A Generalized Expression for the Similarity of Spectra: Application to Powder Diffraction Pattern Classification

R. DE GELDER,[1] R. WEHRENS,[2] J. A. HAGEMAN[2]

[1] Department of Inorganic Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands
[2] Department of Chemometrics in Analytical Chemistry, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Received 13 June 2000; accepted 17 August 2000

ABSTRACT: A generalized expression is given for the similarity of spectra, based on the normalized integral of a weighted crosscorrelation function. It is shown that various similarity and dissimilarity criteria previously described in literature can be written as special cases of this general expression. A new similarity criterion, based on this generalized expression, is introduced. The benefits of this criterion are that it properly recognizes shifted but otherwise similar details in spectra and that the resulting similarity measure is normalized. Moreover, the criterion can easily be adapted to specific properties of spectra resulting from various analytical methods. The new criterion is applied to the classification of a series of crystal structures of cephalosporin complexes, based on comparison of their calculated powder diffraction patterns. The results are compared with those obtained using previously described criteria. © 2001 John Wiley & Sons, Inc. J Comput Chem 22: 273–289, 2001

Keywords: similarity/dissimilarity; structure classification; powder diffraction pattern; correlation function; pattern comparison

Introduction

Many chemical and physical methods for the analysis of compounds in solution and/or the solid state yield one-dimensional spectra or diagrams that consist of isolated and/or (partly) overlapping sharp peaks. Examples of such spectra are 1D-NMR spectra and powder diffraction patterns. In the case of powder diffraction patterns, the positions of the peaks are very sensitive to small deviations in unit cell parameters. This means that in the case of crystal structure prediction, or *ab initio* structure determination, strongly related structures may give (calculated) powder diffraction patterns that look similar from an overall point of view but

Correspondence to: R. de Gelder, Crystallography Laboratory, University of Nijmegen, Toernooiveld 1, 6525 ED Nijmegen, The Netherlands; e-mail: rdg@sci.kun.nl

differ significantly on a more local scale. The same situation may occur for isomorphous compounds that differ only slightly in unit cell volume or unit cell shape. These compounds may give experimental or calculated powder patterns that by visual inspection are definitely related and recognizable as isostructural compounds. Unfortunately, in all those cases the calculation of a reliable and objective measure of similarity or dissimilarity, even when the whole pattern is considered, is nontrivial.

A conventional method for the comparison of powder diffraction patterns is calculating the difference pattern and quantifying the dissimilarity between the patterns as the sum of the differences or the sum of the squared (and weighted) differences. In the field of Rietveld refinement such "criteria of fit" are known as R_p (R-pattern) and R_{wp} (R-weighted pattern).[1] Because these measures are based on a pointwise comparison of patterns, small shifts in peak positions may result in a large (undesired) increase of the dissimilarity measures.

In the present article, it is shown that there is a simple relationship between the conventional criterion based on squared differences, the Pearson product-moment correlation coefficient[2] and the overlap integral described by Lawton and Bartell[3]—who define a measure for the similarity of powder diffraction patterns on an absolute scale—when these criteria are written in terms of the correlation function. With respect to the correlation function, the drawback of these criteria is that they only consider one point (the value at the origin) from the auto- and crosscorrelation functions and neglect the information that is present in the remainder of the auto- and crosscorrelation functions.

Karfunkel, Rohde, Leusen, Gdanitz, and Rihs[4] describe a method, that is based on the work of Stephensen and Binsch,[5] in which they not only compare powder patterns pointwise but compare a point of one diagram with the neighbourhood of the corresponding point on the other diagram and vice versa. As a matter of fact, Stephensen and Binsch transformed the conventional squared difference criterion into a criterion based on correlation functions, by introducing this neighborhood concept in the comparison of patterns. This can easily be shown by rewriting their so-called "fold" in terms of auto and crosscorrelation integrals. A drawback of their criterion is that the resulting values are not on an absolute scale so that limits for acceptance can not easily be defined.

Although in principle all the information on the similarity of two patterns is contained in the crosscorrelation function, an additional function must be used to extract this similarity information. Such a function defines the effective neighbourhood and related weigths in the target pattern that should be compared with the corresponding point on the reference pattern. The "fold" used by Karfunkel et al. contains such a function in the form of matrix $\mathbf{F}$; however, it will be shown that a more convenient function can be used for this purpose. This alternative function is easier to "tune" with respect to the particular properties of the spectra of interest, because it contains only one adjustable parameter.

Because powder diffraction patterns, and also 1D-spectra from many other analytical methods, are on an arbitrary scale, an obvious question is how to scale the patterns before applying a (dis)similarity criterion. Karfunkel et al. scale the patterns by equalizing the total number of counts, which is the same as normalizing the area under the patterns. This choice was not based on specific arguments, although it is clear that for closely related structures the patterns should have a similar number of counts. In this article it is shown that on the basis of crosscorrelation and autocorrelation integrals you should scale the patterns according to the self-similarities of the patterns. In practice, this leads to almost the same scaling as proposed by Karfunkel et al. However, in principle, scaling according to self-similarities may lead to a different sum of counts for each individual pattern.

The similarity and dissimilarity criteria described above can be written as a special form of a generalized expression for similarity, based on normalized weighted auto and crosscorrelation functions. Using this generalized expression it can easily be seen that the various criteria only use a different weighting function and/or different normalization factors. One of the major advantages of this generalized form is that it shows how to obtain a similarity measure on an absolute scale, given a suitable weighting function. Another advantage is that prescaling of the patterns becomes unnecessary.

Although the power of a newly proposed similarity criterion that is based on this generalized expression is demonstrated for powder diffraction patterns corresponding to a series of crystal structures of cephalosporin complexes, the similarity concept presented in this article provides a general method for quantifying the match between spectra of various physical and chemical techniques for the analysis of matter. The applicability of the generalized expression for similarity lies in the field of pattern classification, data base searching and optimization problems. Its recent successful use in the direct determination of molecular constants from rovibronic spectra with genetic algorithms is described by Hageman, Wehrens, De Gelder, Meerts, and Buydens.[6]

Auto- and Crosscorrelation Functions

A function that describes the similarity (or overlap) of two patterns, which are here expressed as two continuous functions $f(x)$ and $g(x)$, as a function of the relative shift r between the patterns, is the correlation function. The maximum and minimum value of the relative shift r is determined by the interval for which the patterns $f(x)$ and $g(x)$ are measured or calculated. The autocorrelation function $c_{ff}(r)$ for a given reference pattern $f(x)$ is given by:

$$c_{ff}(r) = \int f(x)f(x+r)\,dx \tag{1a}$$

The integral of $c_{ff}(r)$, the autocorrelation integral, is given by (see Appendix A.1):

$$\int c_{ff}(r)\,dr = \left(\int f(x)\,dx\right)^2 \tag{1b}$$

This shows that the area under the autocorrelation function $c_{ff}(r)$ is equal to the square of the area under the reference pattern $f(x)$.

