best spectrum of each compound for the "unique" part of the file and for the separate publication in book form [20]. If all isomers of the compound were represented in the file, plus a spectrum of the compound of undesignated isomeric identity (which could thus have a different CAS Registry Number), the latter spectrum was not included in the unique file unless its QI was 0.5 higher (maximum QI value is 1.0) than that of any isomer. To aid in identifying other spectra of the same compound, each spectrum was compared with all others of the same elemental composition using the Probability Based Matching (PBM) algorithm [9, 14, 15, 22]. Previously [16], if two or more spectra of the same compound were very similar, as shown by a close PBM match [14], only the spectrum of highest QI [21] was retained in the collection; spectra removed earlier for this reason were readded so that the data base would be more representative of the statistical variation of spectra of the same compound run under different conditions. ("Exact duplicates" [16] that must represent separate collection of the same original spectrum are still eliminated.) Data on the 12,897 compounds in the file represented by multiple spectra are given in Table 1.

**Table 1.** Compounds represented by multiple spectra

| Number of spectra of the compound | Number of compounds 1982 | 1989 |
|---|---|---|
| 2 | 4772 | 8842 |
| 3 | 1292 | 2097 |
| 4 | 548 | 878 |
| 5 | 276 | 424 |
| 6 | 136 | 233 |
| 7 | 91 | 167 |
| 8 | 47 | 96 |
| 9 | 35 | 59 |
| 10 | 28 | 37 |
| 11 | 11 | 32 |
| 12 | 6 | 18 |
| 13 | 2 | 8 |
| 14 | 2 | 1 |
| 15 | | 2 |
| 16 | 1 | 1 |
| 17 | | 1 |
| 20 | | 1 |
| Total | 7247 | 12897 |

*Minimizing data storage requirements.* The full data base of 139,859 mass spectra requires 100 Mbytes of storage space in its normal form. To conserve space and, especially, to make possible its use with small computer systems, an alternative file was prepared. Data not directly needed for PBM, such as descriptions of the instrument, temperatures (ion source, inlet, sample), electron energy, ion accelerating voltage, pressure, and substructures, were eliminated. Values of mass-to-charge ratio representing a unitary increment from the previous value were deleted. The 279,457 names were sorted to find the most common subunits

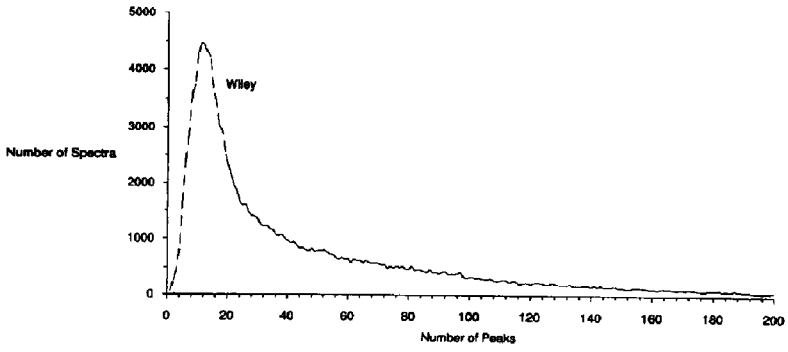


**Figure 1.** Number of spectra in the *1989 Registry* as a function of the number of peaks tabulated for each spectrum.

(methyl, phenyl, etc.), which were then replaced by bit codes ordered by occurrence frequency (Huffman code procedure) [23]. Only the first listed name is included for a compound. With these condensations the data base requires 23 Mbytes of storage. The condensed spectrum file containing only those peaks used in the initial PBM matching requires 10 Mbytes, but the "forward searching capability" [22] of PBM requires the complete spectrum of the best matches. For efficient reference exploration of the spectra by the mass spectrometrist ("browsing"), all names and molecular formulas can be rapidly accessed by using an inverted file structure; these files and the data for full matching capability usable on a personal computer require less than 74 Mbytes of disk storage. (Palisade Corporation, 31 Decker Road, Newfield, NY 14867).

