# A Comparative Study of Point-to-Point Algorithms for Matching Spectra

Jianfeng Li[a], D Brynn Hibbert[a*], Stephen Fuller[b], and Gary Vaughn[b]
a School of Chemistry, The University of New South Wales, Sydney, Australia
b Environmental Forensic & Analytical Science, Department of Environment and Conservation (NSW), Lidcombe, Australia
* Author for correspondence.
b.hibbert@unsw.edu.au

## Abstract

Matching spectra is necessary for database searches, assessing the source of an unknown sample, structure elucidation, and classification of spectra. A direct method of matching is to compare, point by point, two digitized spectra, the outcome being a parameter which quantifies the degree of similarity or dissimilarity between the spectra. Examples studied here are correlation coefficient squared, and Euclidean cosine squared, both applied to the raw spectra and first difference values of absorbance. It is shown that spectra do not fulfill the requirements for a normal statistical interpretation of the correlation coefficient; in particular they are not normally distributed variables. It is therefore not correct to use a Student-*t* test to calculate the probability of the null hypothesis that two spectra are not correlated on the basis of a correlation coefficient between them. We have investigated the effect on the similarity indices, of systematically changing the mean and standard deviation of a single Gaussian peak relative to a reference Gaussian peak; and of changing one peak, and of changing many peaks, in a simulated ten-peak spectrum. Squared Euclidean cosine is least sensitive to changes and the first difference methods are most sensitive to changes in mean and standard deviation of peaks. A shift of the center of a peak has a greater effect on the indices than increases in peak width, but a decrease in peak width does lead to significant changes in the indices. We recommend that if these indices are to be used to match spectra, appropriate windows should be chosen to avoid dilution by regions with no significant change.

**Keywords:** Matching spectra, correlation coefficient, Euclidean cosine, similarity index

## 1. Introduction

The comparison of two spectra is necessary for classification of a spectrum [1], searching a database of spectra to identify an unknown sample [2], to decide if two materials come from a common source[3], in process control against the target spectrum of an acceptable product [4], or to elucidate the structure of a compound [5]. It is realized using a measure of the similarity between the spectra, or, conversely, the distance of one spectrum from the other in some measurement space. If the queried spectrum is in the database a perfect match can be achieved, but if only part of spectrum can be found, the result might be a number of partially matched hits. In environmental analysis, when material spilled in the environment is exposed to weathering, chemical, physical and biological processes will happen [6]. If so, the spectrum of a spill will not always make an exact match with the spectrum of its source, and so to correctly identify a spill for forensic applications requires some allowed tolerance to be applied. In quality control of herbal medicines, due to the changes of season, place of harvest, pre-processing and the conditions of analyses, chromatographic fingerprints of the same herbal medicine are not always the same [7]. Therefore any method of matching spectra will need to distinguish between the same material that has been changed, and different materials with similar spectra.

Methods for comparing spectra can be divided into direct and indirect methods. Direct matching methods use the spectral data directly, and indirect matching methods use derived information from spectra. The latter rely on identification of selected peaks and the extraction of information from them, and have been used by human experts employing visual comparison [8], old computer spectral databases or comparison by simple mathematical calculations such as the measurement of ratios [9]. Multivariate data analysis techniques [10], artificial neural networks [11] and distance/angle [12] methods are direct methods which treat digitized spectra directly without any prior identification of peaks. (Note that it is also possible to use multivariate methods on peak area, or ratio data).

Vibrational and electronic spectra of mixtures can rarely be deconvoluted and assigned to individual components in contrast to the output of other methods such as nuclear magnetic resonance (NMR), chromatography or mass spectrometry, in that individual molecules do not give a single, or a small number of, identifiable peaks. Small informative peaks and overlapped peaks in Fourier Transform infrared (FTIR) spectra are not easily identified by computer software and the shape of a peak, which is important for comparison, is difficult to describe accurately. These difficulties can be partially avoided by using point-to-point matching methods because all the data points in a spectrum are used. Similarity/distance methods based on point-to-point matching also have the distinct advantage, compared to pattern recognition techniques, that they only require two spectra, and not a set of spectra belonging to different classes. Point-to-point matching is a direct method in which equal length vectors describing two spectra (intensities, absorbances or detector response) are compared point by point, and a single statistic calculated. The Pearson correlation coefficient is an example of such a similarity index.

In our previous work [13] on matching spectra of petroleum oils, we have found that although different oils can exhibit very different spectra, they can also be very similar. A spectrum of a slightly weathered oil is almost identical to the spectrum of a fresh sample, but it is possible that the difference between the spectrum of a fresh oil and its weathered derivative is greater than the difference between this spectrum and the spectrum of another, highly similar, fresh oil. If we draw the distributions of a similarity measure of such a situation, we see a broader distribution of the spectral similarity of different oils, a narrower distribution for spectra of the same oil but there is often an overlap region leading to false positive or false negative assignments. The success, or otherwise, of a matching method, therefore rests on its ability to discriminate subtle differences in samples that are inherently similar. The task becomes harder

with real samples from the environment because of weathering and introduction of interfering species such as water.

Measures of similarity usually have a defined range, for example, the Pearson's correlation coefficient lies between –1 to 1, or its square between 0 and 1. The minimum or maximum similarity is not always met in the real world, nor is the distribution of values normal. The meaning of the actual value of a similarity index depends on the situation in which it is applied. A correlation coefficient of 0.99 does not mean a match in all situations. It is the analyst's responsibility to decide whether a pair of spectra match according the actual situation. This cannot be done without a knowledge (explicit or from experience) of the distribution of the index, against which a particular result is judged.

An IR spectrum not only depends on the particular functional groups, it also reflects the arrangement of these functional groups within a molecule. An IR spectrum is thus, in contrast to NMR or mass spectra, predominantly a property of the whole molecule and not just the sum of the properties of it's constituents. The characteristic band of a functional group and the shifts when it connects to different neighboring structures have been described [14,15]. Not only is there not a complete spectral library of the form of bands arising from a particular group in all chemical environments, but also the simple summation of the contributions of all the bands of functional groups in a molecule does not give the real spectrum. It is therefore not possible to predict the spectrum of a complex environmental sample, even if the constituent compounds are known. The only thing we can do is to investigate the effect of the change of peaks of a spectrum itself. To deconvolute an IR spectrum into Gaussian peaks is more difficult than to fit, for example, an x-ray photoelectron spectroscopy (XPS) spectrum. It is impossible to start from a real spectrum and decompose it into small Gaussian peaks. We have therefore conducted the study reported here by simulating increasingly complex spectra, which have been compared pair-wise to yield distributions of similarity indices. Starting from a two simulated Gaussian peaks we investigate the effect, on a number of similarity measures, of differences in the position and width. The study is extended to changes in a single peak among a simulated spectrum of ten random Gaussian peaks, then to changing more peaks. Finally we report the distribution of similarity indices for real spectra, augmented by simulated spectra derived from the variance of Fast Fourier Transform (FFT) coefficients.

