Department of Computing Science Working Paper Series

Faculty of Engineering and Information Sciences

1978

A mass spectrum dictionary for fast library searching

R. Geoff Dromey
University of Wollongong

A MASS SPECTRUM DICTIONARY FOR FAST LIBRARY SEARCHING

R. Geoff Dromey

Department of Computing Science, University of Wollongong,
P.O. Box 1144, Wollongong, N.S.W. 2500, Australia.

Brief:

A technique for precisely ordering a mass spectral library that requires on average less than 1% of the spectra to be examined in searchin for an unknown is presented. The method is economical on storage and it accommodates spectrum variability that would be expected in a search library environment.

Abstract:

When data can be arranged according to some ordering principle (e.g. numerical or alphabetical) powerful searching techniques can be applied to retrieve information. An explicit set of procedures is proposed for constructing a precisely ordered mass spectrum dictionary. Performance tests on the proposed system show that on average less than 1% of the spectra need to be examined in searching for a given unknown. The mass spectrum dictionary is economical on storage and it will accommodate spectrum variability likely to be found in a library search environment.

INTRODUCTION

Searching a large library of mass spectra for the best-match candidates with some particular query spectrum can be a time consuming process. A number of systems that perform well on this task have been proposed[1-7]. One of the most successful of the library search techniques is the method due to Heller[4] which employs an inverted file structure. One of the only real criticisms of this system is that it is very costly both in terms of file generation and file maintenance. The other really serious problem is the size of the file (although there are ways to alleviate this problem[8]).

These problems have prompted the development of a system that is comparable in performance but which requires neither the overhead of file maintenance nor large amounts of storage for the search file.

Searching large amounts of data can be efficient provided the items can be ordered in some way. For instance searching a file as large as a telephone directory by computer can be made a simple and efficient task using a binary search algorithm[9]. The binary search algorithm guarantees to find any item in an ordered file of N items in at most $\log_2 N$ steps. For example only 16 comparisons are necessary to locate any item in an ordered file of 32000 items. This efficiency is derived by taking advantage of the order in the file.

Returning to the mass spectrum context we find that searching a library of mass spectra for the best match with some unknown spectrum is comparable to searching a telephone directory in which the names are i random order. Obviously such a search system is very inefficient because there is no way of taking advantage of any inherent order in the data. Series displacement indices[7] and ion series analysis[5,6] have been suggested as ways of reducing the portion of the file that needs to be searched. These methods do not however possess the specificity that is desirable in searching large files; that is they do not eliminate enough candidates from the search.

This leads to the question as to whether we can impose some ordering principle on a library of mass spectra. Any attempt at such a task must take into account the inherent variability in the data. In particular,there can be considerable differences in intensity information for representations of a mass spectrum measured on different instruments.

A careful examination of a large number of mass spectra reveals that, in general, there are considerable differences in both mass and intensity representations for compounds that are closely similar structually. When intensity information is discarded altogether we still have a very selective representation. It might therefore seem reasonable that the mass information could be used in some way to order a file of spectra. However, using the mass information alone still does not completely remove the intensity variation problem. It will be shown in the next section that certain intensity related constraints can be placed on the selection of masses for the ordered representation.

MASS SPECTRUM DICTIONARY CONSTRUCTION

In order to construct a mass spectrum dictionary a fixed range for selecting masses must be predetermined (for this to work the mass range 40 → 99 is used). This range is divided up into 6 ranges of 10 mass units (i.e. 40 → 49, 50 → 59,...). The selection rules that are then applied to cope with spectrum variability are as follows:

[1] Select the mass of the most intense peak in each range.

[2] Make the following checks to determine if there is any possibility of ambiguity in the selection made

(a) if the most intense peak in a range is less than or equal to 1% assign it to the zero mass for the range and mark it as not ambiguous.

(b) if the most intense peak in a range is greater than 1% and less than or equal to 10% mark it as ambiguous.

(c) if the ratio of the most intense to the second most intense peak in a given 10 amu range is less than 1.4 flag it as ambiguous.

(d) all other peaks that are the most intense in their respective ranges are considered as unambiguous.

The preceding steps divide the ranges and selected peaks into two categories: (1) those that are definitive and unambiguous, and (2) those for which instrumental variations may lead to either no mass or another mass being selected as the most intense for that particular range.

