(11) Small, C. W.; Rasmussen, G. T.; Isenhour, T. L. *Appl. Spectrosc.* **1979**, *33*, 444–448.
(12) Delaney, M. F.; Uden, P. C. *Anal. Chem.* **1979**, *51*, 1242–1243.
(13) de Haseth, J. A.; Azarraga, L. V. *Anal. Chem.* **1981**, *53*, 2292–2295.
(14) Azarraga, L. V.; Hanna, D. A. "ERL GC/FT-IR Software and User's Guide (USEPA/ERL)"; GIFTS: Athens, GA 1979.
(15) Milne, G. W.; Heller, S. R. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 204–208.
(16) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1981**, *53*, 889–893.
(17) Lowry, S. R.; Huppler, D. A. *Anal. Chem.* **1983**, *55*, 1288–1291.
(18) "The Coblentz Society Specifications for Evaluation of Research Quality Analytical Infrared Spectra (Class II)". *Anal. Chem.* **1975**, *47*, 945A.
(19) Griffiths, P. R.; Azarraga, L. V.; de Haseth, J. A.; Hannah, R. W.; Jakobsen, R. J.; Ennis, M. M. *Appl. Spectrosc.* **1979**, *33*, 543–548.
(20) Kowalski, B. R.; Bender, C. F. *Anal. Chem.* **1972**, *44*, 1405–1408.
(21) Cover, T. M.; Hart, P. E. *IEEE Info. Theory* **1967**, *IT-13*, 21.
(22) Leary, J. J.; Justice, J. B.; Tsuge, S.; Lowry, S. R.; Isenhour, T. L. *J. Chromatogr. Sci.* **1973**, *11*, 201–206.
(23) Woodruff, H. B.; Smith, G. M. *Anal. Chem.* **1980**, *52*, 2321–2327.
(24) Tomellini, S. A.; Stevenson, J. M.; Woodruff, H. B. *Anal. Chem.* **1984**, *56*, 67–70.

Performance Analysis of a Simple Infrared Library Search System

MARTIN RUPRECHT and JEAN T. CLERC*

Department of Pharmacy, University of Berne, Berne, Switzerland

Received January 31, 1985

The performance of a commercial microcomputer-based IR library search system has been evaluated. In particular, the effects of sample preparation, concentration and path length, base-line correction, and impurities on the similarity score were studied.

INTRODUCTION

Identification of organic compounds through comparison of infrared spectra is a well-known and often used technique. Many infrared library search systems are offered by instrument vendors or have been described in the literature.[1,2] In order to be useful in practical applications a library search system has to meet several criteria. First of all, if the spectrum of the unknown at hand is part of the library, the system should be able to retrieve the respective reference spectrum. If the unknown is not documented in the reference library, suitable reference compounds structurally similar to the unknown should be retrieved. This, of course, is only possible if suitable reference compounds are part of the library. If the reference library does not contain any reference compound sufficiently similar to the unknown, the system should be able to inform the user about this fact. The similarities between the unknown and the reference spectra retrieved by the system should thus be quantified by an appropriate similarity measure. The system compares spectra, whereas the user thinks in chemical structures. Therefore, the similarity measure has to map similarities in the spectra domain into the structure domain. The similarity measure used by the system should thus as far as possible conform to the user's similarity measure for chemical structures. Furthermore, the system should be insensitive to variations in the spectral data due to different sample preparations. It should also tolerate slightly impure samples. The handling of the system should be easy; the search results should become available within reasonable time and should be presented in a form easily interpreted by the user.

Some aspects of the performance of a library search system are quite easy to specify. The search time for instance can be accurately measured, and whether a reference compound identical with the unknown at hand has been retrieved is easily seen. Other aspects of the performance, however, are extremely difficult to quantify objectively, if this is possible at all. The user's own similarity measure for chemical structures depends heavily on the problem at hand; it is extremely context sensitive. Whether the similarity index given by the system to the top ranking reference compound should be interpreted as indicating identity with the unknown sample or rather as a high degree of similarity is a matter of subjective judgement. Evaluation of the tolerance against slight variations due to sample preparations and/or impurities requires an arbitrary decision as to what should be considered as "slight". User comfort and presentation of the results are again highly subjective. Despite these difficulties, we attempted to evaluate the possibilities and limitations of a simple infrared spectra search system. The results are given and commented in the following.

