1979

Optimum scaling of mass spectra for computer matching

R. Geoff Dromey
University of Wollongong

OPTIMUM SCALING OF MASS SPECTRA FOR COMPUTER MATCHING

R. Geoff Dromey

Department of Computing Science, The University of Wollongong,

P.O. Box 1144, Wollongong, N.S.W., 2500, Australia.

Summary

A scaling procedure that minimizes effects due to mass discrimination and other instrumental distortion in computer-matching of mass spectra is described. It is shown how spectra should only be matched when they have been scaled to be at their minimum "distance" with respect to the similarity index being used for the measurement.

Introduction

Computer-matching of mass spectra is a widely used technique for identifying unknown compounds [1-5]. Variability in instrumental performance can considerably complicate this task. For example, we may find that spectra of the same compound measured on mass spectrometers with different mass discrimination and focussing biases can show appreciable differences in their intensity profiles. This in turn can lead to uncertainties when attempting to identify unknown spectra. The problem is particularly relevant when a large library of spectra derived from a number of sources is being employed as the standard for computer-matching.

These circumstances might prompt us to ask the question: Is there any way in which the matching procedures that are currently employed can be modified to compensate for instrumental distortion and variability? As we will see there is a relatively straightforward way to minimize this problem and hence place us in a position where we can assume a greater confidence in the "degree-of-match" measures that we obtain. The method that is to be described relies upon an optimum scaling procedure that involves a least squares model and analysis.

The Matching Surface and Scaling

In preprocessing spectra for comparison one of two standard approaches is generally used. Probably the most common method is to normalize the spectra being compared so that the most intense ion in each case is set to a constant (usually 100 or 1000). The other approach is to compute the total ion sum for each spectrum and then adjust the peak intensities so that this sum is always a constant (usually 1 or 100). It might be argued that the second method is preferable to the first because it relies on all peaks to derive the normalization factor. At a glance this would seem better than relying on the accuracy of just one

peak, the most intense in the spectrum. Beyond this there is no clear cut argument to suggest that total-ion-sum normalization is the best way to approach the problem.

The approach that is suggested as a viable alternative is based on a somewhat sounder premise. It is conjectured that spectra should only be compared when they have been systematically scaled to be at their minimum distance apart as measured by the similarity index being employed (as an example the distance-apart or similarity index might be the sum of the squared differences of peak intensities of two spectra).

We find, if we vary the scaling of one spectrum with respect to another over a suitable range, that for some value of the scaling parameter(s), the two spectra are at their closest "distance" with respect to one another. That is, for all smaller and larger scaling parameters, the two spectra are further **apart.** Or, in other words, we could observe that the matching surface for the two spectra was "parabolic" with a well-defined minimum value. Figure (1) shows schematically the distance apart of two spectra as a function of the scaling parameter applied to one of the spectra. The matching surface is generally not symmetric about the minimum and so is not strictly parabolic.

The largest-peak normalization and total-ion-sum normalization distance will be arbitrarily located on the matching surface in relation to the minimum distance and optimum scaling. *This further implies that the amplitude of our distance index taken by either of these methods of scaling will be somewhat arbitrary and inconsistent because the matching surfaces for each pair of spectra are independent.* That is, the "parabola" for the two spectra A and C will be *different* to the parabola for A and B.