With this information in hand it is possible to consider how representations that can be ordered may be constructed.

CONSTRUCTION OF ABBREVIATED MASS SPECTRAL REPRESENTATIONS SUITABLE FOR ORDERING

The representation for each unambiguous range is just the value of the integer mass for that range modulo 10 (e.g. mass 43 has the representation 3, and 57 has the representation 7).

The representations for those ranges that have been flagged as possibly ambiguous must be treated differently. They are given two possible representations (1) the representation they would assume if they were not classified as ambiguous (e.g. if the most intense peak in the range 40 → 49 amu is mass 43 at 5% then 3 is its representation) (2) the second representation for the range is a standard default setting of Ø. *This representation is vital for linking together spectra of the same compound that may be distorted with respect to one another due to instrumental variations.* The process must be taken a step further before we can construct suitable entries for an ordered mass spectrum dictionary.

Obviously we cannot expect to have single dictionary entries for any spectra that have one or more of their ranges flagged as ambiguous. Also we cannot expect that another representation of the same spectrum will have the identical ranges flagged as ambiguous as in the dictionary entry (e.g. the dictionary entry may have a peak of 9% [ambiguous] whereas another representation may have the same mass as 15% [unambiguous] and vice versa. *For this reason the dictionary must contain all possible permutations of the ambiguous and unambiguous representations.* The final representation for a given spectrum is therefore a set of six figure numbers. *It follows that if n of the ranges have been flagged as possibly ambiguous or uncertain then 2^n representations for the spectrum will need to be placed in the dictionary.* These 2^n representations should be sufficient to cover all instrumental variability apart from impurity peaks.

The mass range 40 → 99 was chosen for the representation on the statistical grounds that most spectra tend to show considerable fragmentation in this region. Obviously other ranges may be more appropriate in specialized circumstances.

The choice of a six figure representation was made because it was the smallest representation that gave considerable discriminating power. Longer representations could also be used if better resolution was required. However, they introduce the risk of many more dictionary entries per spectrum because of the way uncertain peaks are handled.

The six figure representations can be sorted into numeric order along with their accompanying pointers to the corresponding original spectrum.

As an example consider the dictionary entries for the spectrum of 5-Hydroxyoctanoic acid lactone shown in figure 1[10].

Applying the selection rules previously described we obtain the representation $25\underline{914}9$ where the numbers underlined are possibly uncertain. The peaks at 69 (9) and 84 (4) are ambiguous because they are less than 10%. The peak at 71 is classified as uncertain because its ratio with the peak at mass 70 (the second most intense peak) is less than 1.4. There are therefore three of the six mass regions that contain representations which cannot be guaranteed to withstand variations due to instrument distortion. Obviously the peaks at mass 42, 55 and 99 should withstand any distortion. In this particular case we must therefore generate eight $(2^{\text{no. of uncertain}} \equiv 2^3)$ dictionary entries for this spectrum to accommodate all possible representations. The representations are given in Table I. These representations are entered into the dictionary at the appropriate place correspondinq to ascending numeric order. Each entry is tagged with an address or label corresponding to the location of the complete stectrum in the master file.

PROCEDURE FOR GENERATING COMPLETE SETS OF DICTIONARY ENTRIES

The procedure for generating the permutations for any given spectrum can be formulated as a simple algorithmic sequence which has the following steps

(1) Count the number of uncertain digits N in the basic representation generated using the prescribed selection rules.

(2) Calculate $2^N - 1$ and assign it to the variable X. The value of X is one less than the total number of dictionary entries to be generated for the spectrum.

(3) Calculate the binary representation of X (e.g. $7_{10} \rightarrow 111_2$).

(4) Moving left to right in both the binary representaion of X and the decimal representation of the spectrum generated by the selection rules multiply the uncertain decimal digits by the corresponding binary representation (see Table I).

TABLE I: Mass Spectrum Dictionary Entries for 5-Hydroxyoctanoic acid lactone generated by binary multiplication.

X = 7	$259\underline{149}$ 111	259149
X = 6	$259\underline{149}$ 110	259109
X = 5	$259\underline{149}$ 101	259049
X = 4	$259\underline{149}$ 100	259004
X = 3	$259\underline{149}$ 011	250149
X = 2	$259\underline{149}$ 010	250109
X = 1	$259\underline{149}$ 001	250049
X = 0	$259\underline{149}$ 000	250009

The uncertain digits are underlined

(5) Subtract one from X (e.g. X = X - 1) and if the new value of X is greater than or equal to Ø go back to step (3) and calculate another permuted dictionary entry. If X is less than zero all the necessary representations have been generated and so the process can halt.