## 2. Theory

### 2.1 Similarity indices

A number of measures of similarity have been proposed that can be classed as a Minkowski distance

$$D_{1,2} = \left( \sum_i \left| x_{1,i} - x_{2,i} \right|^m \right)^{1/m} \quad (1)$$

The spectra are described by vectors of equal length with individual elements $x_{1,i}$ and $x_{2,i}$ . The Euclidean distance is given by $m = 2$, and Manhattan (city block) distance is when $m = 1$. Statistical measures include the correlation coefficient, and for approaches based on binary variables the best known is the Tanimoto index, which counts the proportion of points that are mutually above or below a threshold [16]. Similarity indices for use with infrared are discussed by Varmuza et al. [20].

Four point-to-point similarity indices are studied here: squared correlation coefficient (Cor), squared first difference correlation coefficient (DCor), squared Euclidean cosine (Euc) and squared first difference Euclidean cosine (DEuc). Their definitions can be found in Table 1.

It is seen that the difference between correlation coefficient and Euclidean cosine is that the data is mean centered in the calculation of correlation coefficient. For a symmetrical peak, on

taking the first difference, the mean of the spectrum is zero and so DCor = DEuc. The first derivative of a spectrum is often taken to remove the effect of a sloping baseline.

**Table 1. Definitions of similarity indices. $A_1$ ($A_2$) is the vector of intensities or absorbances of spectrum 1 (2), with individual element $A_{1,i}$, ($A_{2,i}$)and mean $\overline{A}_1$ ($\overline{A}_2$)**

| | |
|---|---|
| Squared correlation coefficient (Cor) | $\left(\frac{Cov(\mathbf{A}_1,\mathbf{A}_2)}{s_{\mathbf{A}_1}s_{\mathbf{A}_2}}\right)^2 = \frac{\left(\sum_i \left(A_{1,,i}-\overline{A}_1\right)\left(A_{2,i}-\overline{A}_2\right)\right)^2}{\sum_i \left(A_{1,,i}-\overline{A}_1\right)^2 \sum_i \left(A_{2,i}-\overline{A}_2\right)^2}$ |
| Squared first difference correlation coefficient (DCor) | $\frac{\left(\sum_i \left(\Delta A_{1,,i}-\overline{\Delta A}_1\right)\left(\Delta A_{2,i}-\overline{\Delta A}_2\right)\right)^2}{\sum_i \left(\Delta A_{1,,i}-\overline{\Delta A}_1\right)^2 \sum_i \left(\Delta A_{2,i}-\overline{\Delta A}_2\right)^2}$ where $\Delta A_{1,i} = A_{1,i+1} - A_{1,i}$ |
| Squared Euclidean cosine (Euc) | $\frac{\left(\sum_i A_{1,i}A_{2,i}\right)^2}{\sum_i A_{1,i}^2 \sum_i A_{2,i}^2}$ |
| Squared first difference Euclidean cosine (DEuc) | $\frac{\left(\sum_i \Delta A_{1,i}\Delta A_{2,i}\right)^2}{\sum_i \Delta A_{1,i}^2 \sum_i \Delta A_{2,i}^2}$ where $\Delta A_{1,i} = A_{1,i+1} - A_{1,i}$ |

## 2.2 Simulation of spectra

The squared correlation coefficient or other similarities defined by the formulae in Table 1 do not have any chemical meaning as such and cannot offer insights into changes in composition or structure. It is well known that the points in a spectrum are not random and they can be described as part of a Gaussian peak or the summation of Gaussian peaks. But it is by no means easy to deconvolute a real spectrum into a series of peaks of given distribution. Our simulation, therefore, is of a spectrum for which we have all details of each peak from the beginning of the investigation.

A Gaussian peak can be described with three parameters: the center = $\lambda$ ($\mu = \lambda$), the peak height ($h$) and the full width at half maximum (FWHM) = $w$, ($\sigma = 2.35\ w$) and an IR spectrum $\mathbf{s}$ at a vector of wavelengths ($\mathbf{x}$) with $n$ Gaussian peaks can be formulated as:

$$\mathbf{s} = \sum_{i=1}^{i=n} \mathrm{G}\left(\mathbf{x}, \lambda_i, h_i, w_i\right) \tag{2}$$

where

$$\mathrm{G}\left(\mathbf{x}, \lambda_i, h_i, w_i\right) = h_i \exp\left(-4(\ln 2)\left[\frac{\mathbf{x}-\lambda_i}{w_i}\right]^2\right) \tag{3}$$

The spectral simulations were performed on a Pentium 4 personal computer using Matlab 6 R12 for Windows (The Mathworks Inc., USA). The spectral comparisons, calculation of similarity indices and plotting of distributions were carried out using programs in Matlab,

Excel (Microsoft Corporation, Redmond) and custom programs which were developed in our laboratory. We investigated two single peak spectra, ten-peak spectra, and simulated spectra based on experimental data, with details given in the following sections.

### 2.2.1 Single peak

For a single peak spectrum, because there are only three parameters (height, width and position), an exhaustive investigation of the effects of the changes of the parameters is possible. A new Gaussian peak was generated according to Equation 3 from a reference peak $\lambda = 0$; $w = 2.35$ (corresponding to a standard deviation of 1) with each parameter changed according to

$$P_{new} = P_{ref} * (1 + \Delta) \quad (4)$$

where $0 \leq \Delta \leq 1$, and $P$ is $\lambda$ or $w$.

The similarity indices used are insensitive to simple scaling (changing $h$ in Equation 3), and so this parameter is not investigated. For each parameter, $\Delta$ was varied and the effect on the similarity indices investigated.

### 2.2.2 Ten-peak spectra

A ten-peak spectrum was created at random in the interval $\lambda = 0$ to $\lambda = 10$, calculated at intervals of $\lambda = 0.1$ as the reference spectrum for the study (Figure 1).