*Minimization of errors.* The combined file is based on an original collection of approximately 180,000 different mass spectra. Similarly, for this update, ~ 75,000 spectra were collected, with ~ 15,000 discarded as totally unreliable or because far better spectra of the same compound were already in the file. Approximately 20,000 errors were also corrected in all the spectra, new and old, during this update. Thousands of errors must remain, as this represents only a tiny fraction of the data in the file, which contains $7.4 \times 10^6$ mass and abundance pairs. Many errors were found in the human inspection of all spectra of the same elemental composition, aided by the PBM matching of the spectra. The QI program [21] itself is designed to locate anomalies such as impurity peaks, illogical neutral losses, and incorrect isotopic abundances; some 8,000 of the latter alone were corrected. In a final inspection by the senior author the spectra of more common elemental compositions were sorted as to compound type [20], with spectral correlation of similar compounds for error identification. However, all data thought to be valid were retained; extensive experience has shown that a reference spectrum containing only the most important peaks can still be invaluable for unknown identification.

*Quality Index.* Our original algorithm for calculating the QI value [21] was modified following the proposal of Terwilliger et al. [24] to identify spectra containing "saturated peaks." If the signal for several peaks exceeds the capacity of the data system, all will appear to have the maximum (100%) abundance, and thus the relevant abundances of the other peaks will be erroneously high (this would also be true if only the base peak is "saturated," but our method will not detect this artifact). It was shown earlier that the occurrence probability as a function of peak abundance closely follows the expected log normal relationship, with the probability of a peak of 1–3.4% abundance being $2^5$ times that of one of 73–100% abundance [25]. To identify spectra that possibly contain such saturated peaks, the number of peaks of $\geq 1\%$ abundance in the spectrum is divided by 64, and this number is subtracted from the number of peaks of $\geq 95\%$ abundance, giving the difference value $N$. The distribution of such spectra for $N > 0$ is shown in Table 2. Inspection of individual spectra of higher $N$ values where other spectra of the same compound were in the file showed that most apparently do contain "saturated" peaks. The new Quality Factor 8 is calculated by eq 1, with $QF8$ as unity for $N \leq 1$.

$$QF8 = (11 - N)/10 \quad (1)$$

*Matching multiple spectra of the same compound.* A key policy for the *Registry* and its predecessors has been to collect multiple spectra of the same compound [1, 16, 20]. This allows an optimum selection from those spectra available for specific experimental conditions. Further, leaving such other spectra of the same compound in our previous data base [16] increased the probability of finding a correct match much more than it increased the probability of retrieving false positives. Thus, in matching 385 randomly selected unknowns not in the data base against 64,376 spectra (not isotopically labeled), each of a different compound, the first match was correct (class I: same compound or a stereoisomer [9]) for 61% of randomly selected unknowns. Increasing this file to 76,673 spectra through the addition of other spectra of these compounds increased the number of correct answers to 74%. With the new data base, PBM [9, 14, 15] improved with our new forward-searching procedure [22] was used to match the same 385 randomly selected unknowns [26]. For the data base of 114,418 different unlabeled compounds, the first answer was correct (class I) in 69% of the cases. Adding the other spectra of these compounds to give a data base of 135,953 mass spectra increased the proportion of correct first answers to 82%, surely a dramatic demonstration.

A more comprehensive evaluation of retrieval procedures [27, 28] compares the recall (*RC*, the fraction

**Table 2.** Spectra with excess peaks of large abundance (> 95%)

| N[a] | Number of spectra[b] |
|---|---|
| 1 | 11870 |
| 2 | 1192 |
| 3 | 227 |
| 4 | 69 |
| 5 | 30 |
| 6 | 16 |
| 7 | 5 |
| 8 | 3 |
| 9 | 5 |
| 10 | 3 |
| 11 | 3 |
| 12 | 1 |
| 28 | 1 |
| 47 | 2 |

[a](Number of peaks of ≥ 95% abundance − number peaks of > 1% abundance)/64.
[b]Number for N ± 0.5 (rounded value).

of correct answers retrieved) observed at matching criteria producing different values of reliability (*RL*, the fraction of retrieved answers that are correct) by using eq 2, where $P_c$ and $P_f$ are the possible number of

$$RL = P_c \cdot RC/(P_c \cdot RC + P_f \cdot FP) \qquad (2)$$

correct and false answers, and *FP* (false positives) is the fraction of incorrect answers retrieved. This "recall-reliability" plot shows (Figure 2) in more detail the value of using multiple spectra of the data base compounds in matching. At > 90% predicted reliability (30% recall) using class IV criteria (structural differences for which mass spectrometry is insensitive) [9, 26], the use of multiple spectra reduces the number of wrong answers to less than half.

