

(21) Application No 8925772.9

(22) Date of filing 15.11.1989

(30) Priority data
(31) 8826816 (32) 16.11.1988 (33) GB

(71) Applicant
United Kingdom Atomic Energy Authority
(Incorporated in the United Kingdom)
11 Charles II Street, London, SW1Y 4QP,
United Kingdom

(72) Inventors
A T Chadwick
C G Windsor

(74) Agent and/or Address for Service
Paul Austin Wood
Patents Branch, Building 329, Harwell Laboratory,
Oxfordshire, OX11 0RA, United Kingdom

(51) INT CL⁴
G06F 7/04 15/336

(52) UK CL (Edition J)
G4A ACL
U1S S2135 S2137 S2140 S2142

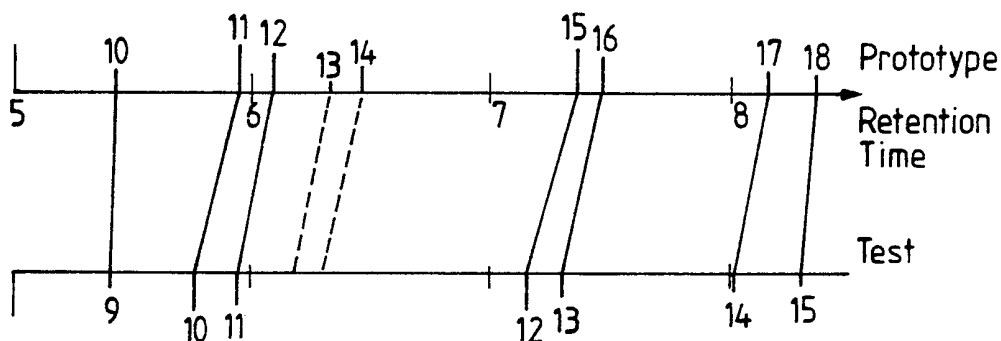
(56) Documents cited
GB 1431207 A US 4008388 A

(58) Field of search
UK CL (Edition J) G4A ACL ACX AUB AUX
INT CL⁴ G06F
Online database: WPI

(54) Method for spectrum matching

(57) A method for matching one spectrum to another, when one spectrum has smoothly varying positional and amplitude instrumental distortions which are superimposed on amplitude differences which represent intersample variations, comprises creating a one-to-one mapping (as shown) between peaks in the two spectra which maximises the correlations between the local structures around the peaks and simultaneously minimises the curvature of the distortions.

Fig. 3.



1/2

Fig.1.

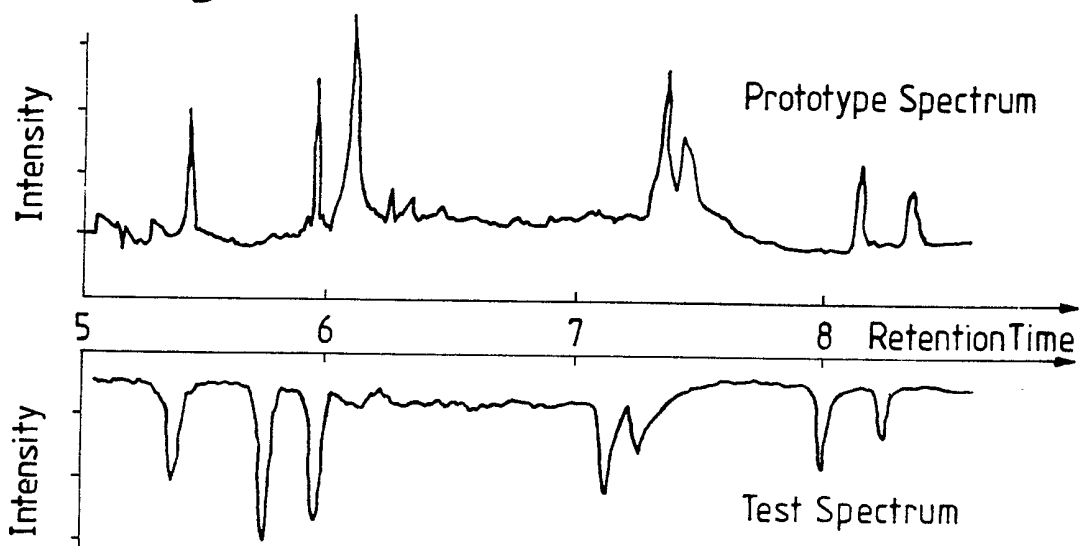


Fig.2.