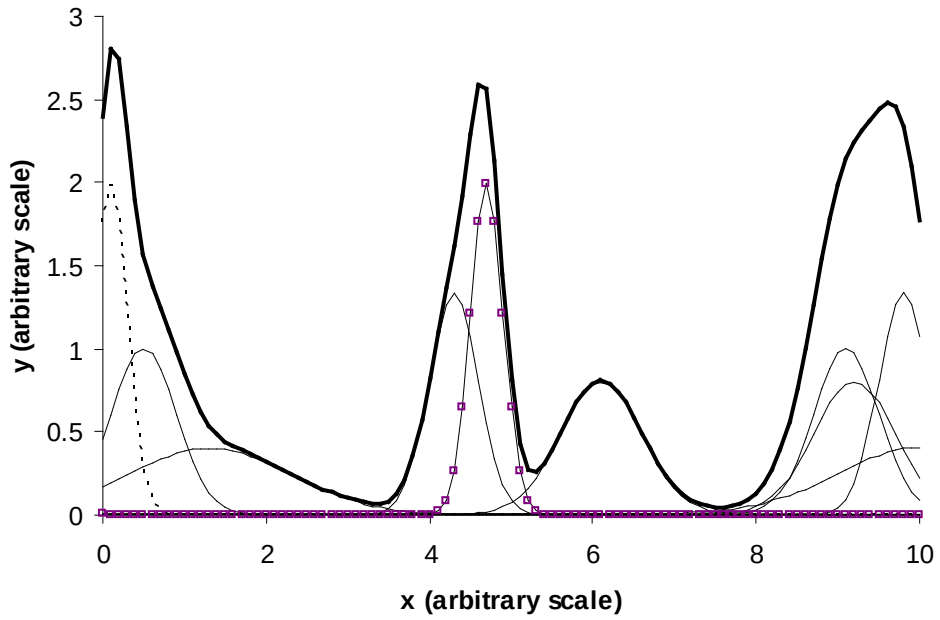



**Figure 1. Simulated ten-peak spectrum (solid line) and it's constituent peaks (dotted lines). The peak indicated by marker points is the reference peak used to determine the effect of changes on similarity indices ($\lambda = 4.7$, $\sigma = 0.2$).**

Because of the number of peaks the number of possible combinations of the changes of parameters is huge. As a first example, we chose a middle peak (circles in Figure 1, $\lambda = 4.7$, $w = 0.477$) and created new spectra with changes in the position and, separately, the width of the peak. The position was varied between $\lambda = 3.7$ and $\lambda = 5.7$, and the width between $w = 0.03$ to $w = 1$. In the presence of other peaks, changes do not have a symmetrical effect and so values were varied either side of the reference peak values. With a single value changing it is possible to graph the effect on the similarity indices. When a number of peaks were changed, a Monte Carlo approach was taken and the distributions of indices recorded. Each parameter was

changed in a uniform random range of $\Delta$ (Equation 4) from –0.5 to +0.5. The change of the center of a peak was limited to ±FWHW. Using this procedure 10,000 spectra were generated, compared with the reference spectrum, and the distribution of values of the similarity indices calculated.

### 2.2.3 Simulation based on a Fast Fourier Transform of real spectra

To generate simulated spectra based on a real spectrum, ten replicate spectra of the same sample of calcium carbonate were collected and FFT was used to decompose each of them into frequency information. The means and standard deviations of each of the first 255 coefficients of the FFT were calculated. From these one hundred normally distributed random numbers were generated for each coefficient and thus one hundred simulated IR spectra were generated. The pair wise similarity indices were then computed of the 110 spectra, giving 110 x 109/2 = 5995 values.

## 3. Experimental

Infrared spectra of $CaCO_3$ (Analytical reagent purity, dried and ground before use) were collected on a Fourier transform infrared spectrophotometer (Excalibur FTS 3000, Bio-Rad). The samples were analyzed under the same conditions using the same KBr cell, which was cleaned between samples, and the spectra were recorded from 4000 to 652 $cm^{-1}$. 32 scans at a resolution of 4 $cm^{-1}$ were collected and averaged for the background and for each sample. A 0.05 mm spacer in the cell ensured consistent thickness of the oil sample. The data for the calculation of similarity indices for these samples was a vector of 837 mean absorbances, and the calculations were performed using programs written in the Matlab (release 12.1, The Mathworks Inc, USA) environment.

## 4. Results and discussion

### 4.1 The applicability of the Student-*t* test for matching spectra

The significance of the linear correlation of two vectors given a Pearson's correlation coefficient can be tested using

$$t = \frac{r}{\sqrt{1-r^2}} \qquad (5)$$

to transform the correlation coefficient to a Student- *t* value [17]. The prerequisites of using the *t*-test, for the null hypothesis that the two vectors are uncorrelated, are that the elements of each vector are normally distributed; (i.e. each vector is a random variable) and each pair of elements of the vectors is independent (i.e. the order of the pairs of elements does not affect *r*). Testing *r* at the 95% level ($\alpha = 0.05$, one tail) means that we reject the hypothesis that the two spectra are not correlated at all when the probability of finding the particular value *t* (calculated from *r* by Equation5) falls below 0.05. For infinite degrees of freedom $\Pr(T \geq t) < 0.05$ when $r^2 > 0.74$, and this is only increased to $r^2 > 0.77$ for ten degrees of freedom. This is shown in Figure 2.

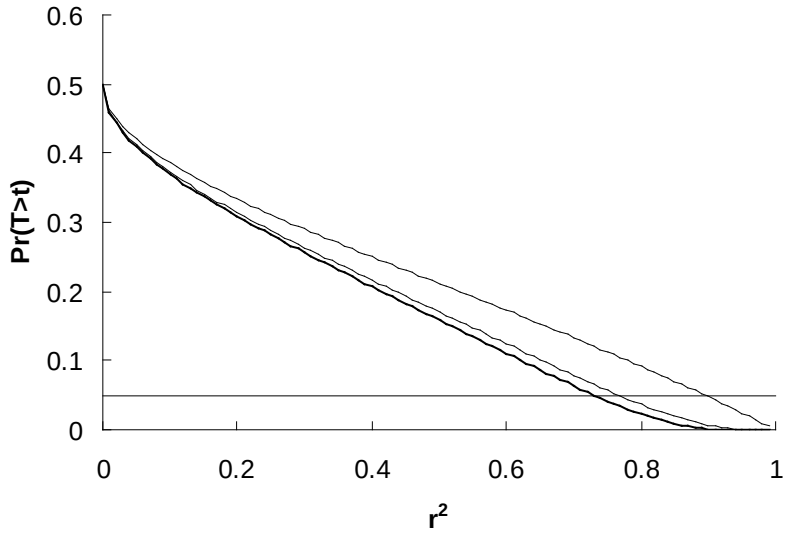



**Figure 2. The probability of $H_0$ that two vectors are not linearly correlated as a function of the squared correlation coefficient. Degrees of freedom: solid line – infinity, dashed line – ten, dash-dot line – two. A horizontal dotted line is drawn at Pr($T \geq t$) = 0.05.**

We first note that the distribution of intensities of a Gaussian peak is not at all normal, thus violating the first assumption of the use of a correlation coefficient (see Figure 3 ).