DESCRIPTION OF THE SEARCH SYSTEM

The system evaluated in this study was the Infrared Library Search Software Package marketed by Pye Unicam Ltd., Cambridge, U.K. It operates on the Pye Unicam SP3-080 data console. For this study the data console was connected to a Pye Unicam SP3-300 grating infrared spectrometer. In order to get full control over the reference library and to have unlimited access to the source spectra, a specially prepared reference library was used in this study rather than the library tapes supplied with the system. This reference library consists of 270 spectra of relatively simple organic compounds covering a wide range of compound classes. All spectra were recorded in KBr wafers at concentrations giving about 10% transmittance for the strongest band in the 4000–600 cm^{-1} wavenumber range, with instrument parameter settings commonly used in routine work. The samples were all of analytical grade and were used as received from the supplier. As the data console uses tape cartridges for mass storage, space for the reference library is somewhat limited (a library tape will hold about 1000 reference spectra), and the search time is determined by the access time to the tape. The spectral data may be entered automatically from digitized spectra obtained on the data console or manually from a chart recording. The data is then normalized to a standard form and encoded. The vendor does not give information about the algorithms used for encoding, searching, and similarity calculation. It is claimed, however, that the algorithms used provide compensation for wavenumber or transmittance inaccuracies associated with data read from photoreduced chart records, chart readings from instruments from other manufacturers, or poorly maintained instruments.

USER COMFORT AND PRESENTATION OF RESULTS

Once the spectrum of an unknown compound is recorded and stored in the data console memory, there are few operations to be performed for searching the library. The system assumes that all "cosmetic" operations on the data set (e.g., smoothing, compensation for sloping base line) have been performed before entering the search mode. First of all, the system selects spectrally significant data from the fully di-

0095-2338/85/1625-0241$01.50/0 © 1985 American Chemical Society

Martin Ruprecht was born in 1951. He is a graduate student at the Department of Pharmacy, University of Berne. He received his B.S. degree in Pharmaceutical Sciences in 1979 from the University of Berne. After 2 years of work in a pharmacy, he rejoined the university for a thesis on infrared library search systems under the supervision of J. T. Clerc. Mr. Ruprecht's main research interests are in qualitative and quantitative analysis of pharmaceuticals and related drugs.

Jean T. Clerc was born in 1934. He is a professor of analytical chemistry at the Pharmacy Department of the University of Berne, Switzerland. He graduated in 1958 in chemical engineering at the Swiss Federal Institute of Technology (ETH) in Zürich. In 1964 he received his Ph.D. in technical sciences at the Organic Chemistry Department of the same institution, where he subsequently worked as a research scientist. In 1978 he changed to the currently held position. His research interests are computer applications in analytical chemistry, in particular for the identification and structure elucidation of organic compounds. Prof. Clerc has served for 10 years as President of the Swiss Society for Instrumental Analysis and Microchemistry, and he is Editor of the computer section of the journal *Analytica Chimica Acta.*

gitized spectrum by applying a sliding noise filter automatically matched to the noise level of the spectrum at hand. The full spectrum is then displayed on the visual display unit for the user to select the base-line level. Next, regions of the spectrum where there is no information or where data should be ignored (e.g., solvent bands) are identified. The spectrum is then transferred into the mix area, where it is mixed with any data already there. Thus, spectra resulting from two different solutions of a sample can be mixed together to produce a complete spectrum without solvent bands. Manually entered peak data cannot be mixed with digitized data. This special feature of the system makes cheating rather difficult. Next, the data in the mix area are normalized and compacted to achieve efficient data storage and to minimize the computation time required for data comparison. Then, the code for the sample is sequentially matched against the codes of all reference spectra in the library. An ordered list of the best matches is then presented on the visual display unit. Each hit is identified by its name (16 characters only) and its sequence number. Similarity is given by a score out of 100%.

Operation of the system is quite simple and straightforward. The presentation of the results is adequate considering the limitations due to the somewhat outdated hardware. It takes about 5 min to search through a reference library tape with 1000 reference spectra. This may be acceptable in routine applications. In a research environment, however, the system is definitely too slow.

REPRODUCIBILITY OF THE SIMILARITY SCORE FOR IDENTICAL COMPOUNDS

One expects a library search system to be able to compensate for variations in the spectral data due to different sample preparations. To test for this ability of the system, several samples of three arbitrarily selected comounds [(3,4-dimethoxyphenyl)acetic acid, 4-aminobenzoic acid ethyl ester, and 4-aminobenzoic acid butyl ester] were prepared, using the same concentration range. Their spectra were recorded and stored in the library. The scores obtained by searching each individual spectrum against all others are representative for the score variations to be expected for spectra of identical compounds. The mean score found was 97.8%, the standard error for a single score being 1.7% (95 degrees of freedom). The 95% confidence interval amounts to 3.3%; the 95% confidence region thus extends from 100% down to 94.5%.