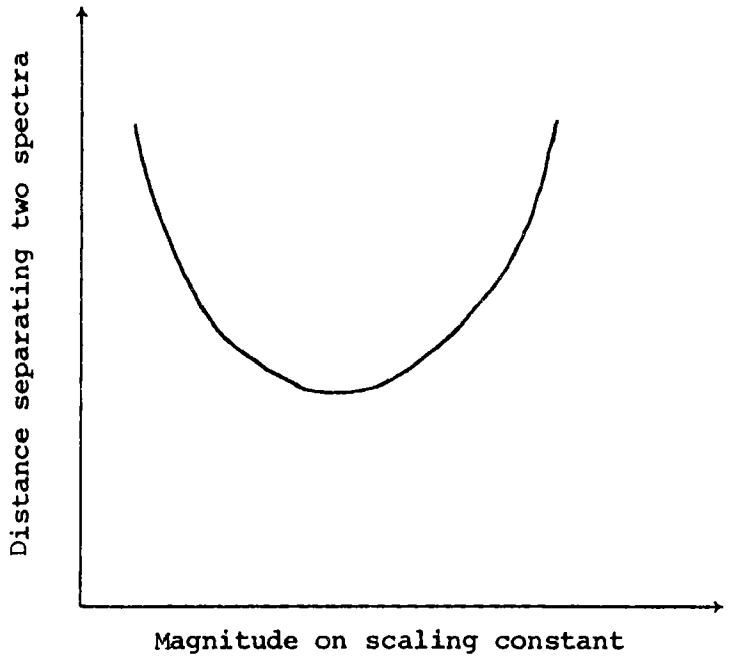


Figure 1

"Parabolic" matching surface showing distance between two spectra as a function of a scaling parameter.

The idea of only evaluating the distance index between two spectra when they have been scaled to be at their minimum distance seems to be the only way to avoid inconsistencies in similarity index measurements. The three matching surfaces depicted in figure 2 illustrate and emphasize the nature of the scaling problem. Clearly there is an *arbitrary* relativity for the distance indices H_A, H_B and H_C that have been scaled with respect to the unknown so that their most intense peaks are equal in scale value (that is, $A_{max} = B_{max} = C_{max}$ = constant (e.g. 100) and the scaling factor $C = 1$). In contrast the similarity indices O_A, O_B and O_C, taken at their respective optimum scalings, are relative because they are always calculated at their minimum distance.

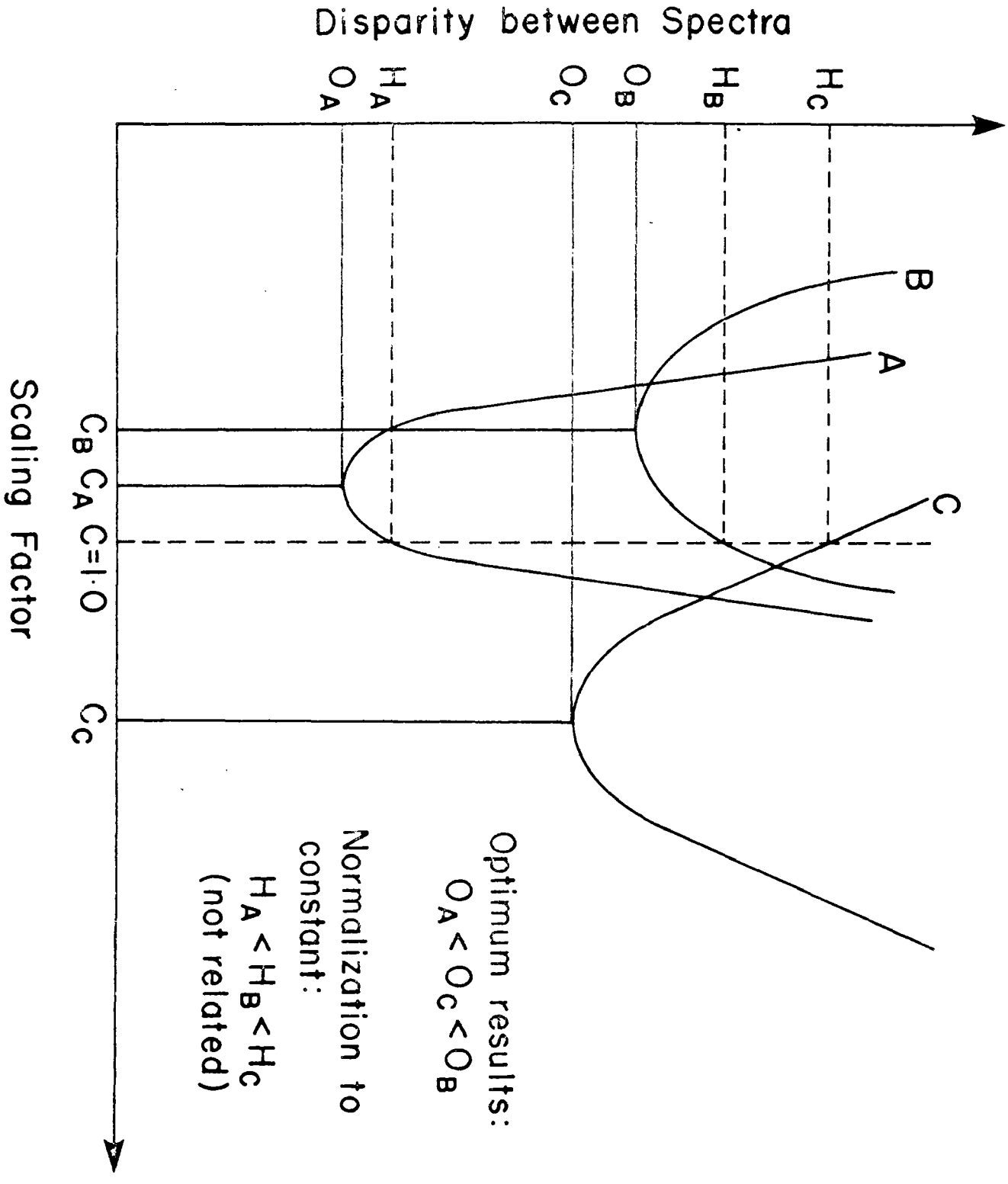
Derivation of Optimum Scaling Conditions for Matching Spectra