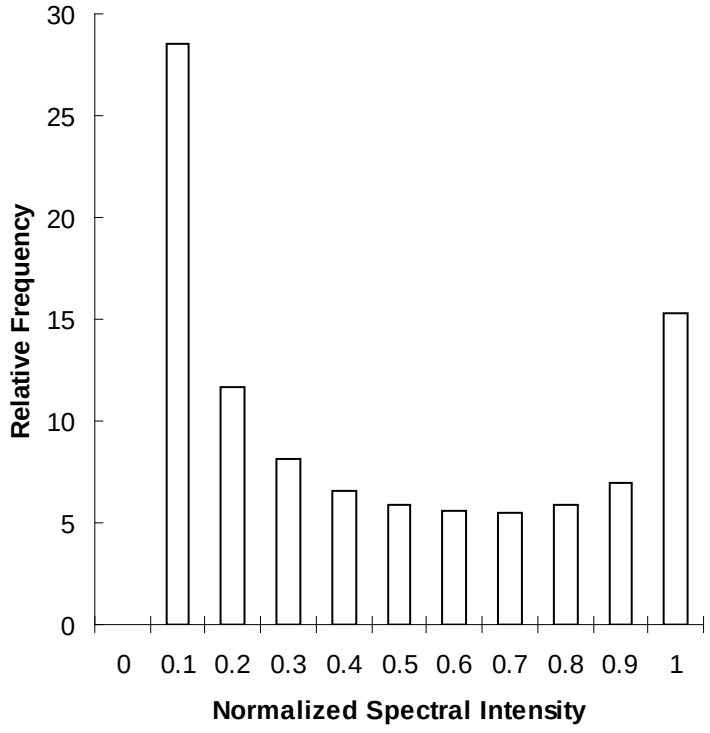



**Figure 3. The distribution of intensities of a Gaussian peak with mean 0 and standard deviation 1, calculated at 600 points between $x$ = -3 and $x$ = +3**

The classical Student-*t* test for Pearson's correlation coefficient does not fit well for the comparison of IR spectra because of the violation of the assumptions of the distribution. Other distributions of correlation coefficient under other conditions have been studied and the exact analytical resolutions are provided in texts by Anderson [18] and Muirhead [19]. Unfortunately, the distributions are complicated and not easily used for spectral comparison.

Also because the distribution of spectral intensity is unknown, but clearly not a Gaussian distribution for real spectra, it is not possible to predict the distribution of the correlation coefficient of two spectra using classical statistical methods.

## 4.2 Comparison of single peak spectra

The effects of changing the width of a single peak and its position are very different. For the simple Gaussian peak a change in position by 0.5 $\sigma$ leads to a value of $r^2$ of 0.74 which has a *t* value (Equation 5) of 1.67 and probability of 0.05. In this case, therefore, a peak shift of half a standard deviation (or FWHM of 1.17) leads to rejection of the hypothesis that the spectra are perfectly correlated. On the other hand, for a second peak centered on the reference peak but with greater width, it is impossible to broaden the peak sufficiently to lead to a non-significant *t*-value, i.e. no amount of broadening can destroy the conclusion that the peaks are correlated. A narrower peak does become significantly different (i.e. the correlation coefficient no longer fails the t-test) when $\sigma = 0.48$. Figure 4 shows the changes in the peak that have to occur to conclude that there is a significant difference between it and the reference.

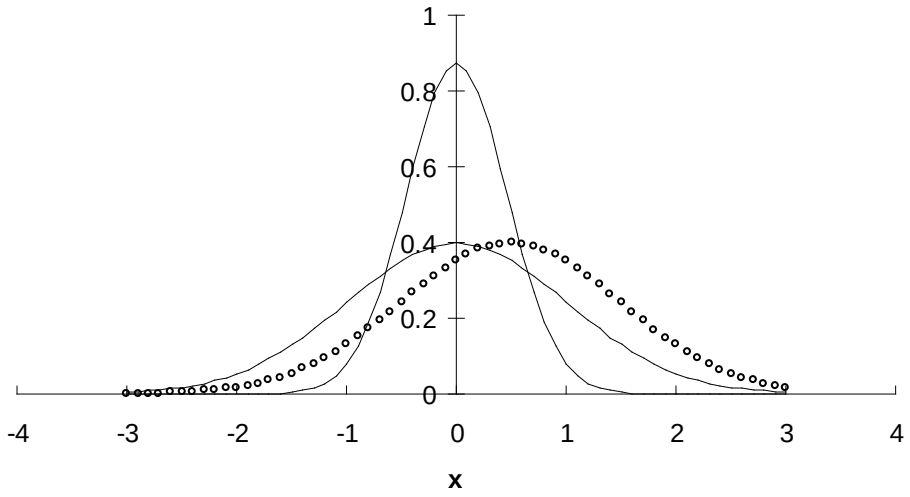



**Figure 4. A reference Gaussian peak (solid line, $\mu = 0$, $\sigma = 1$), solid line, with peaks for which the correlation coefficient just fails the significance test of Equation 5 at the 95% level ($\alpha = 0.05$). Dashed line: $\mu = 0$ and $\sigma = 0.46$, circles: with $\mu = 0.50$ and $\sigma = 1$.**

The Euclidean cosine and difference statistics behave in a similar manner. The effects of changing the mean and standard deviation on the different indices are shown in Figure 5.