Therefore, if a similarity score below 94.5% is found, the statement "the unknown at hand is not identical with the respective reference compound" will be true in 19 out of 20 cases. However, for the inverse case, where the score is above the limit, no probability value can be assessed to the statement "the unknown at hand is identical with the respective reference compound" (error of the second kind[3]). Unfortunately, the second situation has much more practical relevance than the first one. For the decision to consider the unknown at hand to be identical with the top hit of the search list, one has always to take the context into account. Thus, no strict rules can be given.

To test whether the sample concentration has any appreciable effect on the similarity score, a series of samples with varying concentrations were prepared and searched against the standard concentration samples used in the previously described experiment. The concentrations of the standard samples were selected to give approximately 10% transmittance for the strongest absorption band, corresponding to an absorbance value of 1, whereas in the test samples the absorbance of the strongest band varied between 0.2 and 2. No significant variation in the similarity score was observed between these extreme values. The mean similarity score was 96% with a standard error of 1.2% (47 degrees of freedom) for a single value. This is seen not to be significantly different from the result of the previous experiment. Thus, wide variations in sample concentration and/or path length are perfectly compensated for by the system, as claimed by the vendor.

BASE-LINE CORRECTIONS

Solid-phase spectra often exhibit sloping base lines. This is predominantely due to scattering phenomena at the sample particles embedded in the matrix. This effect can often be at last partially removed by regrinding the sample. This, however, amounts to an additional investment of time and labor. The data acquisition system provides means to correct for sloping and otherwise irregular base lines. The respective command accepts a series of wavenumber values specified by the user as laying on the base line. The system linearly interpolates the base line between these specified points. The new base line is then shifted to 100% transmittance. A spectrum treated in this way is referred to as a "flattened" spectrum.

With spectra exhibiting a reasonably level base line, one does not expect any significant changes in the similarity score when flattened spectra are used. Experiments do confirm this. The similarity scores generally are found to be above 98.5% if either the unknown or the reference spectrum is flattened. If a flattened spectrum is compared to itself, the score is of course 100%.

With heavily sloping base lines the situation becomes entirely different. The results are not easy to evaluate, as "sloping" is difficult to quantify precisely. However, the following generalizations can be made. Whether the original spectrum or a flattened spectrum of an unknown exhibiting a sloping base line is matched against a library of untreated spectra is of no importance, the score values do not vary significantly. If, however, the original spectrum is matched against a library of flattened reference spectra, significant drops in the score may be observed. In many cases the similarity score drops from 100% down to well below 90%. In other cases the loss in similarity is small. In all cases, however, the loss induced by flattening the reference spectra is higher than the loss induced by flattening the spectra of the samples. This assymmetry is puzzling at first. One would naively assume that if A is similar to B to a certain extent, then B is similar to A to the same extent. However, the library search system treats the reference spectra and the sample spectra differently. The selected reference spectra should show all relevant features exhibited by the unknown. Absence of a feature of the unknown in the reference is heavily penalized by a sizable deduction from the score. Features exhibited by the reference in excess to those shown by the unknown are of little concern and lead to only a small penalty in the similarity score. Through this assymmetric treatment one ensures that reference compounds considered similar to a given unknown do contain the unknown as a part, rather than being a part of the unknown. Thus, the unknown is characterized by the intersection ("and" combination) of the best reference compounds rather than by their union ("or" combination). Flattening decreases peak absorption intensities; thus, it can result in the loss of spectral features. If spectral features are lost in the sample spectrum, the respective reference compounds will get only the small penalties for excess features. If, on the other hand, the reference spectra lose features by being flattened, they will accumulate the large penalities associated with missing features.

Therefore, if the spectrum of a sample exhibits a sloping base line, it is generally not worthwhile to invest too much work into improving the data, as this will not improve the search results to a significant extent. However, only spectra of the highest attainable quality should be included into the library.

The peak intensities are measured relative to a horizontal base line, the exact position of which is specified by the user. As long as the base-line level is set reasonably, the score results are hardly affected. The scores drop down only when the base line is grotesquely misplaced by 10% transmittance or more.