*Effect on performance of increased file size.* The search time requirement is certainly increased substantially by increasing the size of the data base by 76%. However, this increase is more than offset by gains in computer speed since 1982. By using a 33 MHz 80486-based personal computer (Palisade Corporation), PBM search times for the 139,859 spectra average less than 3 s using our two-level file-ordering technique [14, 29].

Increasing the file size must also increase the possibility of false answers. Although this might appear to be a disadvantage, reducing the reference file to zero spectra to retrieve no false answers is surely not advantageous. However, doubling the size of the reference file does not necessarily double the *FP* value; this should happen only if the new spectra have the

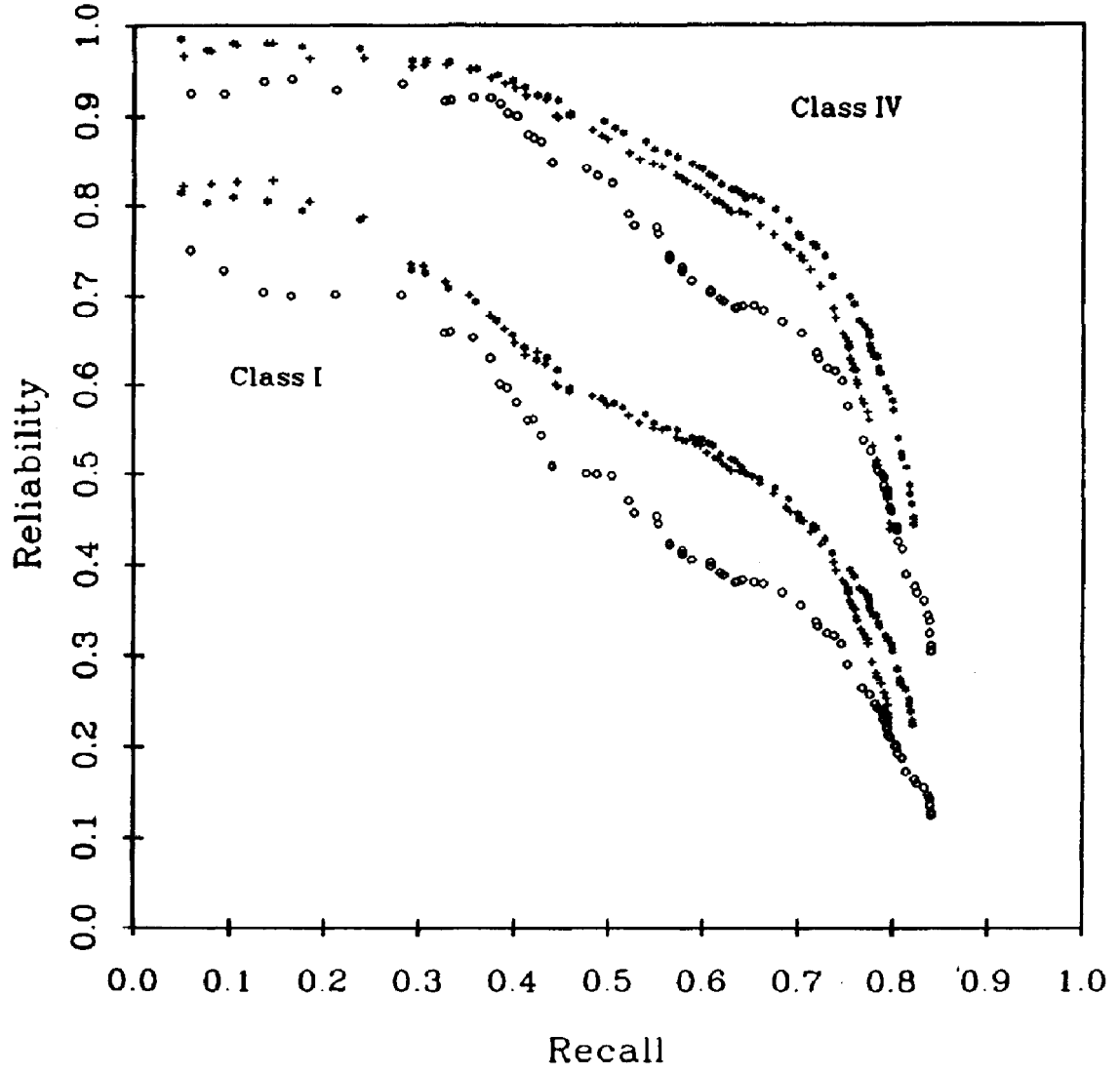


**Figure 2.** PBM reliability versus recall of: + + + +, *1989 Registry* file; oooo, 1989 file limited to only one spectrum of each compound; and ****, 1982 file.

same average data distribution, that is, fall in the same areas of the multidimensional mathematical hyperspace representing the reference data [8]. To test this, the *FP* values for the 61,989 possible false answers ($P_f$) of the 1982 data base and for the 52,429 additional false answers in the 1989 data base were determined from the PBM [22] retrievals of the 385 randomly selected unknowns [26]. By using matching criteria that recall 80% of the correct (class I) answers, 1,210 incorrect answers were retrieved from the 1982 spectra (*FP* = 2.0%) but, surprisingly, only 352 from those added in 1989 (*FP* = 0.67%). It follows that two thirds of the new spectra are substantially different from the original spectra, occupying to a corresponding extent new areas of the multidimensional data hyperspace representing mass spectra.

The expected deleterious effect of the larger data base on the reliability as a function of *RC* (Figure 2) is surprisingly small. However, for this to be a lower *FP* value for the new spectra, from eq 2 the change in the $P_c$ and *RC* values should be insignificant, yet $P_c$ actually increased from 1,180 to 1,450 possible correct answers with the increased file size. To minimize this effect, a separate recall-reliability plot was calculated limiting the data base to only one spectrum of each of the unknown compounds; for the 385 unknowns, the $P_c$ values for the 1982 and 1989 data bases are 478 and 506 (class I criteria, including stereoisomers) as matches. The resulting plots (Figure 3) still show a surprisingly small adverse effect of the 76% size increase on class I matching reliability at any *RC* value; the larger difference for class IV criteria reflects, as