Scaling spectra to be at their minimum distance is usually a relatively straightforward task. It does however depend on the distance metric employed.

The simplest approach that we can use to optimum scaling is to find the constant factor by which all peaks in a spectrum must be multiplied to minimize the distance from some other spectrum. In this case it is easiest to use the mean square distance. In general the distance between the unknown spectrum U and the reference spectrum R is then given by

Figure 2

Three independent matching surfaces are shown which indicate the arbitrary relationship among the similarity indices H_A, H_B, and H_C evaluated at constant normalization (C = 1.0). The indices O_A, O_B and O_C taken at their optimum scaling are directly related.



$$D = \sum (U_m - R_m)^2 \qquad (1)$$

where the sum is taken over all peaks in both spectra.

If "c" is the scaling factor we have

$$D = \sum (U_m - c.R_m)^2 \qquad (2)$$

and "c" will be optimum in a least squares sense when $\frac{\partial D}{\partial c} = \emptyset$

and so we have

$$\frac{\partial D}{\partial c} = \sum 2.c.R_m^2 - \sum 2.U_m.R_m = \emptyset \qquad (3)$$

$$\therefore \qquad c = \frac{\sum U_m R_m}{\sum R_m^2} \qquad (4)$$

Because of the effects of mass discrimination and variations in sample concentration (i.e. in GCMS) it is almost always more desirable to work with a mass-dependent scaling factor. This enables us to compensate for trend differences in intensity that are a function of mass. The simplest way to introduce this compensation is to use scaling factors that are linearly dependent on mass. A model of this kind should be expected to give reasonable compensation against intensity distortions with either increasing or decreasing mass.