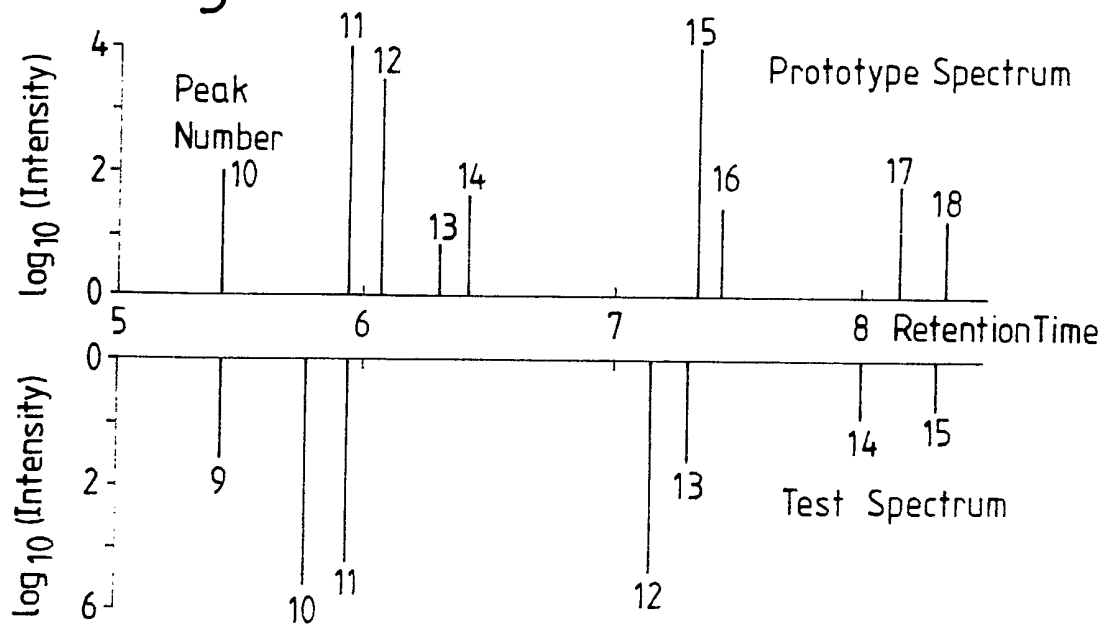
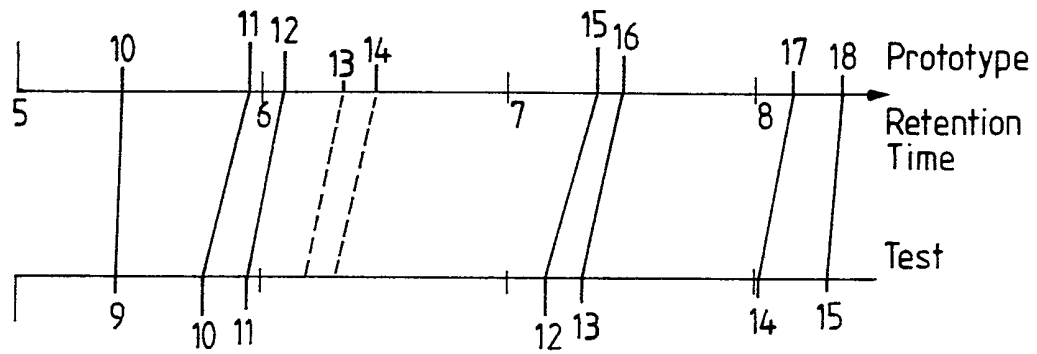
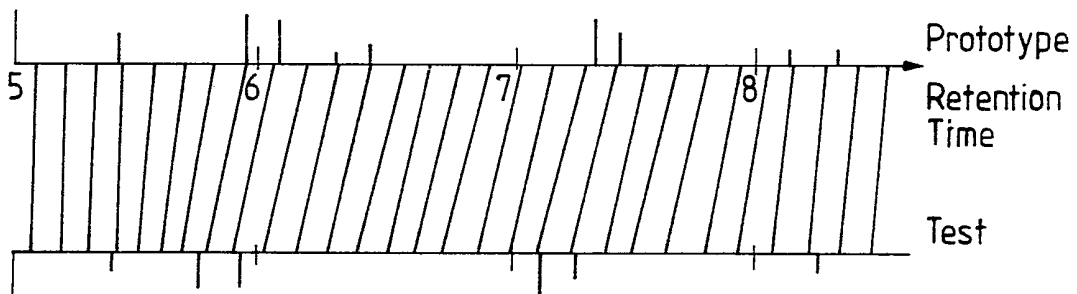