RETRIEVAL OF SIMILAR COMPOUNDS

The performance of any library search system is limited by the spectra library it is based upon. The system can put out only chemical structures represented in the library. If, for a given unknown no suitable reference compound is in the library, even the most powerful and sophisticated system is helpless. Thus, performance in retrieval of similar compounds depends critically on the composition of the library. The probability of having useful reference compounds in the library increases with the size of the library. Thus, performance is expected to increase with library size. On the other hand, with too large a library the performance tends to decrease due to the following reasons. First of all, search time and storage costs increase approximately linearly with library size, and maintenance problems tend to grow even faster.[4,5] Furthermore, if an unknown triggers the retrieval of a given reference compound, it will generally also select for retrieval all other reference compounds belonging to the same compound class. They will then occupy a large part, if not the full hit list. Thus, potentially useful reference compounds from other groups are pushed out of the list and will never come to the attention of the user. In such a situation only the first entry from a given compound group provides useful information. It tells the user that the system believes the unknown to be similar to the respective reference compound, i.e., to belong to the same compound class. The second and all the following entries just repeat that message over and over again. The fact that a whole group of similar reference compounds is retrieved by the system does in no way improve the reliability of the result. If the system blunders with the first reference compound of a group, it will do so with all others.

Thus, too small a library as well as too large a library will have detrimental effects on the performance for retrieval of similar compounds. A small library will often not contain a suitable reference compound. In a library too large groups of similar compounds will be retrieved together and will displace other useful references from the hit list. Groups of compounds too similar to each other under one similarity measure may split up under an improved similarity measure. Thus, the optimal size of the library depends on the quality of the similarity measure employed. One should definitely avoid having in the library groups of compounds that cannot be discriminated with the similarity measure used.[5]

The reference library used in this study comprises 270 compounds. This is certainly much too small a number to give optimal performance. However, to get an estimate of the recall with reasonable effort a library of limited size is absolutely required. Even though about 40 compound classes are represented, 70% of the compounds documented in the library are drawn from nine major categories only (carboxylic acids, esters, amides, alcohols, phenols, amines, amine salts, ethers, and nitro compounds).

To estimate the ability of the system for retrieval of structurally similar compounds a set of 110 sample spectra was searched against the library. Each unknown was represented once in the library. In 86 cases the library spectra were used for searching, and in 24 cases the spectra were rerecorded. In every case the identical compound was retrieved as the top entry in the hit list. For the newly recorded spectra, the scores were between 94% and 100%. Careful inspection of the results of nonidentical compounds allows for the following generalizations. Reference compounds scoring 90% or higher tend to be either homologous compounds or positional isomers of the sample. Only in two cases an apparently nonrelated compound got a score above 90%. 4-Hydroxybenzoic acid ethyl ester retrieved 3,5-dimethoxy-4-hydroxybenzaldehyde with a similarity score of 91.7%, and 4-hydroxybenzoic acid butyl ester retrieved 4-hydroxybenzoic acid (91%). About 40% of the reference compounds retrieved with scores between 90% and 85% were homologous compounds or positional isomers of the respective samples. Other reference compounds falling in this similarity range, however, did not exhibit any useful and obvious structural similarities.

It is true that, for example, aromatic carbonic acids will retrieve other aromatic carbonic acids. However, one also finds a significant number of apparently unrelated references interspersed between the carbonic acids. In a test case, where the sample is only assumed to be unknown, one can easily identify such vaguely similar reference compounds and count them as hits. In real life situations, however, where no prior information is available, they are useless.

Recall is difficult to estimate, even when using a library of limited size. Moreover, the estimate is of doubtful value as it depends heavily on the composition of the library and on the similarity concepts of the persons doing the estimation. In our case, we estimate that in roughly half of all cases studied at least one potentially useful reference compound was missed.

These findings may be generalized into the following recommendations. Among the reference compounds scoring more than 95% similarity, there may be one identical with the unknown at hand. Compounds retrieved with a score above 90% generally exhibit skeletons and/or functional groups identical with the sample. Below 90% the proportion of garbage increases rather fast. In real situations there is no way to differentiate between garbage and potentially useful spectra. It is thus not worthwhile to look at references retrieved with scores below 90%.

In all 24 cases where the sample spectrum was not identical with the library spectrum, the compound identical with the unknown was found in the top position of the hit list. This is just a happy coincidence, however. If two or more compounds closely similar to the unknown are in the library, the ranking sequence may sometimes become garbled. This is illustrated in the following example. For two very similar compounds, namely, 4-aminobenzoic acid ethyl ester and butyl ester, five spectra were recorded and introduced into the library. Then, each of these 10 spectra was used in turn as an unknown. The resulting hit lists, equivalent to 200 single searches, were analyzed. The mean scores for the identical compounds were found to be 98.9% and 97.8%, respectively. The similar compounds averaged 96.3% and 95.1% only. However, in 3% of all cases the identical compound reached but the second rank in the hit list. This result indicates that in most cases the identical compound, if present in the library, will occupy the top position in the hit list. However, if the score values of the top ranking compounds do not differ appreciably, the compound identical with the unknown may not lead the list.