We can proceed with the derivation of optimum mass dependent scaling in a manner similar to the method used above. The distance now becomes

$$D_M = \sum \left[U_m - (c + m.d).R_m \right]^2 \qquad (5)$$

The distance will be minimum when $\frac{\partial D_m}{\partial c} = 0$ and $\frac{\partial D_m}{\partial d} = \emptyset$. Applying these two conditions we get

$$\sum R_m^2.c + \sum R_m^2.m.d = \sum U_m.R_m \qquad (6)$$

$$\sum R_m^2.m.c + \sum R_m^2.m^2.d = \sum U_m.R_m.m \qquad (7)$$

which yield the following optimum values for c and d.

$$c = \frac{\sum U_m . R_m - \sum R_m^2 . md}{\sum R_m^2} \qquad (8)$$

$$d = \frac{\sum U_m . R_m . \sum R_m^2 . m - \sum U_m . R_m . m . \sum R_m^2}{\sum R_m^2 . m \sum R_m^2 m - \sum R_m^2 . m^2 . \sum R_m^2} \qquad (9)$$

Now if we calculate the distance between spectrum U and R by substituting c and d into equation (5) we get the minimum linear mass-dependent distance between the two spectra.

Characteristics of Matching Profiles

Before evaluating and comparing the optimum scaling methods that we have derived it is appropriate to have a closer look into the characteristics of matching surfaces. There is a need for this slight detour in order to gain an insight into the relevance of optimum scaling. Two types of matching surface are relevant in this consideration, one where the two spectra being matched are clearly disparate, and the other where the two spectra are quite similar.

As we might expect the characteristics of the matching surfaces differ considerably in the two instances. For clearly disparate spectra figure (3a) the parabolas are shallow and broad about their minima, suggesting that scaling has only a small influence on the magnitude of the distance between spectra. On the other hand when two spectra are very similar the distance index is very sensitive to how they are scaled with respect to one another. This is borne out by the deep and narrow matching surfaces figure (3b) observed for closely similar pairs of spectra. As we are generally only interested in the integrity of the distance indices for spectra that are very similar to our unknown it is clear that only in these circumstances should we consider applying optimum scaling.

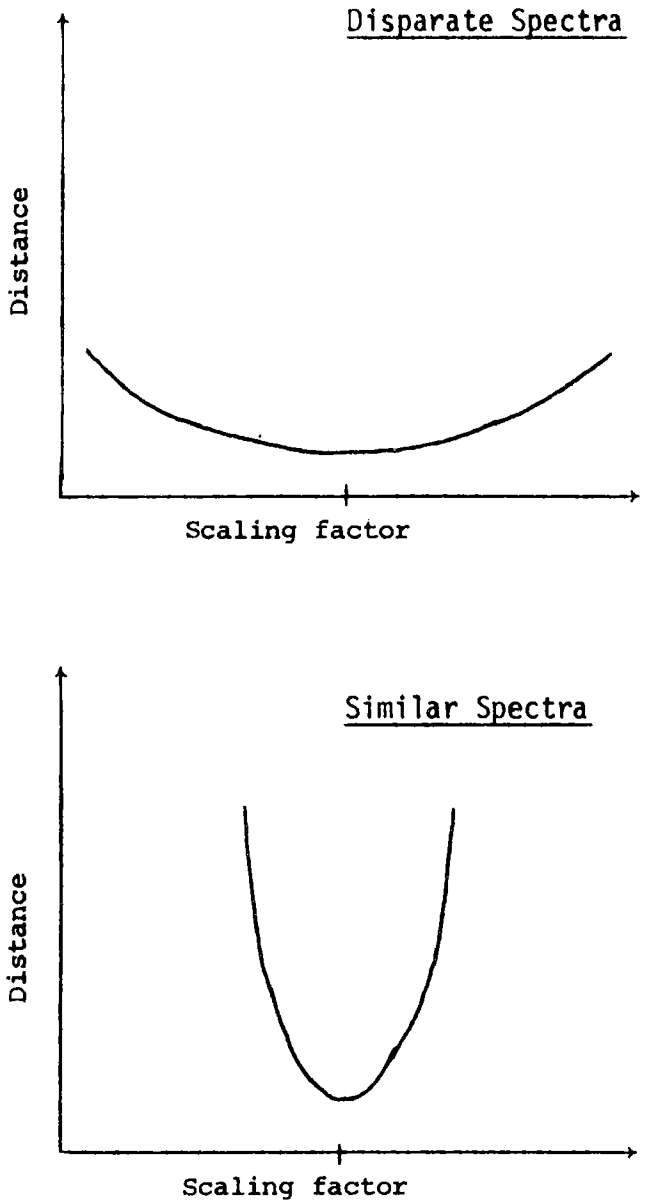


Figure 3

(a) Matching surface for two disparate spectra.