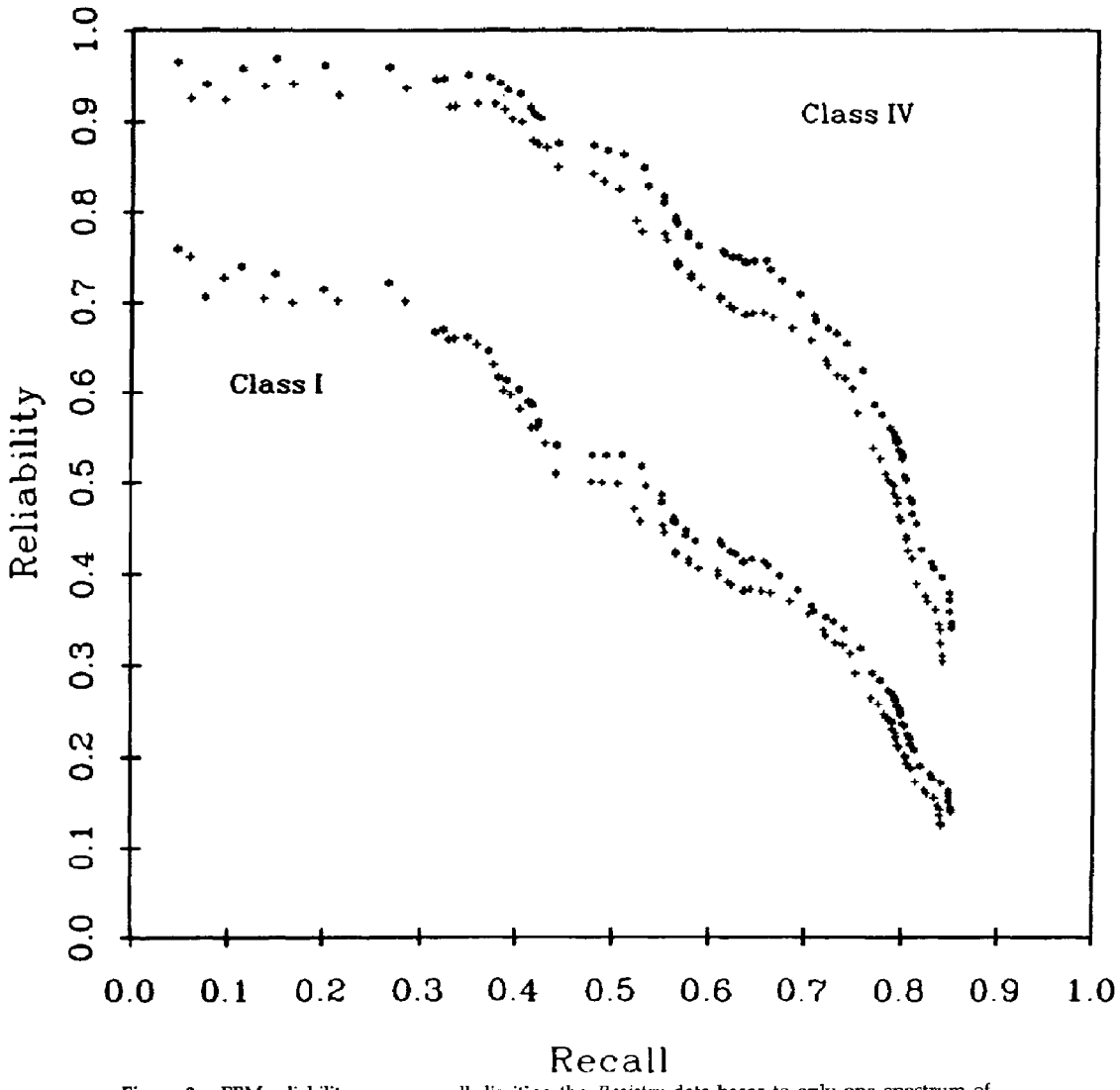


**Figure 3.** PBM reliability versus recall, limiting the *Registry* data bases to only one spectrum of each unknown of: ****, 1982 file; + + + +, 1989 file.

expected, the larger areas of the data hyperspace allowed in the class IV matching requirements. This expansion in data distribution is also shown by the 64% increase in the number of different compound elemental compositions.

The further reduction of these deleterious effects of the larger data base shown for the recall-reliability plot of all spectra (Figure 2) thus indicates an increased *RC* value for the 270 new possible correct answers (eq 2). For all but the lowest reliability values, the class I performance is at least equivalent to that of the smaller data base, and the class IV performance has been degraded to a relatively small extent.

## Conclusions

This 76% expansion of the data base has mainly involved new areas of mass spectral information, thus improving the capability for identifying a broader range of unknown compounds. This also justifies further efforts to increase the size of the data base.

## Acknowledgments

Literally hundreds of mass spectrometrists made important contributions of spectra to expand the data base. Special acknowledgments are due D. Henneberg of the Max-Planck Institute; P. Z. Chang and co-workers of the Chinese Academy of Medical Sciences; D. Sparkman of Nermag; W. Shackelford, J. M. McGuire, and W. L. Budde of EPA; T. Wachs and J. D. Henion of Cornell; L. E. Abbey and D. E. Bostwick of Georgia Tech; F. Turecek and V. Hanus of the Heyrovsky Institute, Prague; M. Buchanan, Oak Ridge; and C. Shackelton and A. L. Burlingame of the University of California, San Francisco. A more comprehensive list of acknowledgments is given in ref 20.

For data evaluation we are grateful to I. J. Amster, D. E. Drinkwater, R. Feng, J. J. P. Furlong, C.-J. Guo, J. A. Loo, M. A. Sharaf, W. Staedeli, B. H. Wang, E. R. Williams, M.-Y. Zhang, and, especially, C. Wesdemiotis. Research on the Probability Based Matching algorithm was supported by the National Science Foundation under grants CHE-8303340 and CHE-8620293, and the data collection by John Wiley and Sons, Inc., Electronic Data Division (605 Third Avenue, New York, NY, 10158).