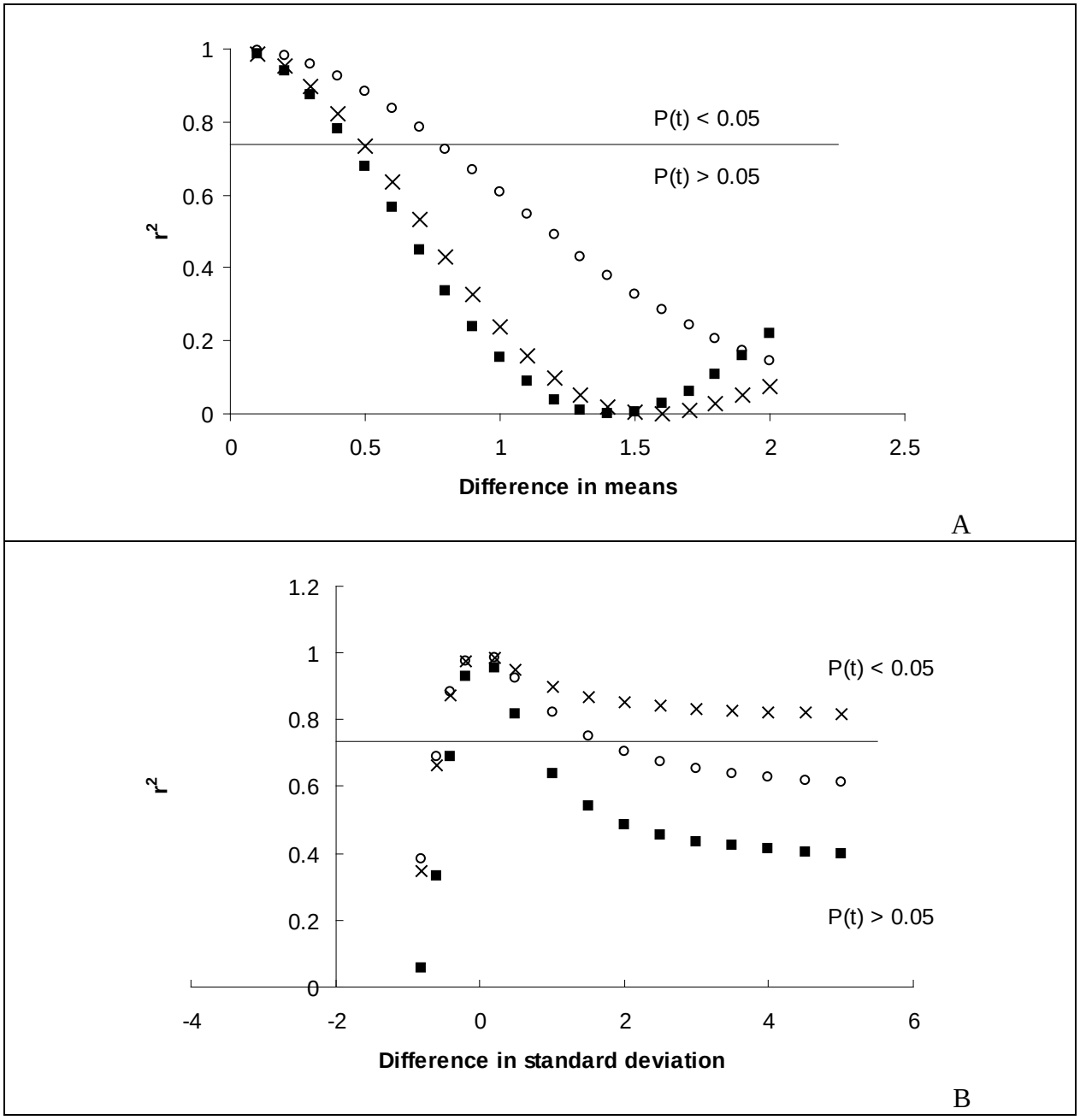



**Figure 5. Values of similarity indices as a function of changes in the parameters of a single Gaussian peak relative to a reference peak with $\mu = 0$ and $\sigma = 1$. ○, Euclidean cosine squared, ×, correlation coefficient squared, ■, first difference correlation coefficient squared.**
**A: Change in the position of the peak ($\mu$); B: Change in the width of the peak ($\sigma$).**

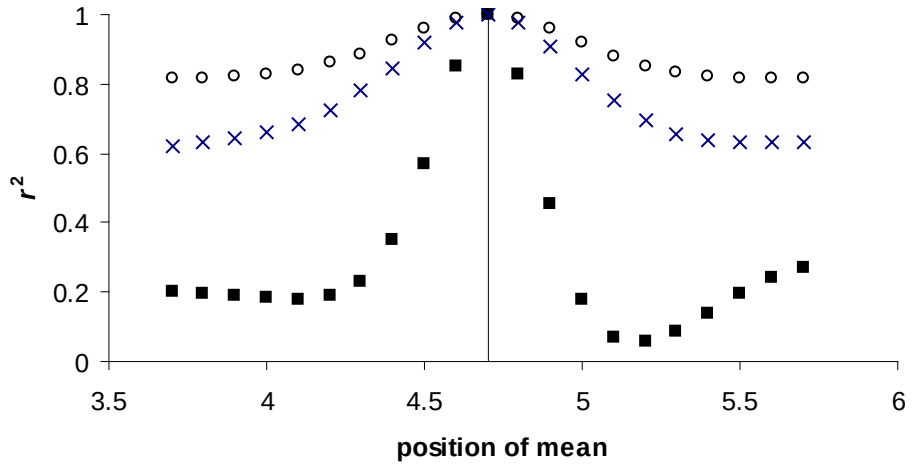
A single-peak spectrum is a simplified case in which there is no significant extent of baseline, and all measurements are informative. From the results shown in Figure 5, it can be seen that the position of the center of a peak strongly influences the similarity index of two spectra. A sharper peak (than the reference peak) can have a low index, but broader peaks limit to an index value that might still indicate a significant match, if they are centered on the reference peak.

As the mean shifts away from the reference peak, the indices eventually turn over. This is because the correlation coefficient (or cosine) goes through zero and then becomes negative, and when squared now appears as a positive index. For a symmetrical single peak the first difference has a mean of zero, with equal positive and negative components. This leads to the

equality of correlation coefficient and Euclidean cosine (see definitions in Table 1, with $\overline{\Delta A_1}$ = $\overline{\Delta A_2}$ = 0). The correlation coefficient is mean centered, and the cosine is not, so if the mean is zero, they will be the same.

### 4.3 Changing a single peak in a ten-peak spectrum

A middle peak of the simulated ten peak spectrum (Figure 1) was chosen as an example to show the influence of the change of a peak in a more realistic spectrum on the distribution of similarity indices. The target peak is overlapped with another peak and is in a unsymmetrical environment having another peak close to it's right. The similarity indices were calculated between the whole reference spectrum of Figure 1 and the spectrum with the changed peak. The effects of changing the position of the peak at constant width, and the width of the peak at constant position, are shown in Figure 6.