(b) Matching surface for two closely similar spectra.

That is optimum scaling should only be applied to rank the small subset of spectra that are most like our unknown. We will come back to these considerations later after examining the effects of optimum scaling with some concrete examples.

Comparison of an Unknown with Library Spectra

The two questions "which library spectrum is most like a given unknown?" and "which library spectrum is a given unknown most like?" are distinct and separate questions when considered in relation to optimum scaling and optimum matching. This may seem a subtle distinction but it is important to be aware of when we are considering spectra that are similar.

When we compare a set of library spectra with an unknown, and in doing so scale the reference spectra in relation to the unknown then the distance indices are meaningfully related in a relative sense (provided each comparison has been made at the minimum distance between the spectra). However in the complementary case when we scale the unknown to each of the library spectra in turn the situation is completely different. The distance indices for the matches of the unknown with the different library spectra are no longer meaningfully related in terms of magnitude - even when the unknown has been optimally scaled in each case. *That is, what we are saying is that a distance of say 100 between spectrum A and spectrum B does not carry the same weight as the distance 100 between spectrum A and spectrum C.* Or in other words relativity is lost because the *same* point of reference (the unknown) is not used in each case.

The results in tables I and II bring home this point concerning the two different approaches to matching spectra and computing distance indices of similarity between and among spectra.

Table I shows results for four different methods of scaling the reference spectra with respect to an unknown spectrum. The distance measure is based on equation (1). The "unknown" spectrum (3-hydroxy benzoic acid methyl ester) is shown in figure 4a and three closely related spectra are given in figure 4b, 4c, and 4d. From these results we see that only when optimum mass-dependent matching is used does reference spectrum 3 come seriously into consideration. That is it assumes a distance of 43.27 compared with that of 40.06 for reference 1. From this example we can see that improvements in matching due to optimum mass-dependent scaling can make the difference between a successful match and failure. Also optimum mass dependent scaling will always tend to cancel out and minimize instrumental intensity distortions. The results for optimum constant scaling will always be better than the two non-optimum methods but in many instances optimum constant scaling is not really as effective as the mass-dependent method.

The results in table II are for the same set of spectra but here the unknown has been scaled to each reference spectrum in turn. There are significant differences in these results as compared to the results in table I. Obviously there can be no change for constant highest intensity normalization. Unlike the results in table I the relative magnitudes of the distance indices in table II for optimum matching are not properly related in magnitude.

Comparison of the results in tables I and II clearly illlstrates the difference between scaling spectrum A to spectrum B as opposed to scaling spectrum B to spectrum A. The magnitude of the differences between the results in the two tables serves to emphasize the need for careful consideration of relative scaling to obtain reliable performance in computer-matching of mass spectra.