CONSTRUCTION OF AN EFFICIENT SEARCH FILE

There are two basic options open for constructing an effective retrieval system based on the abbreviated spectrum representations that have been described. The first and perhaps most obvious approach is to sort and place the representations into ascending numeric order. The other alternative is to use what is known as a hashing method to determine where in the file a particular dictionary entry should be placed. At the expense of some additional storage the latter approach can be guaranteed to be more efficient.

(a) Spectrum Retrieval from a Numerically Ordered File

The most practical and efficient way to search a numerically ordered file is to apply what is known as a binary search algorithm[9]. This search procedure uses information about the order of the file to remove from further consideration half of the remaining entries in the file with each comparison made. The basic procedure is to take the number sought and compare it with the number in the middle of the file. If it is less than the middle value we can discard from further consideration all the values in the top half of the file and vice versa if it is greater than the middle value. We then repeat the process with the remaining half of the file by dividing it in turn in two. With just two comparisons three quarters of the file entries will be eliminated. The bisection procedure continues until we find the number we have sought or until we have established that it is not in the file. Analysis of the binary search procedure tells us that we never need to make more than $\log_2 N+1$ comparisons and that on average only $\log_2 N-1$ will be needed[9]. If N is say 60,000 which is probably typical for a large library of mass spectra then on average only 16 entries in the file will need to be examined to find the address of any particular representation.

(b) Spectrum Retrieval from a Hash-Stored File

It is common practice in computing science to use a hashing technique to retrieve data from a large file. It has also been used in chemistry for molecular formula retrieval[4]. Hashing is usually favoured because of its high efficiency for retrieval[9,11]. The basic idea of hashing is to take the numeric representation of the data to be stored and transform it in some way to produce a representation-dependent location at which the data is to be stored. Such hashing transformations are usually not unique. It is however usually easy to find a suitable transformation that will map the original data into a retrieval file that contains about twice as many locations as representations to be stored. Theoretical analysis of such systems indicates that on *average less than two positions in the table will need to be examined to locate any particular representation.* Detailed procedures for implementing hash files can be found in Knuth[9] and Severance[11]. A linked overflow storage area[11] is the most suitable method for handling mass spectral data because of the fact most representations occur more than once in the file.

To locate an item in a hashed file the representation that is sought is hashed to produce a location at which to begin the search. The value stored in the calculated location is then compared with the number sought. If they match the search terminates otherwise adjacent locations are examined until the desired match if found. If in this subsequent search an empty location is encountered this signifies that the representation sought is not present in the file.

(c) The Search and Matching Procedures

The following steps must be carried out to locate the set of spectra most like some unknown spectrum being sought.

(1) The selection rules are applied to the unknown spectrum, establishing its numeric representation and its uncertain peaks.

(2) The permuted set of dictionary entries are then generated.

(3) The addresses of all the dictionary entries are then located by hashing or binary search depending on the file representation chosen. These addresses are pointers to where the complete spectra are stored.

(4) The selected set of complete spectra are then retrieved and compared against the unknown spectrum using some standard matching procedure.

The merging process is easily implemented with a bit vector which has as many bits as there are spectra in the file. As each spectrum number is found the corresponding bit is set in the bit vector.

RESULTS AND DISCUSSION

To test out the feasibility of the procedures that have been outlined a set of 10000 mass spectra were studied and analyzed. The first parameter of interest for files of this type is the average number of representations per spectrum that are needed to cover all possibilities introduced by peak uncertainties. It was found that 60894 representations were generated for the 10000 spectra. That is on average just over 6 dictionary entries per spectrum are needed to construct the dictionary. The size of the file processed to obtain this result should be large enough to guarantee its statistical significance for any general file of mass spectra.

The other important parameter for gauging the effectiveness of this method of spectrum retrieval is the average number of complete spectra that are required to be matched for a given query spectrum. To estimate this performance parameter 100 spectra from the file of 10000 were selected at random to remove any bias. Dictionary entries for each of these spectra were then generated and the corresponding number of spectra retrieved in each instance was calculated. The search results for the 100 spectra are summarized in Table II. Averaging the results for the 100 spectra searched for, it was found that 0.98% (∿1%) of the file had to be matched.