A

B

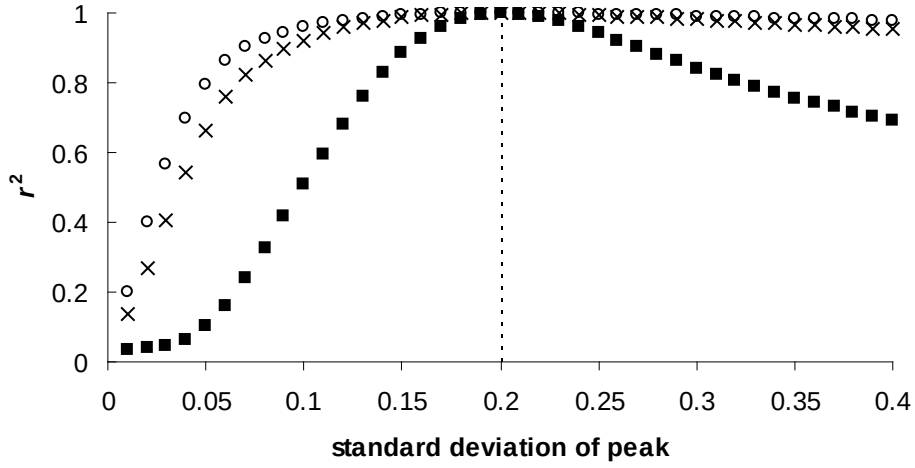
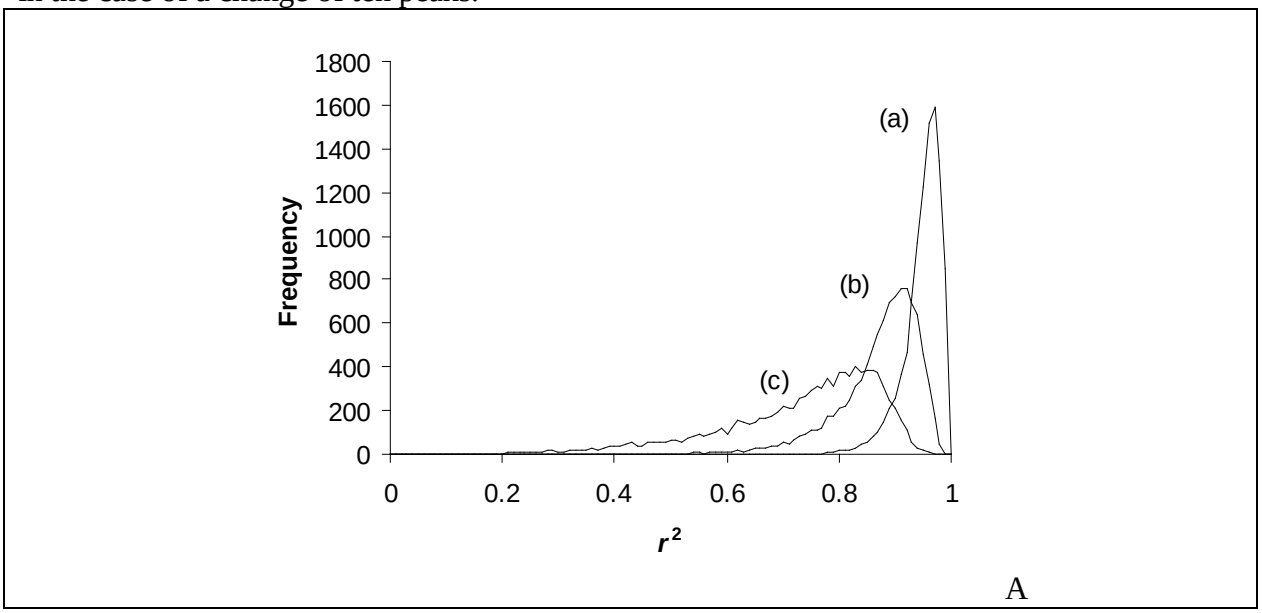


**Figure 6. Values of similarity indices as the parameters of a Gaussian peak with base values $\lambda = 4.7$, $\sigma = 0.2$ are changed in a ten peak spectrum. ○, Euclidean cosine squared, ×, correlation coefficient squared, ■, first difference correlation coefficient squared.**
**A: Change in the position of the peak ($\lambda$); B: Change in the width of the peak ($\sigma$).**

As with a single peak spectrum, the effect of changing the position of a peak is the greatest, but when embedded in a multi-peak spectrum, the effect is diluted by the presence of the other peaks. As a peak is moved away from the position of the reference peak, it can come within the domain of other peaks to which it can correlate. The ability of the difference methods to subtract the effects of the baseline is shown in Figure 6, where the absolute methods give a limiting index much greater than that of the difference methods. As with a single peak, the first difference correlation coefficient is very similar (but no longer identical) in value to that of the first difference cosine.

## 4.4 Changing many peaks in a ten-peak spectra

For changes in many peaks, the influence on a similarity index of the change of any one peak is not easy to identify from a general knowledge of the influence of a change, and it is not possible to analyze the influence of a given parameter of a single peak on the spectral similarity. The distributions of similarity indexes are not normal and although they are different from method to method the same pattern is found throughout. Figure 7 shows the distributions for changing two, five and all ten peaks at random as described above, for the squared Euclidean cosine and first difference squared correlation coefficient, as examples. As more peaks differ between the spectra the distribution moves from high index values to lower index values. Squared Euclidean cosine gives the highest similarities, and the smallest changes, in all four investigated situations. Even for 10 peaks changed at the same time, no squared cosine is 0 and the distribution has a maximum at 0.8. In contrast to the squared Euclidean cosine, the squared first difference correlation coefficient is the most sensitive to any change of peak. In the case of the change of two peaks it has some zeros and no value is greater than 0.7 in the case of a change of ten peaks.



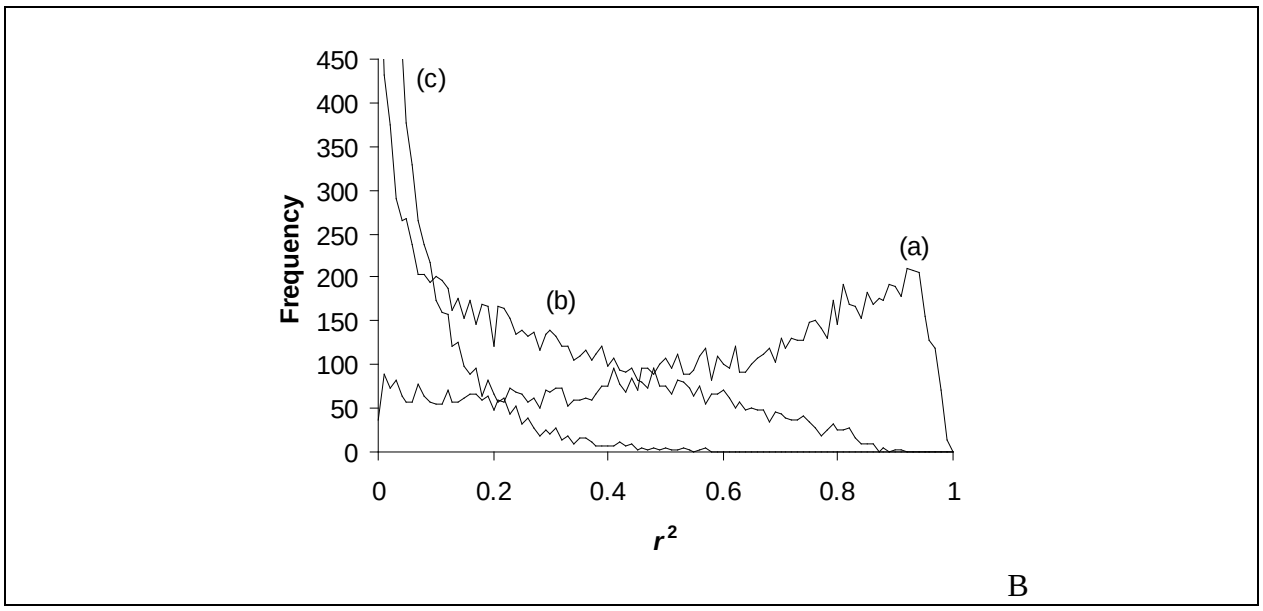



**Figure 7. Distributions of similarity indices with 10,000 random changes in peak parameters in a ten-peak simulated spectrum. (a) Changing two peaks, (b) changing five peaks, (c) changing ten peaks. A: Squared Euclidean cosine; B: Squared first difference correlation coefficient.**

## 4.5 FFT simulation based on real spectra

One hundred simulated spectra and their ten template real spectra are shown in Figure 8. The simulated spectra keep very well not only the general profile of the parent spectra but also show appropriate noise.