Fig.3.*Fig.4.*

Method for Spectrum Matching

Many modern analytical techniques provide an output in which molecules or molecular components of different types ("chemical components") give a signal at characteristic values of one or more continuous variables, the "analytical space". For example, in chromatography the continuous variable is a retention time determined in part by the differential partition of molecules between a mobile and a stationary phase. In electrophoresis the continuous variable is a distance migrated in unit time, determined by the relation between molecular charge, size and interaction with the stationary phase. In spectroscopy the continuous variable is usually a wavelength but may also be a magnetic field strength.

The signal usually consists of a peak whose appearance is determined by the extent of spreading in the analytical space, and may be asymmetric ("tailed"). Conventional peak fitting algorithms are available for finding a discrete position for such a peak in the analytical space together with a "peak size", usually the area but sometimes the height, which is approximately proportional to the amount of a chemical component. The present invention is concerned with assignment of peaks in chemical analysis data, which is required to identify chemical components and if required find their relative amounts. Even where individual peaks cannot be assigned to any known standard, it is possible through the present invention to correlate results from two or more samples in such a way that the relative sizes and positions of peaks due to chemical components in the samples may be used as a means of identification or classification of samples, or as a means of determining the similarity of samples. Samples may for example be biological material or forensic specimens such as traces of explosives, or artefacts whose place or means of manufacture is to be determined by chemical methods.

The chemical component which is determined depends on the analytical technique used and may for example be an entire molecule or molecular complex (in chromatography and electrophoresis), an atom (in X-ray photoelectron spectroscopy), a functional group within a molecule (eg the first approximation of results from infra-red spectroscopy, or mass spectroscopy), or a combination of properties of constituent atoms together with their arrangement in the molecule (eg nuclear magnetic resonance - nmr). Analytical techniques considered here also include those in which information is obtained on the environment of a molecule (eg fluorescence relaxation times, or nmr in polymers) and those in which the unknown is not analysed directly but instead subjected to chemical procedures in such a way that the chemical component characterised by the analytical apparatus is a secondary measure of the required information - for example a gene probe may be used to give chemical information on DNA.

In addition to analytical methods where the data output has literal peaks, we are concerned also with methods where such a peak position and size may be obtained by the mathematical transformation of primary data which has a different functional form, which may for example be a continuously increasing or decreasing function, as in photoelectron spectroscopy or electron energy loss spectroscopy.

The analytical space may have more than one dimension, for example in crossed or 2-dimensional electrophoresis methods, or in chromatography when the eluant from one column is sampled and run through another column under different conditions. In such cases the separation is multi-dimensional but a given chemical component will give

only one peak at some position in the full analytical space. The present invention may be applied directly to an analytical space of arbitrary dimension.

5 This is distinguished from multiple-column chromatography in which duplicate samples may be run separately under different conditions with more than one type of solvent or stationary phase. As in 2-D electrophoresis, results are represented on a 2-dimensional
10 plot. However a given chemical component may give zero output under one or more of the different conditions. Where a component gives a peak under every condition then the various peak sizes for a given chemical component have to be combined by some means into a single number. The
15 present invention could be used as a means for correlating peaks between the different analytical conditions.