TABLE II: Number of sample spectra requiring less than X% of the file to be searched.

No. of spectra requiring less than X% of file to be searched	Percentage of file searched (X%)
26	0.1%
41	0.25%
50	0.5%
68	1.0%
82	2.0%
92	3.0%
100	over 3%

The results in Table II indicate that we expect a quarter of the searchs that are made to require searching of less than 0.1% of the file. The table also shows that on average half of the spectra will require less than 0.5% of the file to be searched and that a total of 92% of the spectra require less than 3% of the file to be searched.

One spectrum in the set took 5.88% and another spectrum required 7.92% of the file to be searched. The latter spectrum had 5 of its 6 peaks flagged as uncertain (e.g. 113791) with only the mass 77 being unambiguous. There were therefore 32 (i.e. 2^5) dictionary entries for this compound and so became of the considerable number of aromatic compounds present 7.92% of the file had to be searched.

There were nearly 1000 representations that had all zeros as their representation (e.g. 000000). The high number of these representations is due to a number of factors. A number of spectra in the data set do not have mass measurements taken until after mass 600. Ideally these anomolous data should be removed. Other spectra (mostly aromatics) have only a very few low intensity peaks below mass 100 while some have low molecular weights with only small peaks in the mass range 40 to 99.

The latter groups obviously have 000000 as a legitimate representation. The most practical approach to this problem is to use another range (e.g. 100 → 159). The 000000 representation then disappears from many of the dictionary entries for the spectra in this group. An effective, although not strictly correct approach to the problem, is to ignore the oooooo representation when performing retrievals.

Where library spectra have not been recorded over the prescribed range (e.g. down to mass 40) obviously we cannot expect their representation to be accurately described in the file. This problem is an inadequacy in the data rather than in the search strategy and as such it is the responsibility of those in charge of maintaining the integrity of the data base.

In discussing the performance of the system the preliminary searches of the dictionary have been ignored because of their very small overhead relative to spectrum matching operations. Further the number of comparisons for retrieval from both a hash file and ordered file are known and have been completely characterized by theoretical arguments[9].

CONCLUSIONS

A mass spectrum retrieval system based on a mass dictionary concept has been shown to be viable and competitive with existing systems. *It provides, at very small cost, a mechanism for filtering out on average better than 99% of the spectra in any given search.* The selection rule formalism used accommodates spectrum variability that is likely to be found in a search library environment. The storage cost of the filter dictionary averages out to only 6 entries per spectrum. This is small in relation to the cost of storing a complete fully inverted file. Even better performance from the system could be obtained by using a series of ranges. For example the selection rules could be used to derive representations for spectra that have peaks in say the range 100 → 160.

Alternatively, since the storage cost is small, representations for two ranges, e.g. 40 → 99 and 100 → 160 could be used if it was desirable to reduce the number of spectra to be matched to considerably less than 1%.

ACKNOWLEDGEMENTS

The author wishes to thank Ann Titus for typing the manuscript and Drs. S.R. Heller and G.W. Milne for providing the mass spectral data base.