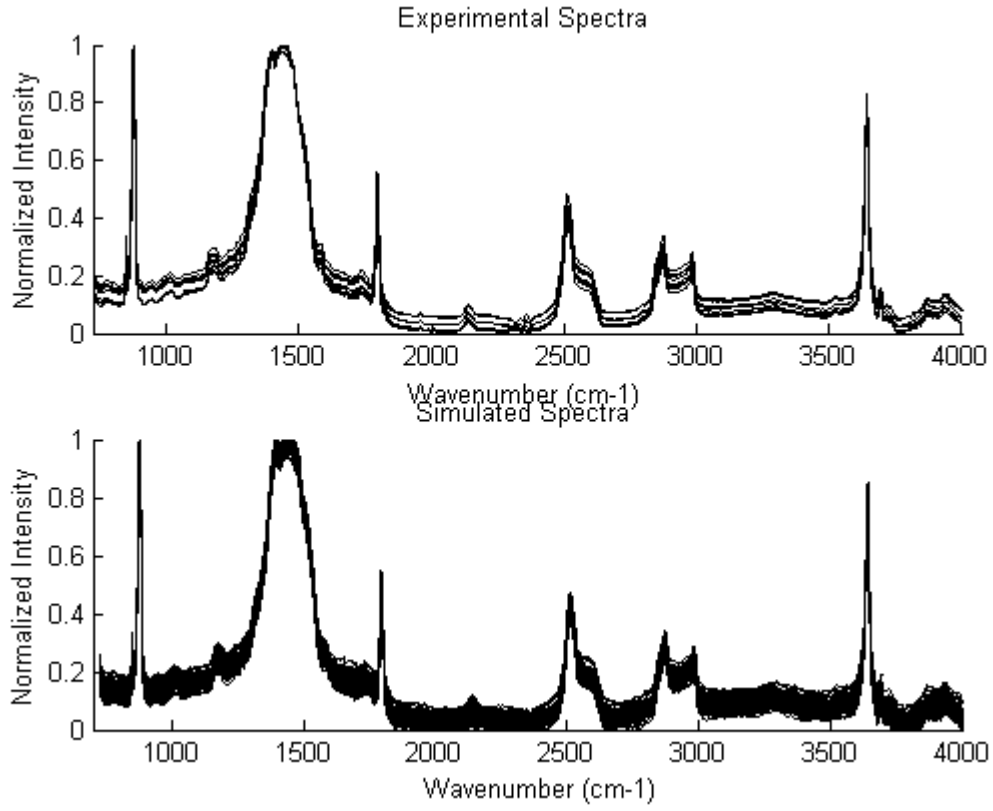



**Figure 8. 100 simulated and 10 template experimental spectra of $Ca_2CO_3$.**

We now consider the pair wise matching of these 110 spectra. The distributions in Figure 9 represent the expected distributions of genuinely matching spectra. The distributions are more narrow than the simulations in which greater changes were made. However the order of the maxima of the distributions does follow the simulations described above, with Euclidean cosine being the least affected by change (and hence giving values near one), and the squared first difference correlation coefficient giving the greatest spread of values, being the index most sensitive to change.

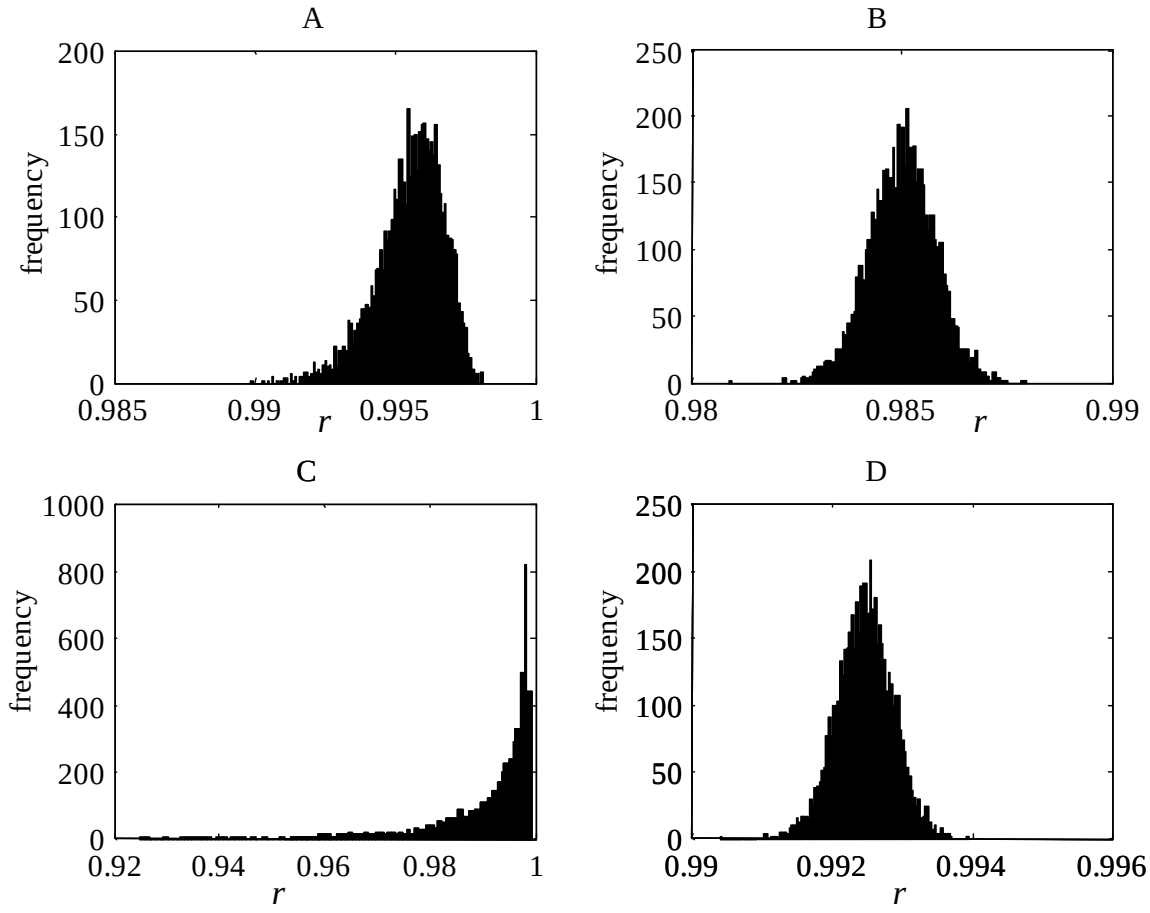



**Figure 9. Distributions of the similarity indices for pairwise comparisons among the simulated and template spectra of $Ca_2CO_3$. A: Squared correlation coefficient, B: squared first difference correlation coefficient, C: squared Euclidean cosine, D: squared first difference Euclidean cosine**