Similar expressions can be defined for the autocorrelation function $c_{gg}(r)$ of target pattern $g(x)$:

$$c_{gg}(r) = \int g(x)g(x+r)\,dx \tag{2a}$$

$$\int c_{gg}(r)\,dr = \left(\int g(x)\,dx\right)^2 \tag{2b}$$

It can easily be seen from expressions (1b) and (2b) that setting the total number of counts for $f(x)$ and $g(x)$ to the same value also results in setting the autocorrelation integrals for $f(x)$ and $g(x)$, which express the self-similarities of the patterns, to the same value. In this way it is possible to put the autocorrelation functions for $f(x)$ and $g(x)$ on an absolute scale (in principle, arbitrarily chosen) and compare the values of the crosscorrelation function $c_{fg}(r)$ with the values of the autocorrelation functions $c_{ff}(r)$ and $c_{gg}(r)$.

The crosscorrelation function $c_{fg}(r)$ for patterns $f(x)$ and $g(x)$ is defined in a similar way by:

$$c_{fg}(r) = \int f(x)g(x+r)\,dx \tag{3a}$$

$$\int c_{fg}(r)\,dr = \int f(x)\,dx \int g(x)\,dx \tag{3b}$$

From (3b) it can be seen that the area under the crosscorrelation function $c_{fg}(r)$, the crosscorrelation integral, is always equal to the product of the areas under the patterns $f(x)$ and $g(x)$. The crosscorrelation function $c_{fg}(r)$ can be normalized by dividing it by the root of the product of the autocorrelation integrals [the product of the areas under the patterns $f(x)$ and $g(x)$], which makes prescaling of the patterns $f(x)$ and $g(x)$ unnecessary:

$$c'_{fg}(r) = c_{fg}(r) \Big/ \left(\int f(x)\,dx \int g(x)\,dx\right) \tag{3c}$$

The integral of $c'_{fg}(r)$ will always be equal to 1. This means, however, that the crosscorrelation integral itself is *not* a measure for the similarity between $f(x)$ and $g(x)$. It is the shape of the correlation function $c_{fg}(r)$ [or $c'_{fg}(r)$] that contains the information on the similarity between patterns $f(x)$ and $g(x)$. In Figure 1, two different powder diffraction patterns are shown (corresponding to entries 2 and 20 of Table II, which will be explained later) with their corresponding (normalized) auto- and crosscorrelation functions. The areas under the correlation functions are the same; however, the different shapes of the curves clearly reflect the differences and dissimilarity of the patterns.

In the next sections it will be shown that various similarity and dissimilarity criteria that are decribed in the literature can be written in terms of auto- and crosscorrelation functions.

Pointwise Similarity and Dissimilarity Criteria

A criterion often used for expressing the dissimilarity between two spectra or diagrams is the conventional pointwise criterion that includes the sum of the squared differences (see, e.g., Harris, Johnston, and Kariuki[7] or Dods, Gruner, and Brumer[8]). The difference criterion d_{fg} is given by:

$$d_{fg} = \int \left(f(x) - g(x)\right)^2 dx \tag{4a}$$

The difference criterion d_{fg} can be rewritten into an expression that only includes terms that are identical to the autocorrelation functions for $f(x)$ and $g(x)$ and the crosscorrelation function for $f(x)$ and $g(x)$ at $r = 0$ (see Appendix A.2):

$$d_{fg} = c_{ff}(0) + c_{gg}(0) - 2c_{fg}(0) \tag{4b}$$

Therefore, d_{fg} is only based on the values of the auto and crosscorrelation functions at $r = 0$ (no relative shift between the patterns is taken into account).