INPUT OF SPECTRA AS PEAK TABLES

The system allows for input of spectra as tables of peak positions and intensity values. This is deemed to be useful for extension of the library with spectra extracted from the literature or for searching with spectra not available in fully digitized form. To check the performance for this application mode, the spectra of 15 compounds were recorded on an instrument from a different manufacturer. Peak tables were then prepared by reading the wavenumber values to the nearest 5 cm^{-1} (10 cm^{-1} above 2000 cm^{-1}) and the intensity values to the nearest 5% transmittance. The search results showed no significant deterioration of the similarity scores for the identical compounds. Score values for tabulated spectra were always lower as for the full spectra but were in all cases above the level of 94.5%, which is expected for identical compounds. Variations in the score for similar compounds did not show a consistent trend.

SENSITIVITY TO IMPURITIES

In real-life situations one has often to deal with impure samples. It is thus interesting to know how the system reacts to compounds containing impurities. One expects the sensitivity against impurities to depend on the spectral similarity between the impurity and the main component. If both compounds show similar spectra, low-purity samples should be well tolerated. With strongly dissimilar spectra, however, the score should become quite sensitive. These expectations are fully corroborated by the experiment. The identity score for glutaconic acid, for example, drops below the 95% level only upon admixture of more than 25% (w/w) of succinic acid. The same holds true for the inverse case, where succinic acid is mixed with glutaconic acid. This behavior is typical for compounds scoring above about 85% similarity when compared to each other in pure form. Results for mixtures of compounds exhibiting dissimilar spectra are less easily generalized. In these cases sensitivity to impurities is increased. Glutaconic acid tolerates less than 10% (w/w) of 4-hydroxybenzoic acid ethyl ester. The sensitivity, if specified in weight percent, obviously depends strongly on the extinction coefficients of the impurity spectrum. An impurity exhibiting weak absorptions only is again relatively well tolerated. Thus, glutaconic acid still scores above 95% similarity even with up to 25% (w/w) of benzilic acid methyl ester, even though the spectra are rather dissimilar.

A third important parameter is the number of bands of significant intensity exhibited in the spectra. A compound exhibiting many bands in its spectrum is not too susceptible to impurities showing only a few important bands, whereas in the opposite case the similarity score becomes sensitive. This effect can become quite pronounced. Glutaconic acid shows few important bands in its spectrum, whereas the spectrum of 4-hydroxybenzoic acid methyl ester has many strong bands. Consequently, glutaconic acid is very sensitive to 4-hydroxybenzoic acid ethyl ester as an impurity, tolerating less than 10% (w/w). However, for the latter compound more than 30% (w/w) of glutaconic acid has to be added to lower the similarity score below 95%.

In summary, three important parameters govern the sensitivity to impurities, namely, the similarity of the spectra, the relative intensity, and the number of bands with significant intensities. High tolerance against impurities is generally found if the compound at hand exhibits in its spectrum a large number of strong bands relative to the spectra of the impurities and if the impurity spectra are similar to the sample spectra. However, interactions between the effects of these parameters make predictions inaccurate and unreliable.

CONCLUSIONS

In summary, it can be stated that the evaluated IR library search system performs well in a routine environment. Its major drawback is the slow search speed. It takes 4–5 min to search a library of 1000 reference spectra. It reliably retrieves from the library compounds identical with the sample at hand, even with imperfect spectra. Compounds identical with the sample are generally retrieved with a similarity score above 94.5%. Reference compounds scoring more than 90% are in almost all cases either homologous to or positional isomers of the unknown. References with similarities below 90% are useless. Differences in concentration and/or path length are perfectly compensated. Base-line corrections may slightly improve the results but are generally not worthwhile in a routine environment. The system exhibits a high tolerance against low-quality spectra and even accepts peak tables as input without significant deterioration of the results. Sensitivity to impurities varies with spectra type. Some 3% (w/w) are always tolerated; in favorable cases this limit may be as high as 30% (w/w). All in all, the system meets or even surpasses the specifications and has proved to be really useful.

REFERENCES AND NOTES

(1) Delany, M. F. *Anal. Chem.* **1984**, *56*, 277R and references cited therein.

(2) Blaffert, T. *Anal. Chim. Acta* **1984**, *161*, 135.

(3) Davies, O. L. "Statistical Methods in Research and Production"; Oliver and Boyd: London, 1961.

(4) Koenitzer, H. Ph.D. Thesis, ETH Zurich, 1980.

(5) Clerc, J. T.; Koenitzer, H. In "Computational Methods in Chemistry"; Bargon, J., Ed.; Plenum Press: New York, 1980; p 1.