 The method of the present invention also is applicable to analytical techniques which give more than
20 one peak from a given chemical component, either in one analytical variable or by use of multiple methods which give simultaneous responses from a single chemical component. As an example of multiple peaks in one analytical variable, in spectroscopy a given functional
25 group gives absorption at more than one characteristic frequency. Examples of techniques giving multiple responses are multi-detector chromatography and gas chromatography/mass spectroscopy (GC/MS). In each of these techniques, molecules eluting at different times each give
30 a spectrum which is scanned. Frequently this spectrum is sampled by two or more detectors each sensitive to a different region of wavelength, which provide multiple channels for determining peak size. In such cases obvious modifications of the method of the present invention will
35 apply; for example, in spectroscopy when the peak separation is uncertain the method can be applied to each

part of a multiple peak separately and the results averaged or decided by a voting method over all the peaks to determine the best correlation of multiple peaks between samples; when the peak separation is known much better than the overall position of the peaks then a search can be made for a set of peaks at fixed separations, using any prior information on the ratios of peak sizes for that chemical component. In multi-detector chromatography and GC/MS there is a primary variable, the chromatographic retention time, and a secondary variable, the adsorbance at several frequencies or the yield of fragments at several mass numbers. In multi-frequency magnetic resonance techniques (eg ENDOR, Overhauser effect) both the primary and the secondary variables are ratios of frequencies and magnetic fields. The method of the present invention could be applied to either the primary or the secondary variable or to both simultaneously. Here a further modification of the method would use functions of peak sizes over a number of secondary analytical variables as a criterion for assigning chemical components to given values of the primary variable. For analytical methods of the type discussed in this paragraph, in the subsequent discussion the term "peak size" and "peak position" should also be understood to apply to information from more than one peak.

An ideal analytical technique would always give the same peak size at the same position in analytical space for a given component in identical samples. If standards of known amount are available then peak assignment becomes a trivial task. Many analytical procedures are subject to drift both in peak position and sensitivity. A skilled operator can assign peaks correctly in the presence of non-linear drift, even without standards, by recognising the patterns of peak sizes which are conserved between different samples of the same basic type - for example biological specimens from animals of the same species.

Automated analytical apparatus is unable to do this well at present. A method is required for use in or with analytical apparatus, which can correct for non-linear drift. If convenient to use, such a method can save the labour of the preparation of multiple or indeed any standards.

Standard methods are available which allow overlapping peaks to be expressed in terms of a mean position and a size. The use of statistical moments (Grushka, 1971) allows peaks to be separated even when closer than their half-widths, provided that either the peak widths or the skewness are sufficiently different for the overlapping peaks. The maximum entropy method may also be used to provide an unbiased representation of the information present in the original data, in a form more suitable for automated peak matching.

The conventional approach to peak identification in analytical processes involves the use of standards. The following discussion makes special reference to chromatography but the same methods apply to peak identification in spectroscopy.

A single standard is adequate if instrumental drift in the analytical space (here retention time) is linear and if the position of zero (the injection time) is known precisely. It is preferable to use more than one standard and to normalise intermediate peak positions. Ayers and Sanford (1969) suggest programming of the recorder chart speed such that the intervals between standards are the same in all specimens. Chilcote (1974) points out that relative retention times calculated between pairs of identical samples, under nominally identical analytical conditions, are often a non-linear function of the absolute retention time and that standards ideally should be chosen

to give the best approximation of this curve in a piece-wise linear fashion. He also points out that for a single-channel instrument, the standards themselves may be confused with peaks in a complex sample. It is therefore
5 desirable to minimise the number of standards as far as practicable. According to Chilcote, the eye is able to match patterns on the basis of patterns of intensity or of peak shape, and is therefore superior to the instrumental methods previously known. For hydrocarbons the retention
10 times vary logarithmically with carbon number and this may also be used as a basis for interpolation, using the "Kovat's retention index". Caddy, Fish and Scott (1973) demonstrate this approach in drug analysis where results from both polar and non-polar stationary phases are used as
15 a basis for comparison with standard drugs and their metabolites.

If the order of peak sizes is well conserved between samples then this can be used as a basis for identification
20 (Issaq and McNitt, 1982). However, the method breaks down when the order of intensities is reversed for some of the chemical components. This may be due either to changes in analytical conditions or to natural variations between samples, or both.