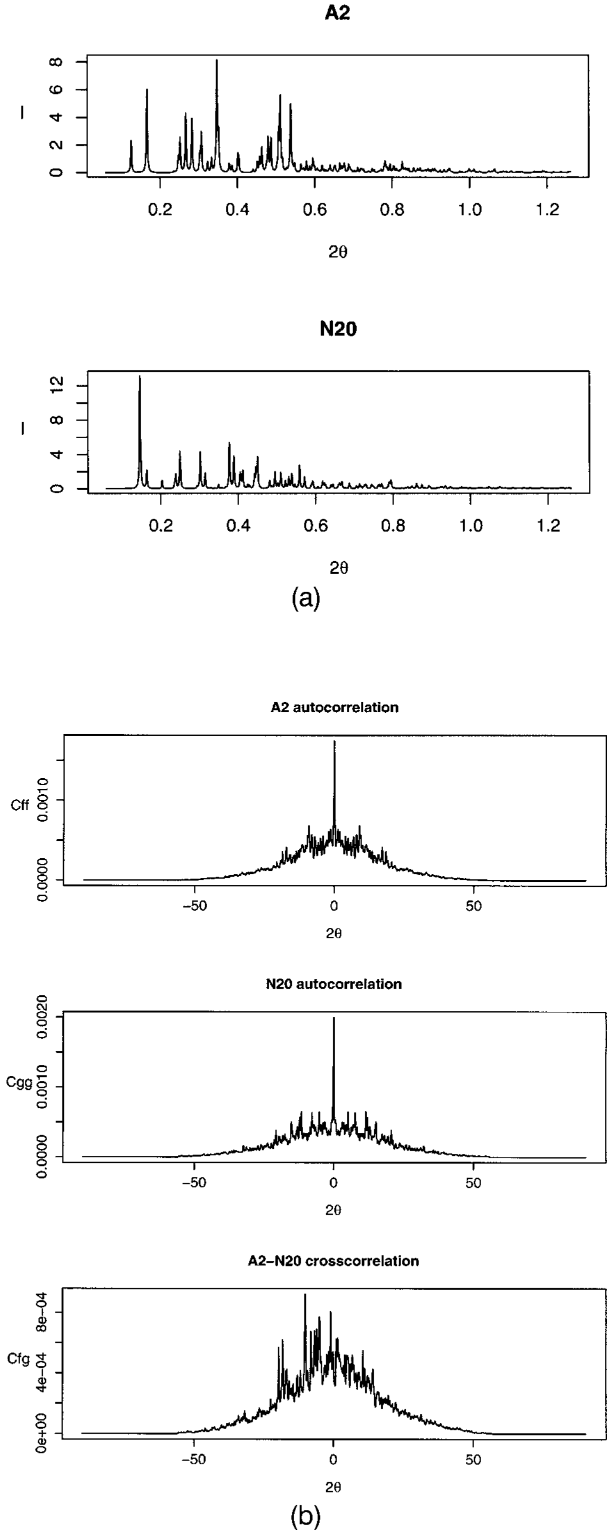

FIGURE 1. (a) Example of two different powder diffraction patterns corresponding to the Cefradine/beta-nafthol complex (A2) and the Cefradine/methyl 3-hydroxybenzoate complex (N20) (see Table II). (b) Auto- and crosscorrelation functions corresponding to the powder patterns shown in (a).

Another criterion often used for expressing the similarity between two patterns or two vectors is the Pearson product-moment correlation coefficient r_{fg}:[2]

$$r_{fg} = \frac{\int (f(x) - \langle f(x)\rangle)(g(x) - \langle g(x)\rangle)\,dx}{\left(\int (f(x) - \langle f(x)\rangle)^2\,dx \int (g(x) - \langle g(x)\rangle)^2\,dx\right)^{1/2}} \qquad (5a)$$

where $\langle f(x)\rangle$ and $\langle g(x)\rangle$ are the mean values of patterns $f(x)$ and $g(x)$, i.e., $\langle f(x)\rangle = \sum f(x_i)/n$ and $\langle g(x)\rangle = \sum g(x_i)/n$ (n being the number of points x_i at which values for $f(x_i)$ and $g(x_i)$ are measured or calculated).

Defining the new patterns $f'(x) = f(x) - \langle f(x)\rangle$ and $g'(x) = g(x) - \langle g(x)\rangle$, this expression can also be written in terms of auto- and crosscorrelation functions (see Appendix A.3):

$$r_{fg} = c_{f'g'}(0)\big/\left(c_{f'f'}(0)c_{g'g'}(0)\right)^{1/2} \qquad (5b)$$

The value of r_{fg} is independent of the scale of the patterns $f'(x)$ and $g'(x)$, and thus independent of the scale of the patterns $f(x)$ and $g(x)$. r_{fg} can directly be used to express the similarity between two "unscaled" patterns $f(x)$ and $g(x)$.

The most important conclusion is that the Pearson product-moment correlation coefficient is also based on the values of the auto- and crosscorrelation functions at $r = 0$ only. However, the terms $\langle f(x)\rangle$ and $\langle g(x)\rangle$ in expression (5a) introduce a shift of the original patterns $f(x)$ and $g(x)$ in the y-direction. Moreover, a specific scaling is implicitly applied to $f(x)$ and $g(x)$, while using criterion (5a). When the values of $\int f'(x)^2\,dx$ and $\int g'(x)^2\,dx$ are normalized to 1, by scaling $f(x)$ and $g(x)$, the values of r_{fg} and $c_{f'g'}(0)$ will become the same. This is not the same as scaling by the total number of counts [setting the areas under $f(x)$ and $g(x)$ to the same value]. Instead, for a simple pointwise comparison of patterns based on the correlation function, $c_{fg}(0)$ divided by the root of the product of the autocorrelation integrals for $f(x)$ and $g(x)$ [see expression (3c)] could be used:

$$c'_{fg}(0) = c_{fg}(0)\Big/\left(\int f(x)\,dx \int g(x)\,dx\right) \qquad (5c)$$

This criterion can be used for "unscaled" patterns and measures the similarity on the basis of the crosscorrelation function in point 0, as if the patterns were prescaled according to the total number of counts. The value of $c'_{fg}(0)$ ranges from 0 to 1.

The Pearson product-moment correlation coefficient is closely related to the overlap integral $S_{\alpha\beta}$ that is described by Lawton and Bartell.[3] In principle, the method they propose is based on peak positions (lines) deduced from powder diagrams.

By representing the diffraction peaks by Gaussian functions they simulate a profile that can be used to calculate an overlap integral. This overlap integral is a direct index of how well two patterns match each other.

It can easily be seen that the overlap integral is a normalized crosscorrelation function at $r = 0$ and is similar to expression (5b), the Pearson product-moment correlation coefficient, when the simulated profile, based on lines, is replaced by the whole pattern, calculated or measured [see Appendix A.4]. The only difference is a base-line shift [via the terms $\langle f(x)\rangle$ and $\langle g(x)\rangle$], which is not incorporated in the overlap integral. In other words, if the terms $\langle f(x)\rangle$ and $\langle g(x)\rangle$ are removed from expression (5a) one obtains the overlap integral of Lawton and Bartell:

$$S_{\alpha\beta} = c_{fg}(0)\big/\big(c_{ff}(0)c_{gg}(0)\big)^{1/2} \tag{5d}$$

To summarize for these pointwise criteria: the difference criterion can be transformed into a criterion, which includes terms that are identical to the values of the autocorrelation functions for $f(x)$ and $g(x)$ and the crosscorrelation function for $f(x)$ and $g(x)$ at $r = 0$. The difference criterion is, in principle, the dissimilarity counterpart of the overlap integral of Lawton and Bartell if the patterns are scaled according to $\int f^2(x)\,dx$ and $\int g^2(x)\,dx$. If a baseline shift is applied to $f(x)$ and $g(x)$, according to $\langle f(x)\rangle$ and $\langle g(x)\rangle$, the overlap integral of Lawton and Bartell is transformed into the Pearson product-moment correlation coefficient.