To further establish the usefulness of optimum mass dependent matching

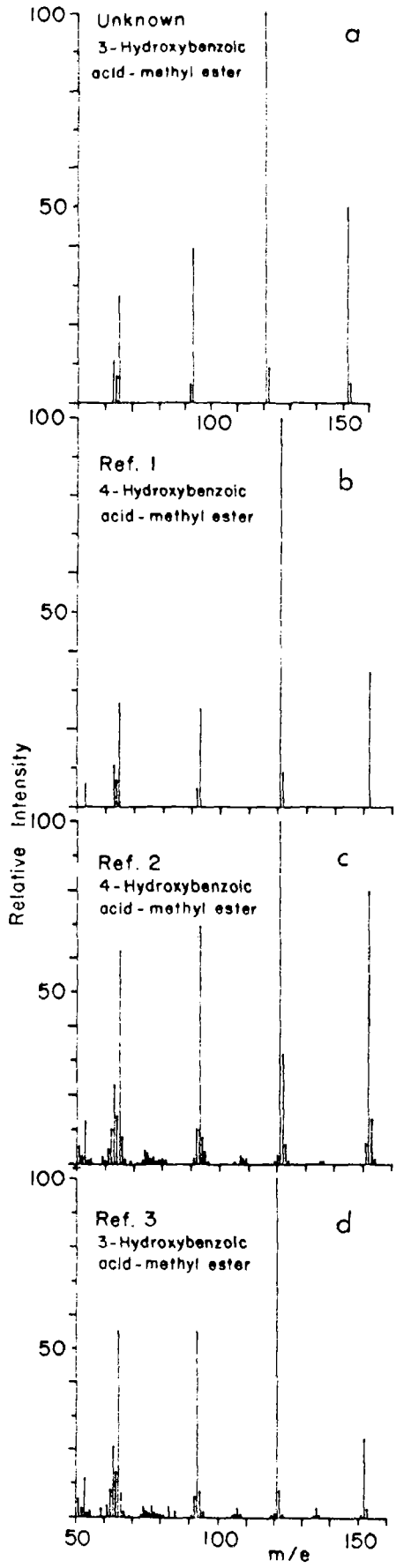


Figure 4

(a) "Unknown" mass spectrum (3-hydroxy benzoic acid methyl ester) used in matching against a set of reference compounds.

(b) Reference spectrum (1) of 4-hydroxybenzoic acid methyl ester.

(c) Reference spectrum (2) of 4-hydroxybenzoic acid methyl ester.

(c) Reference spectrum (3) of 3-hydroxybenzoic acid methyl ester.

(All spectra were taken from the EPA/NIH source library.)

of mass spectra a set of 27 duplicate spectra from the EPA/NIH library were examined. For this group of spectra it was found that the average percentage reduction in distance between the duplicates was 40.1%. This represents a substantial distance reduction and as such underlines the effectiveness of optimum mass-dependent scaling.

Existing Matching Techniques and Optimum Scaling

In the light of the results on optimum scaling it was considered necessary to examine its relevance to other currently available computer-matching methods. To make this study it was chosen to take two closely similar spectra and make several perturbations and study their effects under conditions of optimum constant scaling. The three matching methods considered were the mean square difference method described above (MSD), the absolute difference method (ADIF), and Reed's divergence method (DIV) [5]. The ADIF method involves computing the sum of absolute intensity differences i.e.

$$D_{ADIF} = \sum |U_m - R_m| \tag{10}$$

The divergence method involves computing the distance given by

$$D_{DIV} = \sum \frac{(U_m - R_m)^2}{(U_m + R_m)} \tag{11}$$

For the purposes of this experiment it was decided to make the following three perturbations to the closely similarly spectra fig. (5).

(a) Remove a 10% peak from one spectrum
(b) Change an 82% peak to a 92% peak
(c) Change a 33.4% peak to a 43.4% peak.

To make an absolute comparison among the matching methods their minimum distances separating the two spectra (prior to the three perturbations) were calculated and normalized to 100. Only optimum constant scaling was used for the test because of the difficulty of obtaining it for the