25

Other published methods are based on the expected positions of peaks, make some allowance for drift, and may include decision criteria based on the size of the peak to be identified relative to the mean, but do not take into
30 account the relative sizes of neighbouring peaks. Follain (1973) suggests that the relative error observed between the expected and observed retention times for a given constituent of a mixture be applied in the identification of the following constituent. This method will encounter
35 difficulties when peaks are absent or extra peaks are observed as the peak search will "lose its way". Boege

(1974) claims a method in which the tolerance range for identifying a peak as a known substance (based on a standard) is based on the width of the peak in the unknown sample.

5

According to the invention there is provided a method for matching one spectrum to another when one spectrum has smoothly varying positional and amplitude instrumental distortions which are superimposed on amplitude differences which represent intersample variations, comprising the operations of creating a one to one mapping between peaks in the two spectra which maximises the correlations between the local structures around the peaks and simultaneously minimises the curvature of the distortions.

15

Preferably the method also includes the operation of making allowance for missing or extra peaks in one of the spectra.

20

For the purposes of the present specification, the term "spectrum" includes any form of data representation in which successive values of a parameter are presented sequentially. Thus, by way of example, in addition to a conventional spectrum, the term includes a trace from a chart recorder and the output from a chromatographic column, or the data forms discussed in the introduction to this specification.

25

The invention may be used as a final stage in the matching of spectra which have processed by some other method. It may also be used in the searching of databases from chemical analysers.

30

The invention will now be described, by way of example, with reference to the accompanying drawings, in which:

35

Fig. 1 shows examples of two gas chromatograms

Fig. 2 shows the peak positions and relative intensities of the two spectra evaluated by a conventional data processing method

5 Fig. 3 shows a mapping between the spectra as shown in Figure 2, and

Fig. 4 shows an example of the distortion vector connecting the spectra shown in Figure 2.

10

Figure 1 shows examples of two gas chromatograms, referred to as spectra, although they could as well be other kinds of analytical result. Each peak corresponds to a certain chemical component, and the objective is to compare the intensities of corresponding peaks in the two spectra. Figure 2 shows the peak positions and relative intensities evaluated using a conventional data processing method. It is seen immediately that the spatial and amplitude distortions between the two spectra make the association a non-trivial task. The fact that not all peaks are located, depending on the signal to noise ratio and resolution of each spectrum, complicates the process. The invention seeks to make the association, as does the eye, by comparing the local structure around particular peaks in one spectrum with that around likely peaks in the other spectrum. A search is made for the pair of peaks having the best matching local structure. In mapping the set of peaks from the prototype spectrum onto the test spectrum, a distortion of the analytical variable, the retention time in this case, must be introduced. An arbitrary distortion will fit any set of peaks. The invention provides a method for constraining the search for the best match between prototype and test spectra, such that the distortions vary smoothly along the spectrum. If the distortion is likened to an elastic sheet connecting corresponding points on the two spectra, then the "elastic energy" in the sheet is minimised as far as possible consistent with obtaining the maximum correlation.

15
20
25
30
35

The method explicitly defines a connection vector which maps each peak on the prototype spectrum with one on the test spectrum. Such a mapping is illustrated in figure 3. It will not always be possible to find an associated
5 peak so that the connection vector may have to end on no peak, or may end on two peaks. An additional "connection energy" term is associated with such non-ideal connections. This may either be an adjustable parameter or can be calculated in proportion to peak amplitudes taking a
10 missing peak as equal in size to the integration threshold, and taking the sum of peak amplitudes for a doublet. The method proceeds iteratively towards the final mapping. The initial set of connections is not critical to the method and may be chosen on the basis of associating each peak on
15 the prototype spectrum with the peak on the test spectrum closest in analytical space to the value given by linear interpolation between standard sample peaks.