## 4.6 General discussion

The reason for using spectral matching is that similar structures have similar spectra and vice versa, that observed similar spectra imply similar structures. This assumption is not always true. Besides that, not all structural information hidden in a spectrum can be retrieved by a search system or an analyst. Whether the similarity index is sensitive and reliable is another big issue for a successful spectral comparison. A similarity index depends on the capacity (volume) of the information to be compared and the differences in this information space. When the information space is small, the relative difference can be great, as seen in the single peak simulation. Point-to-point comparison methods do not use any chemical information and therefore the difference revealed by the methods might not give the expected conclusion based on chemical knowledge. When the information space is small, the numerical result follows the chemical difference and drawbacks of the method are hidden. When the information space is

large and there are great differences across the entire spectrum, the deviation is also obvious. However in our study, the squared Euclidean cosine still has relative high values when even 50% of the peaks change at the same time (see Figure 7A).
Varmuza et al. [20] have investigated the relationship between IR spectral similarity (as measured by indices such as those used here) and structural similarity, measured by a Tanimoto index of 1365 sub-structure elements, applied to 13,484 compounds in a database. It is interesting that their distributions of the structural similarity index for pairs of compounds that have high spectral similarity, and for randomly chosen pairs, mirror those observed by us for similarity coefficients of spectra of same and different oils [13], and the results here of distributions with increasing changes of peaks. The work by Varmuza and that presented here highlights the problem faced when trying to establish whether two chemical samples are similar (or the same). Even if they are pure compounds the trail that proceeds: *chemical structures* $\rightarrow$ *spectra* $\rightarrow$ *similarity index* has the potential loss of information going from one term to the next. The measures investigated here do not vary linearly with the changes in the underlying spectra, and the spectra do not change linearly with chemical structure. We must therefore choose the most appropriate spectroscopy that accurately reflects changes of chemical structure for the system under investigation, and then again choose the most sensitive index to allow the correct inference to be made concerning the structures. Although not a part of this investigation, we note that the choice of infrared method, for example between FTIR and ATR, or use of absorbance spectra and transmittance spectra, will have an effect on the similarity indices and discrimination. Absorbance spectra may be preferred because of the linear relation with concentration through Beer's Law.
In forensic analysis for which no database can be used, a local comparison is needed if accuracy is important. Even the average of similarity indexes, a biased estimation of the similarity between spectra, is better than the single number obtained from the whole range of the spectra. This has been demonstrated by our work of the comparison oil spill with suspects using FTIR. Maximal common substructure is another way to screening the spectral searching results [21,22].

## 5. Conclusions

Not all structural information can be retrieved using one analytical technique and not all spectral information can be acquired by a spectral interpretation method. If the contribution of a component is small, it is difficult to identify it from the peak in a spectrum of a mixture. Squared correlation coefficient and squared cosine can lead to false positive results while squared first difference correlation coefficient and squared first difference cosine have false negative results. Based on these realizations, we recommend testing in windows of a spectrum, i.e. by employing regional comparison when comparing spectra for structural elucidation, thus focusing attention on regions that are changing. Using different methods, a balance between accurate match and tolerance of the noise factor, e.g. weathering, small contamination, etc. can be achieved.

## Acknowledgements

This work was supported by a grant from the New South Wales EPA Trust.

## References

1. Laloum, E., N. Q. Dao, Daudon, M. (1998). Applied Spectroscopy **52**(9): 1210-1221.
2. Chen, C.-S., Y. Li, Brown, C. W. (1997). Vibrational Spectroscopy **14**(1): 9-17.
3 Papazova, D., Pavlova, A., Kovacheva, K., (1998) Application of instrumental analytical methods for oil spill identification, Analytical Laboratory, 7, 201-206.
4 Behr, A., Brehme, V. A., Ewers, C. L. J., Gron, H., Kimmel, T., Kuppers, S., Symietz, I., (2004) New developments in chemical engineering for the production of drug substances, Engineering in Life Sciences, 415-24.
5. Stein, S. E. (1995). "Chemical Substructure Identification by Mass-Spectral Library Searching." Journal of the American Society for Mass Spectrometry **6**(8): 644-655.
6. Li, J., S. Fuller, et al. (2004). "Matching fluorescence spectra of oil spills with spectra from suspect sources." Analytica Chimica Acta **514**(1): 51-56.
7. Chau, F.-T., D. K.-W. Mok, et al. (2001). "Fingerprinting analysis of raw herb: application of chemometrics techniques for finding out chemical fingerprint of Chinese herb." Analytical Sciences **17**(Suppl.): a419-a422.
8. ASTM, D 3414-98 Standard test method for comparison of waterborne petroleum oils by Infrared spectroscopy. (1998) Philadelphia, American Society for Testing and Materials.
9. Murphy, B. and R. D. Morrison (2001). Introduction to Environmental Forensics. Amsterdam, Academic Press. p137
10. Praisler, M., J. Van Bocxlaer, et al. (2002). "Chemometric detection of thermally degraded samples in the analysis of drugs of abuse with gas chromatography-Fourier-transform infrared spectroscopy." Journal of Chromatography, A **962**(1-2): 161-173.
11. Lavine, B. K., C. E. Davidson, et al. (2004). "Spectral Pattern Recognition Using Self-Organizing MAPS." Journal of Chemical Information and Computer Sciences **44**(3): 1056-1064.
12. Liebich, V. and G. Ehrlich (1989). "Multivariate comparison of concentration profiles in materials analysis." Mikrochimica Acta **2**(1-3): 39-48.
13. Li, J., D. B. Hibbert, Fuller, S., Cattle, J., Pang Way, C. (2005). Comparison of spectra using a Bayesian approach. An argument using oil spills as an example, Analytical Chemistry,**77**, 639-644.
14. Bellamy, L. J. (1975). Advances in Infrared group frequencies. London, Chapman and Hall.
15. Schrader, B. (1995). Infrared and Raman spectroscopy :methods and applications. New York, VCH.
16. P. Willett, (1987) Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, UK,.
17. Fisher, R. A. (1970). Statistical methods for research workers. Edinburgh, Oliver and Boyd.
18. Anderson, T. W. (1984). An introduction to multivariate statistical analysis / 2nd ed. New York :. New York, John Wiley & Sons, Inc.
19. Muirhead, R. J. (1982). Aspects of multivariate statistical theory. New York, John Wiley & Sons, Inc.
20. Varmuza, K., Karlovits, M., Demuth, W. (2003). "Spectral similarity versus structural similarity: infrared spectroscopy." Analytica Chimica Acta **490**(1-2): 313-324.
21. Wang, T., J. Zhou, J. (1997). EMCSS: A new method for maximal common substructure search, Journal of Chemical Information and Computer Sciences, 37828-834.
22. Chen, L. G., Robien, W. (1994) Application of the Maximal Common Substructure Algorithm to Automatic Interpretation of C-13-NMR Spectra, Journal of Chemical Information and Computer Sciences, 34, 934-941.