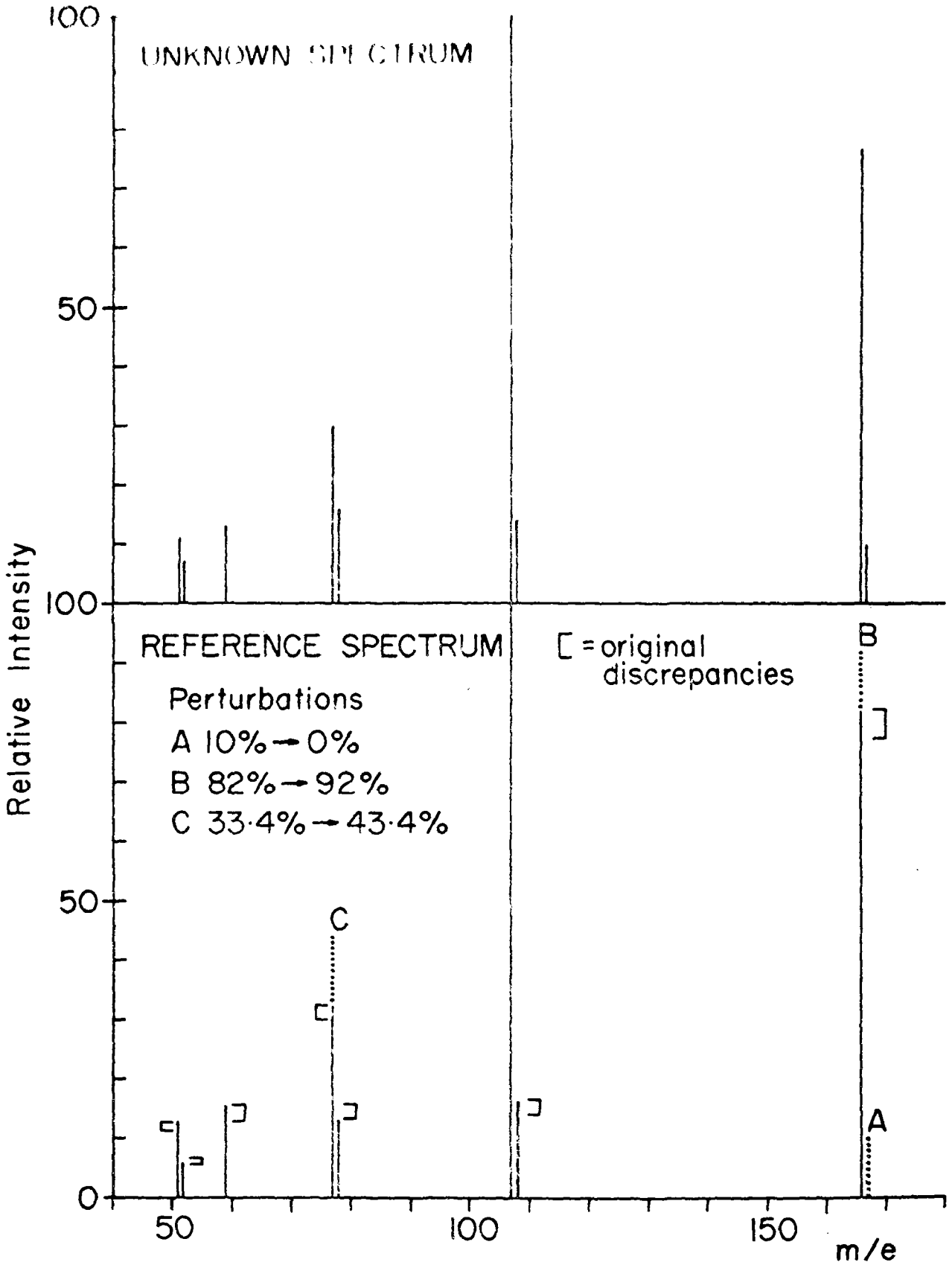


Figure 5

Two similar spectra used for comparison of matching methods are shown. The original differences between the spectra are bracketed and the three perturbations are marked by dotted lines.

absolute difference method (unlike the mean square difference method it does not yield an analytical solution to the scaling problem).

A complete summary of the results using the three perturbations described above is given in table III. What comes most prominently out of these results is that only the divergence method distinguishes between peak absence and a variation in peak intensity of the same magnitude. For example the missing 10% peak causes a change in the distance index from 100 to 1210 whereas a change of 10 from 82% to 92% causes a shift from 100 to 164. Clearly, of the three methods, only the divergence distance focusses on the percentage difference in magnitude between peaks. The small discrepancies that do occur for the three perturbations for the MSD and ADIF methods are caused by differences due to optimum constant scaling.

