Comparative Evaluations of Mass Spectral Data Bases

Fred W. McLafferty
Department of Chemistry, Baker Laboratory, Cornell University, Ithaca, New York, USA

Douglas B. Stauffer and Stanton Y. Loh
Palisade Corporation, Newfield, New York, USA

Recent reports from the National Institute of Science and Technology (NIST) state that its large (53,994) collection of mass spectra is unique in "consisting almost entirely of complete spectra." Our study of the 1989 *Registry of Mass Spectral Data* of 139,859 different spectra shows that its 53,994 spectra containing the most peaks average 108 peaks per spectrum, 48% larger than the NIST data base. Further, in matching unknown spectra of compounds present in both files, by using criteria yielding 68% reliability, 14% of the possible correct answers were recalled with the NIST data base versus 36% with the *Registry*. *(J Am Soc Mass Spectrom 1991, 2, 438-440)*

In a recent report [1] we described the latest expansion of the *Registry of Mass Spectral Data* [2, 3] that now contains 139,859 different spectra of 118,144 different compounds. This contains all spectra of the 1989 collection of the National Institute of Standards and Technology (NIST), which uniquely contributed 17,537 different spectra of 14,271 different compounds. Because of this generous cooperation of NIST, the 1989 *Registry* contains almost all mass spectra available in any public collection.

This report [1] also evaluated the 1989 *Registry* as a reference collection for identifying unknown mass spectra. Of particular interest here is a recent NIST publication [4] raising the completeness of spectra as a key issue in data base utility; this stated that spectra uniquely in the 1982 *Registry* have a mean size of 13 peaks per spectrum versus 60 peaks per spectrum in the NIST collection. For the 1990 NIST data base, released after the original submission of our study [1], publicity [5] claimed that it "is the only large collection of mass spectra consisting almost entirely of complete spectra." For a more careful evaluation of this issue, statistical data for the 1990 NIST data base, which contains 53,994 mass spectra of different compounds, are compared with that reported for the 1989 *Registry* [1].

Completeness of spectra. The 1989 *Registry* averages 53 peaks per spectrum; the larger proportion of new spectra taken from the literature is the main reason for the reduction from the value of 67 in the 1982 *Registry*. Both values are less than that of 73 in the 1990 NIST data base. However, the 53,994 spectra of the 1989 *Registry* containing the most peaks average 108 peaks per spectrum, 48% larger than the value for the NIST data for the same number of spectra. Further, the claim "consisting almost entirely of complete spectra" [5] is not applicable to either data base, as shown by the distribution curve of Figure 1. The number of peaks most commonly found in the *Registry* is 11, representing 4470 spectra, while that for NIST is 10, representing 995 spectra; both values are surprisingly close to the value eight of the *Eight Peak Index* [6] which contains 66,720 mass spectra. In Figure 1, with increasing values of peaks per spectrum above this maximum, the number of spectra for each value decreases quite regularly, but the values for the NIST data remain well below those of the 1989 *Registry*; the *Registry* has 49,103 spectra containing 50 or more peaks, versus 28,513 for the NIST collection.

Multiple spectra of the same compound. The NIST collection contains only one mass spectrum of each compound, while the *Registry* includes all separately measured spectra of the same compound. Although increasing the size of the data base must increase the probability that one of its spectra will incorrectly match that of the unknown, including multiple reference spectra of a compound increases the probability that the experimental conditions used in measuring one of these will be similar to those used in measuring an unknown mass spectrum of that same compound. A

Address correspondence to Fred W. McLafferty, Baker Chemistry Laboratory, Cornell University, Ithaca, NY 14853-1301.

1044-0305/91/$3.50

Received February 11, 1991
Accepted March 12, 1991

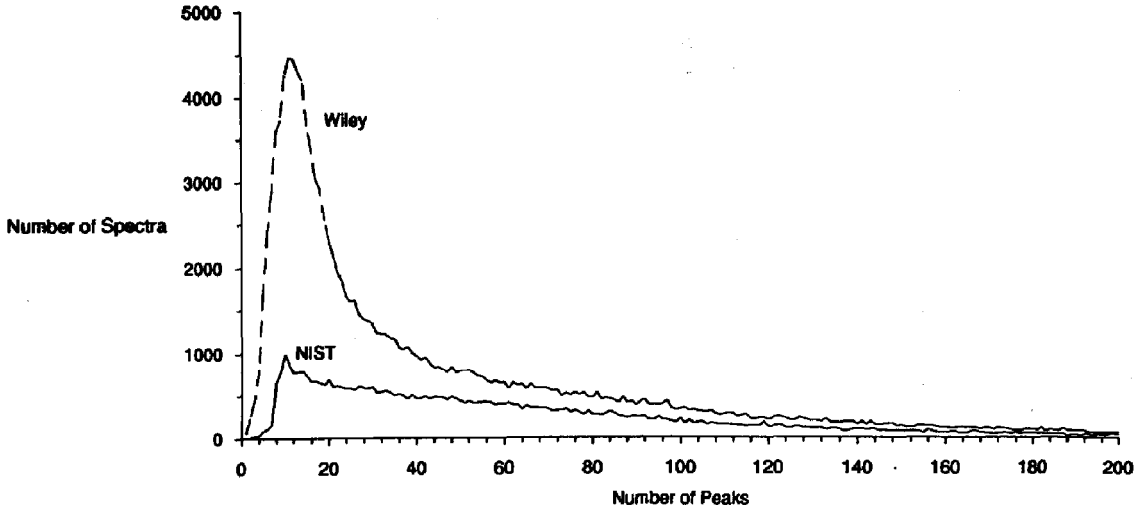

Figure 1. Number of spectra in the 1990 NIST and 1989 Wiley *Registry* data bases as a function of the number of peaks tabulated for each spectrum.