Similarity and Dissimilarity Criteria Including Neighborhoods

The pointwise difference criterion d_{fg} can be extended to a neighborhood criterion by defining an expression $d_{fg}(r)$ in analogy with $c_{fg}(r)$ [see Appendix A.5]:

$$\begin{aligned} d_{fg}(r) &= \int \big(f(x) - g(x+r)\big)^2\,dx \\ &= c_{ff}(0) + c_{gg}(0) - 2c_{fg}(r) \end{aligned}$$

$$\int d_{fg}(r)\,dr = c_{ff}(0) + c_{gg}(0) - 2\int c_{fg}(r)\,dr \tag{6}$$

Although the dissimilarity criterion (6) contains two constants $c_{ff}(0)$ and $c_{gg}(0)$ and the same term as (3b), this criterion is not just the dissimilarity counterpart of (3b). When $f(x)$ and $g(x)$ are scaled on the basis of $\int f(x)\,dx$ and $\int g(x)\,dx$, the sum of the terms $c_{ff}(0)$ and $c_{gg}(0)$ may have a different value for each different pair of patterns, because these terms are related to the sum of the squared pattern values.

It will now be shown that criterion (6) shows close resemblance to the fold used by Karfunkel et al.[4] Their fold criterion is defined as:

$$\mathbf{d}^T\mathbf{F}\mathbf{d}$$

where an element of the vector $\mathbf{d}$, $d(x) = f(x) - g(x)$.

The elements of the matrix $\mathbf{F}$ are defined as:

$$F_{ij} = 1\big/\big(1 + \alpha|i-j|^{\beta}\big)$$

To compare this criterion with the crosscorrelation function we initially assume that all elements of $\mathbf{F}$ are equal to 1. This would mean that we use equal weigths in the comparison of a point of the reference pattern with the neighborhood of the corresponding point on the target pattern. It can easily be shown that in that case the fold criterion is a criterion based on correlation integrals only [see Appendix A.6]:

$$\mathbf{d}^T\mathbf{F}\mathbf{d} = \int c_{ff}(r)\,dr + \int c_{gg}(r)\,dr - 2\int c_{fg}(r)\,dr \tag{7}$$

where $r = i - j$. Note the analogy of expression (7) with expressions (4b) and (6).

If patterns $f(x)$ and $g(x)$ are scaled on the basis of the total sum of the counts, the fold criterion leads to the same but opposite results as criterion (3b), which would mean a fold-value of 0 for any combination of $f(x)$ and $g(x)$, when all matrix elements of $\mathbf{F}$ are set to 1. Where (3b) is a similarity criterion, (7) is the corresponding dissimilarity criterion. By introducing the concept of comparing one point on the reference pattern with the neighborhood of the corresponding point on the target pattern, Karfunkel et al. transformed the conventional criterion, based on squared differences, to a correlation integral criterion. It may now be clear that, before calculating the fold, the scaling of the patterns *must* be the procedure proposed by Karfunkel et al. for reasons related to the normalization of the autocorrelations integrals, the first and second term of (7). However, once the original matrix $\mathbf{F}$ of Karfunkel et al. is used, renormalization of the crosscorrelation integral may be needed, as will be shown now.

Introducing the original matrix $\mathbf{F}$ into the expression for (7) would lead to [see Appendix A.6]:

$$\begin{aligned} \mathbf{d}^T\mathbf{F}\mathbf{d} = \int w(r)c_{ff}(r)\,dr + \int w(r)c_{gg}(r)\,dr \\ - 2\int w(r)c_{fg}(r)\,dr \end{aligned} \tag{8}$$

where $w(r) = 1/(1 + \alpha|r|^{\beta})$ $(r = i - j)$.

The similarity counterpart of (8), expressing the similarity between the patterns using the same function $w(r)$, would be a weighted crosscorrelation integral:

$$\int c^{w}_{fg}(r)\,dr = \int w(r) \int f(x)g(x+r)\,dx\,dr \qquad (9a)$$

and the associated autocorrelation integrals would include the same function $w(r)$:

$$\int c^{w}_{ff}(r)\,dr = \int w(r) \int f(x)f(x+r)\,dx\,dr \qquad (9b)$$

$$\int c^{w}_{gg}(r)\,dr = \int w(r) \int g(x)g(x+r)\,dx\,dr \qquad (9c)$$

An important conclusion is that the fold criterion of Karfunkel et al. can be seen as a dissimilarity counterpart of a weighted correlation integral $\int c_{fg}(r)\,dr$, as given in (3b) [weighted with the function $w(r)$].

If the patterns $f(x)$ and $g(x)$ are scaled on the basis of the total number of counts, it may be clear from expressions (9b) and (9c) that the corresponding autocorrelation integrals do not necessarily result in the same value. On the other hand, it may now be clear that the matrix $\mathbf{F}$, or a different function $w(r)$, is needed to extract the similarity information from the crosscorrelation function. To ensure that the autocorrelation integrals will give identical values, the weighted crosscorrelation integral $\int c^{w}_{fg}(r)\,dr$ must be normalized to obtain a similarity measure C_{fg} on an absolute scale:

$$C_{fg} = \int c^{w}_{fg}(r)\,dr \Big/ \left(\int c^{w}_{ff}(r)\,dr \int c^{w}_{gg}(r)\,dr\right)^{1/2} \qquad (10)$$

This similarity criterion will yield a value of 1 when patterns $f(x)$ and $g(x)$ are identical and a value between 0 and 1 for other cases. The corresponding dissimilarity criterion, which can be obtained from (7) by substituing C_{fg}, C_{ff}, and C_{gg} for $\int c_{fg}(r)\,dr$, $\int c_{ff}(r)\,dr$, and $\int c_{gg}(r)\,dr$, respectively, will yield values between 0 and 2. In fact, this dissimilarity criterion is a renormalized "fold."

A Generalized Expression for Similarity and Dissimilarity

All criteria described before can be summarized by the following expressions for similarity and dissimilarity. The generalized expression for the similarity S_{fg} between patterns $f(x)$ and $g(x)$ is given by:

$$S_{fg} = \int w_{fg}(r)c_{fg}(r)\,dr \Big/ \left(\int w_{ff}(r)c_{ff}(r)\,dr \int w_{gg}(r)c_{gg}(r)\,dr\right)^{1/2} \qquad (11)$$

The corresponding generalized expression for the dissimilarity D_{fg} is given by:

$$D_{fg} = S_{ff} + S_{gg} - 2S_{fg} \qquad (12)$$

The function $w_{fg}(r)$ determines the way in which the similarity information is extracted from the crosscorrelation function and the functions $w_{ff}(r)$ and $w_{gg}(r)$ determine the normalization of the weighted crosscorrelation function via the autocorrelation functions. For obtaining a similarity or dissimilarity measure on an absolute scale the following condition must hold:

$$w_{ff}(r) = w_{gg}(r) = w_{fg}(r)$$

To include the neighborhood in the comparison of points the weighting functions should be defined for $r \neq 0$. Both aspects are important and can easily be combined.