The large emphasis that the DIV method places on peak absences (and large percentage differences between peaks at the same mass) is at once an advantage and at the same time a disadvantage in that the absence of even very minor peaks can cause considerable changes in the similarity index. It is certainly desirable to weight peak absences more heavily than intensity variations of the same magnitude. However it would seem more fruitful to reduce this weighting such that the absence of very minor peaks does not induce significant changes in the similarity index. One way to do this is to reduce the squared term in the numerator of the divergence equation to an absolute intensity difference. In this method (PDIF) percentage differences are independent of intensity and missing peaks are weighted equally. The distance formulated for this method is

$$D_{PDIF} = \sum \frac{|U_m - R_m|}{(U_m + R_m)} \qquad (12)$$

This method focusses on the overall profile of the spectra being compared.

Results for this method are included in table III. Like the absolute difference method it has the drawback of requiring that optimum scaling be done algorithmically rather than analytically although scaling by equation (5) and then evaluating PDIF would be possible.

The point that this peak perturbation study brings home is that for the divergence method and any other sensitive matching methods, it is essential that spectra be optimally scaled. If this precaution is not taken we can end up with very large discrepancies in similarity indices that will be caused by mass discrimination and other distortions. To summarize these results more generally we can say that the more sensitive the matching method and the closer the spectra are together the more critical it becomes to apply optimum scaling before evaluating similarity indices.

Conclusions

The results of the study on optimum scaling of mass spectra before evaluating their distance indices suggest that the method is quite capable of adequately compensating for instrumental intensity distortions. Furthermore it puts on a more sound formal basis the idea of normalization of spectra for comparison purposes. At the same time the scaling method introduces additional computations before the distance indices can be evaluated. For this reason optimum scaling should only be applied to the small subset of spectra most like the unknown. Methods exist [6-8] for extracting such subsets from a large library,and so the scaling method when used in conjunction with these techniques,does not introduce any serious efficiency problems. In conclusion we can expect the performance of computer-matching methods to be improved significantly by the appropriate consideration of spectrum scaling.

Acknowledgements

The author wishes to thank Miss Ann Titus for typing the manuscript.

References

1. L.R. Crawford, J.D. Morrison, Anal.Chem., 40, (1968) 1464.

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

3. S.L. Grotch, Anal.Chem., 42 (1970) 1214.

4. S. Abrahamsson, Sci.Tools, 14 (1967) 29.

5. S. Farbman, R.I. Reed, D.H. Robertson, M.E. Silva, Int.J.Mass Spectrom.Ion Phys., 12 (1973) 123.

6. J.T. Clerc, P.R. Nageli, Anal.Chem., 46 (1974) 739A.

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

8. R.G. Dromey, Anal.Chem., Jan. (1979).

Table I

Comparative results for an "unknown" matched with three similar spectra each scaled with respect to the "unknown".

Reference Spectrum No.	Optimum Mass Dependent Match	Optimum Constant Match	TIC Equal Match	Largest Both 1000
1	40.06	42.71	47.83	48.20
3	43.27	172.83	174.68	198.72
2	134.83	155.86	181.31	390.44

Table II

Comparative results for an "unknown" matched with three similar spectra. The "unknown" is scaled to each reference spectrum in turn.

Reference Spectrum No.	Optimum Mass Dependent Match	Optimum Constant Match	TIC Equal Match	Largest of Both Spectra 1000
1	35.02	36.57	37.56	48.20
3	18.34	198.61	244.77	198.72
2	250.23	280.72	489.97	390.44

Table III

Response of different matching methods to peak absence and intensity variations.

Method	10% Peak Missing	Peak 82% to 92%	Peak 33.4% to 43.4%	Minimum distance before perturbations
MSD	186.8	187.8	205.9	100
ADIF	156.8	147.0	156.8	100
DIV	1210.0	164.0	281.5	100
PDIF	313.4	107.7	128.6	100