The method introduces up to two variables explicitly
20 defining the spatial and optionally, the amplitude distortions of the test spectrum with respect to the prototype spectrum. The spatial distortion is evaluated by averaging the shifts in analytical space between peaks in a region of the prototype spectrum and those in the test
25 spectrum mapped by the spatial connection vector. The size of the region used to obtain this average is a parameter of the method, and must be appropriate to the scale of the changes in the distortion vector. Similarly the amplitude distortion vector which is an optional feature of the
30 method, is the ratio of the amplitudes of peaks in the prototype spectrum with connected peaks in the test spectrum, averaged over the same region. By moving the centre of the averaging region along the prototype spectrum, the two distortion functions may be sampled
35 along the prototype spectrum. Figure 4 shows an example of the distortion vector connecting the spectra shown in figure 2.

The method needs to evaluate the correlation between regions of the two spectra close to pairs of peaks joined by the current connection vector. This may be done in several ways. The present method, applied to continuous spectra as in figure 1, evaluates the sum of the squared difference between amplitudes of points on the prototype and test spectra over the regions of the spectra spanning the chosen peaks being tested, according to the spatial distortion vector. The length of this region, which should be enough to include a few peaks on either side of each of the chosen pair of peaks, is an important parameter of the method. If the spectra have been processed to give peak positions and intensities only, as in figure 2, then the squared differences between the two peak areas are evaluated for all overlapping pairs of peaks in the chosen region. Peaks are defined to overlap if the calculated position of a peak in the test spectrum, calculated from its position in the prototype spectrum, lies within a defined width of its actual value. The width is related to the actual peak width and to the random fluctuating errors on the peak position determination. In practice it is a parameter of the method. If peak width is an important cue to identification then correlation of peak width may also be taken into account.

The method leads iteratively from the initial connection vector to the optimum matching using a Monte Carlo procedure. A peak on the prototype spectrum is chosen at random, and an alternative allowed connection chosen. The change of the connection vector is allowed or not depending on the change in the "energy". In this case the energy has four positive terms whose sum is minimised over the region of the spectrum for which peaks are to be assigned:

- i) an energy depending on the squares of the difference between the regions being correlated.
- 5 ii) an energy depending on the squared difference between the actual spatial displacement between the connected peaks and the averaged distortion vector at that point.
- 10 iii) an energy depending similarly on the squared difference between the actual peak intensity ratio and the averaged value.
- 15 iv) an energy depending on the numbers of connection vectors ending on each peak in the test spectrum. This energy is zero for a single connection but positive for a zero or multiple connection. This term may be included in (i) by appropriate definition.

20

The relative magnitudes of each of these four terms are parameters of the method.

The integral of the energy term (i) may be used as a
25 measure of the difference between samples and hence as a basis for classification. If means or medians are to be calculated as class exemplars then terms (ii) and (iii) may be used as inverse weightings showing the analytical error. If a number of spectra are to be correlated then an
30 efficient procedure is to correlate in pairs in order of a proxy for instrumental drift in one of the analytical variables such as retention time for a standard.

Claims

1 A method for matching one spectrum to another when one spectrum has smoothly varying positional and amplitude instrumental distortions which are superimposed on amplitude differences which represent intersample variations, comprising the operations of creating a one to one mapping between peaks in the two spectra which maximises the correlations between the local structures around the peaks and simultaneously minimises the curvature of the distortions.

2 A method according to claim 1 including the operation of defining a connection vector which seeks to map a peak on the spectrum having the positional and amplitude instrumentation distortions with a peak on the other spectrum, determining the spatial distortion between the spectra in the region of selected mapped peaks in the spectra by averaging the shifts in analytical space, as hereinbefore defined, between the selected mapped peaks and repeating the operation iteratively to minimise the said spatial distortions between the spectra in the said regions.

3 A method according to claim 2 including the operation of defining an amplitude distortion vector by determining the ratio of the amplitudes of peaks in the spectrum having the positional and amplitude instrumental distortions with connected peaks in the other spectrum averaged over the same region as that used to determine the spatial distortion, moving the centre of the averaging region along the other spectrum and evaluating the two distortion functions so as to sample the spatial and amplitude distortions along the other spectrum.

4 A method for matching one spectrum to another when one spectrum has smoothly varying positional and amplitude

instrumental distortions which are superimposed on amplitude differences which represent intersample variations substantially as hereinbefore described and with reference to figures 1, 3 and 4 of the accompanying drawings.