The differences between the various criteria described in literature can simply be explained by a different definition of the weighting functions $w_{ff}(r)$, $w_{gg}(r)$, and $w_{fg}(r)$. In Table I an overview is given of the various criteria and their corresponding use of the three weighting functions. From Table I it can be seen that none of the criteria described in literature include both the concept of neighborhood and a correct normalization to obtain a measure on an absolute scale. To define a similarity measure on an absolute scale in analogy with the fold of Karfunkel et al. the generalized similarity measure S_{fg} could be used with weighting functions:

$$w_{fg}(r) = 1\big/\left(1 + \alpha|r|^{\beta}\right)$$
$$w_{ff}(r) = w_{gg}(r) = w_{fg}(r)$$

In principle, we now have defined a new criterion, a normalized fold, with different characteristics than the original fold. In the next section it is shown that the weighting function $w_{fg}(r)$, as defined for the fold, can be replaced by a simple triangle function.

Application of the Generalized Expression for Similarity to Structure Classification from Powder Diffraction Patterns

It has been shown that the various criteria for similarity and dissimilarity described in the literature can be deduced from the generalized expressions (11) or (12). To obtain a measure on an absolute scale, a suitable normalization should be used, which is easily done by taking identical expressions for the weighting functions $w_{ff}(r)$, $w_{gg}(r)$,

TABLE I.
(Dis)similarity Criteria and Their Corresponding Weighting Functions.

Criterion	Equation	Type	$w_{fg}(r)$	$w_{ff}(r)$	$w_{gg}(r)$
Difference criterion d_{fg}	(4a) and (4b)	D_{fg}	1 if $r = 0$, 0 if $r \neq 0$	1	1
Pearson product moment correlation coeficient r_{fg}[a]	(5a) and (5b)	S_{fg}	1 if $r = 0$, 0 if $r \neq 0$	$= w_{fg}(r)$	$= w_{fg}(r)$
Overlap integral Lawton and Bartell	(5d)	S_{fg}	1 if $r = 0$, 0 if $r \neq 0$	$= w_{fg}(r)$	$= w_{fg}(r)$
Fold criterion Karfunkel et al.	(8)	D_{fg}	$1/(1 + \alpha\|r\|^{\beta})$	1	1
Normalized fold criterion	(10)	D_{fg}	$1/(1 + \alpha\|r\|^{\beta})$	$= w_{fg}(r)$	$= w_{fg}(r)$
Newly proposed similarity criterion	(11)	S_{fg}	$1 - \|r\|/l$ if $\|r\| < l$ 0 if $\|r\| \geq l$	$= w_{fg}(r)$	$= w_{fg}(r)$

[a] A mean centering of $f(x)$ and $g(x)$ should be applied first.

and $w_{fg}(r)$ in expressions (11) or (12). The question of which similarity (or dissimilarity) criterion can best be used for a given application now focusses on the choice of the weighting function $w_{fg}(r)$. This weighting function defines the neighborhood and associated weights in the comparison of two corresponding points on the reference and target patterns.

The conventional difference criterion, the Pearson product-moment correlation coefficient, and the overlap integral of Lawton and Bartell use a delta function for $w_{fg}(r)$ and do not incorporate any contributions from the neighborhood of a point. The fold criterion of Karfunkel et al. includes a weighting function that can be tuned by two parameters α and β. These two parameters define the shape and width of the weighting function.

The differences in discriminating power between the conventional difference criterion, the overlap integral of Lawton and Bartell, the fold criterion of Karfunkel et al., and a newly proposed similarity criterion, which is obtained by using a simple triangle weighting function in the generalized expression for similarity, was investigated in the classification of 20 crystal structures of complexes of cephalosporin antibiotics on the basis of their calculated powder diffraction patterns. From single-crystal X-ray diffraction analyses of these compounds (Kemperman, De Gelder, Dommerholt, Raemakers–Franken, Klunder, and Zwanenburg[9, 10]) it was found that among these 20 complexes six different crystal forms are found. Ten compounds crystallize in form A, four compounds in type B, two compounds in form N and the remaining four compounds in types C, D, E, and F. In Table II the crystal data are shown for these 20 complexes. From this table it can be seen that small differences are present in the cell parameters of the compounds belonging to the same crystal form. These small differences give rise to large peak shifts in the corresponding powder diffraction patterns. In Figure 2, the simulated powder diffractions patterns for the different crystal forms are shown. For forms C, D, E, and F, only one member can be shown. For forms A, B, and N, two representative members are shown, clearly illustrating the large peak shifts resulting from the small differences in unit cell parameters.

The question is now whether these 20 complexes can be classified on the basis of their powder patterns using a dedicated similarity criterion. The different crystal forms should be recognized as dissimilar, however, the complexes belonging to the same crystal form should ideally be recognized as similar compounds, and should somehow be clustered together.

Similarities or dissimilarities [which can be interconverted, see (11) and (12)] were calculated with four criteria (the difference criterion, the overlap integral of Lawton and Bartell, the fold criterion, and the newly proposed similarity criterion) for each pair of patterns. For the parameters α and β corresponding to the fold criterion, the optimized values of the authors[4] were used (10^{-7} and 4, respectively). For the newly proposed similarity criterion the following simple triangle weighting function was used:

$$w_{fg}(r) = 1 - |r|/l \qquad \text{if } |r| < l$$
$$w_{fg}(r) = 0 \qquad \text{if } |r| \geq l$$

The parameter l defines the width (degrees 2θ) of the neighborhood taken into account. This func-

TABLE II.
Crystal Data of the Various Cephalosporin Complexes.