LITERATURE CITED

1. F.W. McLafferty, R.H. Hertel and R.D. Villnock, *Org.Mass.Spec.*, 9 690 (1974).

2. S. Grotch, *Anal.Chem.*, 43, 1362 (1971).

3. H.S. Hertz, R.A. Hites, and K. Bienann, *Anal.Chem.*, 43, 681 (1971).

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

5. L.R. Crawford and J.D. Morrison, *Anal.Chem.*, 40, 1469 (1968).

6. D.H. Smith, *Anal.Chem.*, 44, 536 (1972).

7. R.G. Dromey, *Anal.Chem.*, 48, 1464 (1976).

8. R.G. Dromey, *J.Chem.Inf.Comp.Sci.*, (August 1978).

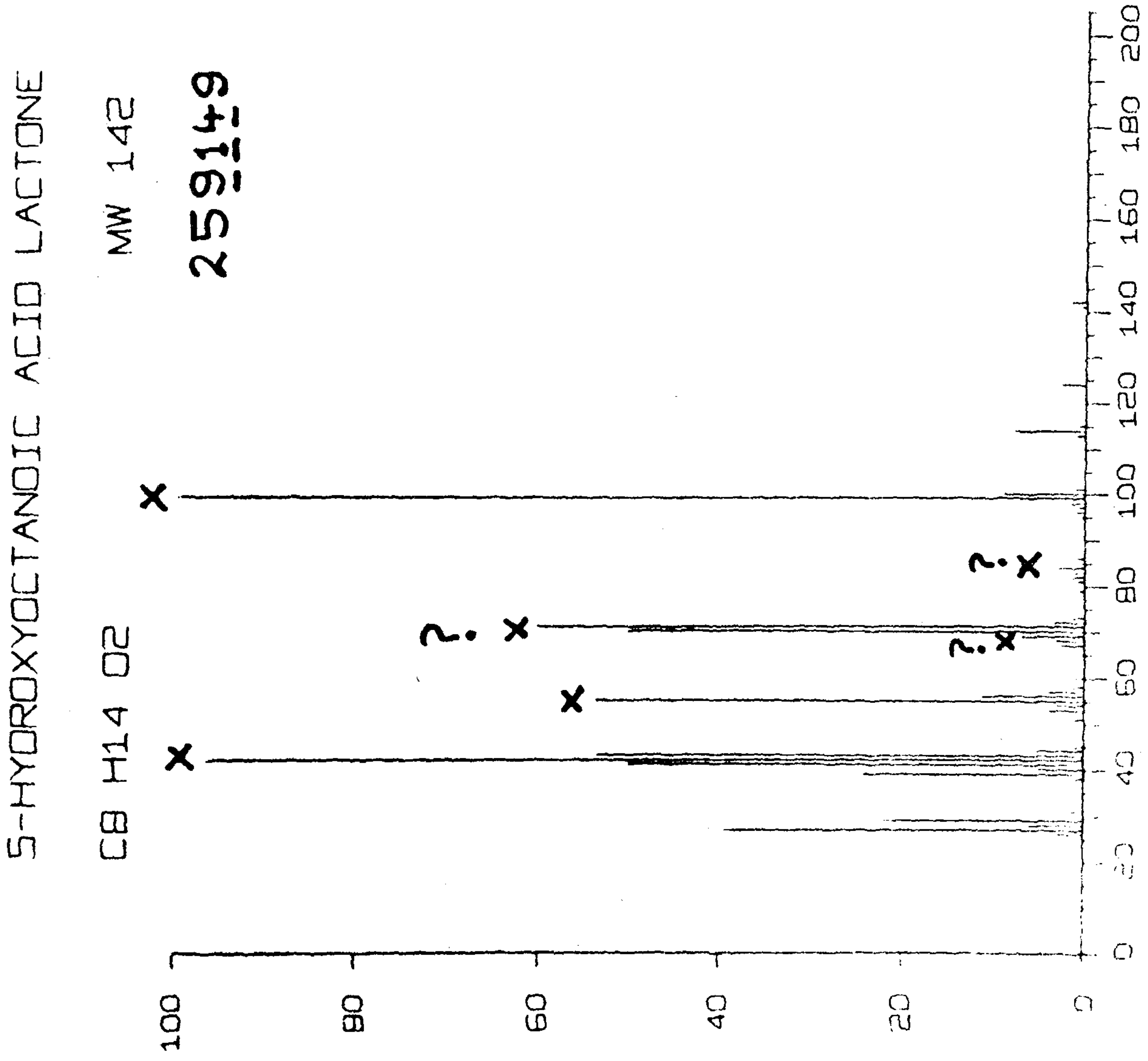
9. D. Knuth, "The Art of Computer Programming, Vol. 3, Addison-Wesley, Reading (1973).

10. B.H. Kennett, K.E. Murray, F.B. Whitfield, G. Stanley, J. Slinpton, P.A. Bannister, "Mass Spectra of Organic Compounds", CSIRO report, (1977).

11. D. Severance and R. Duane, *Comm.ACM*, 19, 409 (1976).

FIGURE CAPTION

Figure 1: Numeric representation for 5-Hydroxyoctanoic acid lactone.


5-HYDROXYOCTANOIC ACID LACTONE
C8 H14 O2
MW 142
X
?. X
?. X
X
?. X
X