## References

1. Abrahamsson, S.; Stenhagen, E.; McLafferty, F. W. *Atlas of Mass Spectral Data*; John Wiley: New York, 1969.
2. Abrahamsson, S. *Sci. Tools* **1967**, *14*, 129.
3. McLafferty, F. W.; Gohlke, R. S. *Anal. Chem.* **1959**, *31*, 1160.
4. Venkataraghavan, R.; McLafferty, F. W.; Van Lear, G. E. *Org. Mass Spectrom.* **1969**, *2*, 1–15.
5. Hertz, H. S.; Hites, R. A.; Biemann, K. *Anal. Chem.* **1971**, *43*, 681.
6. Smith, D. H.; Buchanan, B. G.; Engelmore, R. S.; Duffield, A. M.; Yeo, A.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. *J. Am. Chem. Soc.* **1972**, *94*, 5962.
7. Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. *J. Am. Chem. Soc.* **1973**, *95*, 4185.
8. Justice, J. B.; Isenhour, T. L. *Anal. Chem.* **1974**, *46*, 223.
9. Pesyna, G. M.; Venkataraghavan, R.; Dayringer, H. G.; McLafferty, F. W. *Anal. Chem.* **1976**, *48*, 1362.
10. Dokomos, L; Henneberg, D.; Wiemann, B. *Anal. Chim. Acta* **1983**, *150*, 37.
11. Cleij, P.; van't Klooster, H. A.; van Houwelingen, J. C. *Anal. Chim. Acta* **1983**, *150*, 23.
12. Clerc, J. T.; Szekeley, G. *Trends Anal. Chem.* **1983**, 2, 50.
13. Shackelford, W. M.; Kline, D. M.; Faas, L.; Kurth, G. *Anal. Chim. Acta* **1983**, *146*, 15.
14. McLafferty, F. W.; Stauffer, D. B. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 251.
15. McLafferty, F. W.; Loh, S. Y.; Stauffer, D. B. In: *Computer Enhanced Analytical Spectroscopy Vol. II*; Meuzelaar, H. C., Ed.; Plenum: New York, 1990; pp 163–181.
16. McLafferty, F. W.; Stauffer, D. B. *Int. J. Mass Spectrom. Ion Processes* **1984**, *58*, 139–149.
17. Office of Standard Reference Data, National Institute of Science and Technology, Gaithersburg, Maryland.
18. Lias, S. G. *J. Res. Natl. Inst. Stds. Techn.* **1989**, *94*, 25–35.
19. *Eight Peak Index of Mass Spectra*, Royal Society of Chemistry: Nottingham, UK, 1983.
20. McLafferty, F. W.; Stauffer, D. B. *Wiley/NBS Registry of Mass Spectral Data*; Wiley-Interscience: New York, 1989.
21. Speck, D. D.; Venkataraghavan, R.; McLafferty, F. W. *Org. Mass Spectrom.* **1978**, *13*, 209–213.
22. Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. *Anal. Chem.* **1985**, *57*, 1056–1060.
23. Huffman, D. A. *Proc. IRE.* **1952**, *40*, 1098–1106.
24. Terwilliger, D. T.; Behbenani, A. L.; Ireland, J. C.; Budde, W. L. *Biomed. Environ. Mass Spectrom.* **1987**, *14*, 263–270.
25. Pesyna, G. M.; McLafferty, F. W.; Venkataraghavan, R.; Dayringer, H. G. *Anal. Chem.* **1975**, *47*, 1161–1164.
26. Loh, S.; McLafferty, F. W. *Anal. Chem.* **1991**, *63*, 546–550.
27. Salton, G. *Introduction to Modern Information Retrieval*; McGraw-Hill: New York, 1983.
28. McLafferty, F. W. *Anal. Chem.* **1977**, *49*, 1441–1443.
29. Mun, I. K.; Bartholomew, D. R.; Stauffer, D. B.; McLafferty, F. W. *Anal. Chem.* **1981**, *53*, 1938.