Compound name	a	b	c	α	β	γ	Space Group	Type
Cefradine/alpha-naphthol	23.47079	7.12154	14.93036	90.0	108.26834	90.0	C2	A1
Cefradme/beta-naphthol	23.42119	6.97147	15.00473	90.0	110.40521	90.0	C2	A2
Cefradine/naphthalene	23.45837	7.11788	14.89216	90.0	108.57098	90.0	C2	A3
Cefradine/quinoline	23.41265	7.10910	14.80598	90.0	108.14590	90.0	C2	A4
Cefradine/2-hydroxy acetonaphthone	23.38549	7.19649	14.75882	90.0	108.57974	90.0	C2	A5
Cefradine/bipyridine	23.02269	7.14670	14.55443	90.0	104.64385	90.0	C2	A6
Cephalexine/beta-naphthol	23.39759	7.06229	14.91757	90.0	109.79842	90.0	C2	A7
Cephalexine/alpha-naphthol	23.43432	7.10804	14.87538	90.0	108.19072	90.0	C2	A8
Cefaclor/alpha-naphthol	23.48840	7.07855	14.84551	90.0	108.94678	90.0	C2	A9
Cefaclor/beta-naphthol	23.44692	7.02619	14.84134	90.0	110.55009	90.0	C2	A10
Cephadroxil/beta-naphthol	7.11174	21.71704	30.95857	90.0	90.0	90.0	$P2_12_12_1$	B11
Cephadroxil/4-hydroxy-benzoezuur	6.99921	20.99127	30.69011	90.0	90.0	90.0	$P2_12_12_1$	B12
Cephadroxil/2,6-dihydroxy-naphthalene	7.10786	21.86340	32.30589	90.0	90.0	90.0	$P2_12_12_1$	B13
Cephadroxil/2,7-dihydroxy-naphthalene	7.09018	21.27323	31.00436	90.0	90.0	90.0	$P2_12_12_1$	B14
Ceftadine/4-hydroxy-benzoezuur	14.91661	7.38199	20.50296	90.0	105.77318	90.0	$P2_1$	C15
Cefradine/2-phenylphenol	23.56421	7.13203	18.68928	90.0	109.37986	90.0	C2	D16
Cefradine/hydroquinone	7.07185	10.70306	14.23422	87.15449	78.99942	89.74252	P1	E17
Cefradine/4-methylacetophenone	15.40382	7.29832	23.57355	90.0	99.35406	90.0	$P2_1$	F18
Cefradine/DMF	10.87473	9.51140	12.39035	90.0	98.70461	90.0	$P2_1$	N19
Cefradine/methyl 3-hydroxybenzoate	10.90731	9.40654	12.19924	90.0	98.53256	90.0	$P2_1$	N20

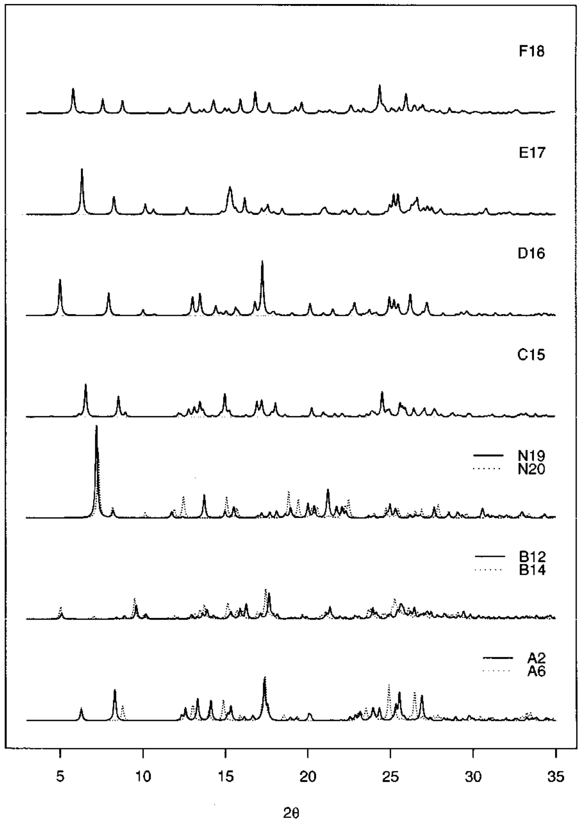

FIGURE 2. Simulated powder diffraction patterns of the various crystal forms found for complexes of cephalosporin antibiotics.

tion extracts information from the central part of the crosscorrelation function with a weight that decreases proportionally to the distance from the origin ($r = 0$). After a number of experiments it was found that values of l between 0.6 and 3.0 lead to stable and comparable results for the powder diffractions patterns of the 20 complexes. A value of 0.6 was chosen for further calculations. In Figure 3, the weighting function of Karfunkel et al. and triangle weighting function for $l = 0.6$ and $l = 3.0$ are shown for comparison. The effect of the triangle weighting function on the crosscorrelation function, that was also shown in Figure 1, is illustrated in Figure 4.

The similarity calculations lead to four (dis)similarity matrices [similarities are eventually converted to dissimilarities using expression (11)] that are shown in Table III. Note that only the similarity matrices obtained with the overlap integral of Lawton and Bartell and the newly proposed similarity criterion, using the triangle weighting function, contain values on an absolute scale (values between 0 and 1).

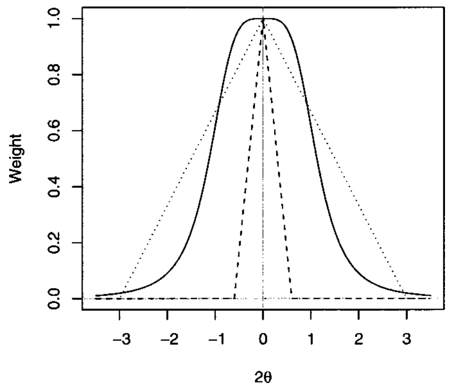

FIGURE 3. The weighting function of Karfunkel et al. (solid line), the triangle weighting function with $l = 0.6$ (dashed line) and $l = 3.0$ (dotted line).

On the dissimilarity matrices, a clustering algorithm was applied to group the patterns that are considered to be similar on the basis of the data in the matrices. A hierarchical agglomerative clustering method was used for this purpose. Initially, each object is viewed as a separate cluster; in each subsequent step, similar objects are joined according to a distance criterion, and the distances of the newly formed cluster to the other clusters or objects are recalculated. This process continues until all objects are joined into one cluster. The criterion that is used is known as "Ward's method,"[11] where elements or clusters are joined in such a way that the sum of heterogeneities of all clusters (defined as the summed squared distance of each member of a cluster to the centroid of that cluster) increases as little as possible. The method performs best in cases where the clusters are approximately spherical in shape and of equal size, and is widely applied. This cluster-

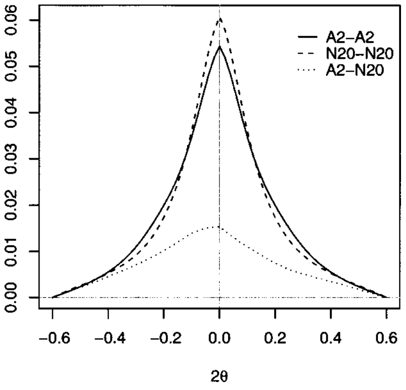

FIGURE 4. Weighted auto- and crosscorrelation functions (triangle weighting function with $l = 0.6$) for A2 and N20.