retrieval comparison using Probability Based Matching (PBM) [7–9] was made by using 371 unknown spectra [1], selected at random, for which other spectra of the same compound are contained in both the NIST and *Registry* data bases. The best matching spectrum was correct (same compound or a stereoisomer) for 36 more (15%) unknowns by using the *Registry* versus NIST data base, despite the fact that the latter contains only 39% as many spectra. The PBM matching performance for these unknowns is also compared in the recall/reliability plot of Figure 2. Inspection of individual retrievals shows that preva-

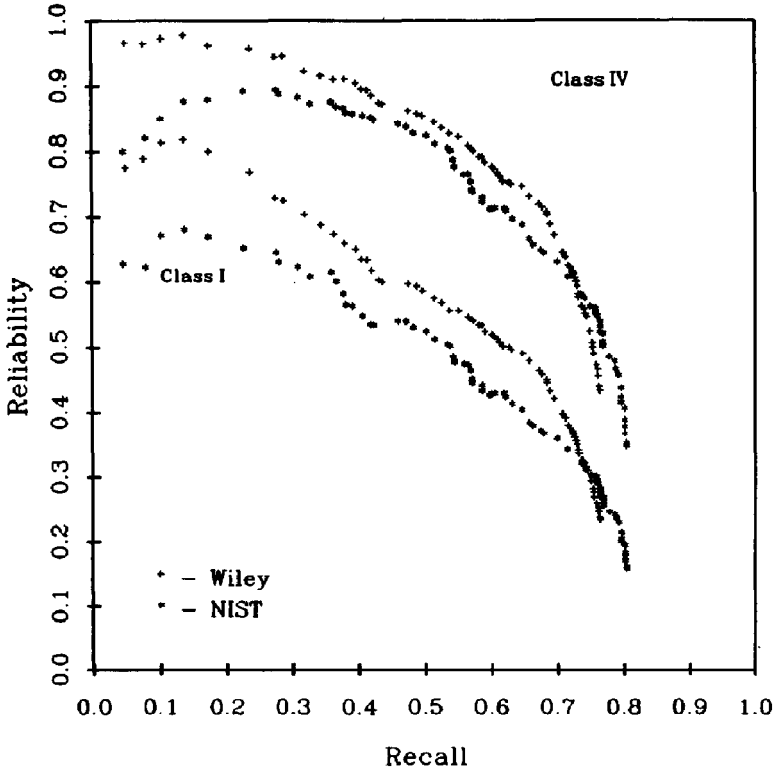

Figure 2. PBM reliability versus recall, 1990 NIST and 1989 Wiley *Registry* data bases. Class I matches are the same compound or a stereoisomer, while Class IV matches only vary structurally by differences for which mass spectrometry is insensitive [7–9].

lence of errors is also a factor, particularly in the high reliability area; under matching criteria that yield 68% reliability, 14% of the possible correct answers are recalled by using the 1990 NIST collection versus 36% by using the 1989 *Registry*.

Search speed. With modern computer capabilities, data base size is no longer an appreciable limitation on search speed. By using a 33 MHz 80486-based personal computer (Palisade Corporation, Newfield, NY) [9], PBM search times for the 1989 *Registry* average less than 3 sec.

Acknowledgments

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.

References

1. McLafferty, F. W.; Stauffer, D. B.; Twiss-Brooks, A. B.; Loh, S. Y. *J. Am. Soc. Mass Spectrom.* **1991**, 2, 432–437.
2. McLafferty, F. W.; Stauffer, D. B. *Int. J. Mass Spectrom. Ion Processes* **1984**, *58*, 139.
3. McLafferty, F. W.; Stauffer, D. B. *Wiley/NBS Registry of Mass Spectral Data*; Wiley-Interscience: New York, 1989.
4. Lias, S. G. *J. Res. Natl. Inst. Stds. Techn.* **1989**, *94*, 25–35.
5. *Chem. Eng. News* **1990**, *July* 2, 20.
6. *Eight Peak Index of Mass Spectra*; Royal Society of Chemistry: Nottingham, UK, 1983.
7. Pesyna, G. M.; Venkataraghavan, R.; Dayringer, H. G.; McLafferty, F. W. *Anal. Chem.* **1976**, *48*, 1362.
8. Stauffer, D. B.; McLafferty, F. W.; Ellis, R. D.; Peterson, D. W. *Anal. Chem.* **1985**, *57*, 1056.
9. 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.