TABLE IIIa.
Pattern Dissimilarity Values Calculated with the Difference Criterion.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	0.00	0.21	0.02	0.06	0.11	0.14	0.13	0.02	0.11	0.22	0.17	0.20	0.17	0.20	0.17	0.16	0.23	0.22	0.39	0.31
A2		0.00	0.18	0.16	0.21	0.15	0.11	0.19	0.14	0.07	0.12	0.14	0.17	0.12	0.17	0.20	0.16	0.17	0.37	0.27
A3			0.00	0.03	0.11	0.12	0.10	0.02	0.09	0.20	0.15	0.18	0.15	0.17	0.16	0.13	0.20	0.20	0.37	0.30
A4				0.00	0.13	0.11	0.09	0.03	0.08	0.19	0.14	0.18	0.17	0.17	0.17	0.15	0.17	0.20	0.38	0.30
A5					0.00	0.19	0.18	0.12	0.17	0.24	0.19	0.19	0.17	0.19	0.16	0.19	0.22	0.20	0.39	0.31
A6						0.00	0.15	0.13	0.15	0.16	0.09	0.14	0.16	0.13	0.16	0.18	0.19	0.18	0.36	0.29
A7							0.00	0.09	0.04	0.08	0.14	0.19	0.17	0.14	0.19	017	0.18	0.21	0.40	0.31
A8								0.00	0.07	0.20	0.15	0.20	0.17	0.18	0.17	0.15	0.21	0.22	0.39	0.31
A9									0.00	0.11	0.14	0.19	0.17	0.15	0.19	0.18	0.20	0.21	0.39	0.31
A10										0.00	0.15	0.18	0.19	0.14	0.20	0.23	0.21	0.21	0.40	0.32
B11											0.00	0.10	0.09	0.04	0.13	0.14	0.16	0.12	0.31	0.24
B12												0.00	0.10	0.08	0.12	0.18	0.14	0.10	0.28	0.23
B13													0.00	0.09	0.11	0.16	0.16	0.12	0,30	0.24
B14														0.00	0.13	0.15	0.14	0.12	0.29	0.23
C15															0.00	0.18	0.18	0.12	0.31	0.25
D16																0.00	0.24	0.21	0.38	0.33
E17																	0.00	0.18	0.35	0.27
F18																		0.00	0.32	0.23
N19																			0.00	0.22
N20																				0.00

TABLE IIIb.
Pattern Similarity Values Calculated with the Overlap Integral of Lawton and Bartell.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	1.00	0.47	0.96	0.85	0.73	0.64	0.69	0.94	0.73	0.47	0.51	0.40	0.53	0.41	0.53	0.60	0.41	0.34	0.20	0.25
A2		1.00	0.51	0.56	0.45	0.56	0.72	0.52	0.63	0.82	0.59	0.50	0.44	0.59	0.44	0.47	0.53	0.40	0.18	0.27
A3			1.00	0.92	0.73	0.68	0.75	0.96	0.78	0.50	0.55	0.41	0.53	0.46	0.52	0.66	0.45	0.36	0.21	0.25
A4				1.00	0.68	0.70	0.77	0.93	0.80	0.53	0.57	0.40	0.49	0.48	0.50	0.61	0.53	0.35	0.19	0.23
A5					1.00	0.50	0.56	0.70	0.58	0.42	0.44	0.38	0.50	0.42	0.54	0.54	0.40	0.40	0.18	0.23
A6						1.00	0.62	0.65	0.61	0.57	0.69	0.47	0.44	0.55	0.47	0.52	0.43	0.38	0.20	0.23
A7							1.00	0.78	0.90	0.79	0.59	0.38	0.49	0.59	0.44	0.57	0.52	0.35	0.18	0.24
A8								1.00	0.83	0.52	0.56	0.39	0.52	0.46	0.52	0.62	0.44	0.34	0.19	0.25
A9									1.00	0.73	0.59	0.39	0.49	0.55	0.43	0.56	0.44	0.35	0.18	0.23
A10										1.00	0.58	0.44	0.43	0.60	0.43	0.43	0.44	0.36	0.17	0.21
B11											1.00	0.57	0.62	0.84	0.47	0.58	0.45	048	0.25	0.28
B12												1.00	0.55	0.63	0.48	0.41	0.49	0.52	0.27	0.25
B13													1.00	0.60	0.54	0.51	0.43	0.51	0.25	0.26
B14														1.00	0.49	0.53	0.52	0.47	0.28	0.28
C15															1.00	0.45	0.40	0.50	0.24	0.25
D16																1.00	0.35	0.33	0.19	0.18
E17																	1.00	0.35	0.21	0.27
F18																		1.00	0.20	0.28
N19																			1.00	0.54
N20																				1.00

TABLE IIIc.
Pattern Dissimilarity Values Calculated with Fold Criterion.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	0.00	−0.12	0.07	0.37	−0.01	0.51	−0.03	0.00	0.35	0.25	0.93	1.73	0.97	1.00	0.51	0.79	2.99	0.69	6.46	4.85
A2		0.00	−0.06	0.17	0.03	0.22	0.05	−0.09	0.36	0.42	0.65	1.34	0.78	0.80	0.33	0.63	2.55	0.59	5.91	4.43
A3			0.00	0.19	0.04	0.33	0.10	0.08	0.49	0.42	0.75	1.47	0.86	0.86	0.43	0.71	2.83	0.53	6.18	4.72
A4				0.00	0.25	0.39	0.41	0.36	0.84	0.84	0.87	1.36	1.00	0.92	0.28	0.80	2.21	0.33	5.51	4.35
A5					0.00	0.49	−0.03	−0.02	0.16	0.17	1.07	1.97	1.23	1.12	0.62	0.95	3.10	0.80	6.37	4.92
A6						0.00	0.61	0.48	0.72	0.73	0.63	1.06	1.05	0.70	0.63	1.18	2.84	0.66	7.08	5.40
A7							0.00	0.00	0.43	0.34	1.06	1.98	1.09	1.20	0.58	0.79	3.35	0.93	6.49	5.03
A8								0.00	0.34	0.26	0.92	1.67	0.96	0.97	0.48	0.80	3.01	0.67	6.38	4.77
A9									0.00	0.01	1.43	1.93	1.58	1.31	0.98	1.48	3.72	1.05	7.07	5.27
A10										0.00	1.51	2.19	1.71	1.48	1.06	1.37	4.00	1.37	7.25	5.68
B11											0.00	0.39	0.30	0.11	0.76	0.69	2.22	0.58	6.68	5.11
B12												0.00	0.56	0.33	1.29	1.56	1.88	0.98	6.26	4.70
B13													0.00	0.28	1.02	0.86	1.85	0.80	6.26	5.05
B14														0.00	0.73	1.02	1.89	0.58	6.56	4.99
C15															0.00	1.16	1.59	0.35	4.11	3.33
D16																0.00	2.90	1.00	6.23	5.22
E17																	0.00	1.29	5.25	4.07
F18																		0.00	4.60	3.19
N19																			0.00	1.18
N20																				0.00

TABLE IIId.
Pattern Similarity Values Calculated with Generalized Similarity Criterion.

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B11	B12	B13	B14	C15	D16	E17	F18	N19	N20
A1	1.00	0.80	0.99	0.96	0.93	0.88	0.93	0.99	0.92	0.82	0.78	0.60	0.78	0.72	0.75	0.80	0.63	0.66	0.34	0.41
A2		1.00	0.82	0.82	0.78	0.78	0.93	0.83	0.87	0.95	0.80	0.75	0.75	0.80	0.76	0.77	0.68	0.69	0.34	0.41
A3			1.00	0.98	0.92	0.90	0.94	0.99	0.93	0.83	0.81	0.63	0.79	0.75	0.76	0.82	0.66	0.67	0.35	0.41
A4				1.00	0.90	0.90	0.93	0.97	0.91	0.81	0.79	0.63	0.76	0.74	0.76	0.78	0.71	0.67	0.34	0.39
A5					1.00	0.82	0.86	0.91	0.85	0.79	0.72	0.61	0.75	0.69	0.81	0.75	0.58	0.73	0.31	0.40
A6						1.00	0.85	0.89	0.85	0.79	0.85	0.69	0.75	0.77	0.74	0.75	0.63	0.68	0.33	0.38
A7							1.00	0.95	0.96	0.94	0.81	0.66	0.77	0.78	0.75	0.81	0.66	0.66	0.35	0.40
A8								1.00	0.95	0.85	0.80	0.61	0.77	0.74	0.75	0.80	0.65	0.65	0.34	0.41
A9									1.00	0.93	0.79	0.66	0.75	0.77	0.71	0.78	0.63	0.64	0.34	0.39
A10										1.00	0.78	0.71	0.73	0.78	0.73	0.77	0.60	0.66	0.32	0.37
B11											1.00	0.84	0.87	0.96	0.73	0.83	0.66	0.69	0.41	0.42
B12												1.00	0.78	0.89	0.70	0.67	0.67	0.74	0.42	0.41
B13													1.00	0.87	0.77	0.83	0.66	0.75	0.42	0.44
B14														1.00	0.73	0.78	0.70	0.70	0.44	0.43
C15															1.00	0.67	0.70	0.75	0.43	0.44
D16																1.00	0.55	0.63	0.36	0.35
E17																	1.00	0.60	0.37	0.45
F18																		1.00	0.42	0.53
N19																			1.00	0.80
N20																				1.00

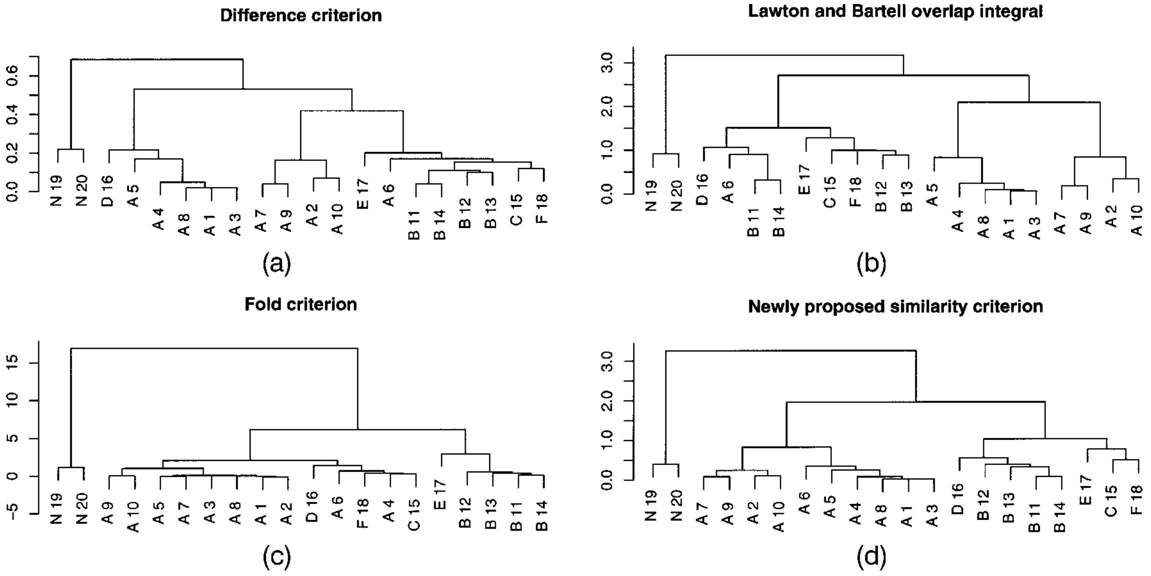

FIGURE 5. (a) Clustering dendrogram found for the difference criterion. (b) Clustering dendrogram found for the Lawton and Bartell overlap integral. (c) Clustering dendrogram found for the Fold criterion. (d) Clustering dendrogram found for the newly proposed similarity criterion.

ing procedure yields the dendrograms as depicted in Figure 5. Clearly, the results are quite different for the four (dis)similarity criteria. The newly proposed similarity criterion, using the triangle weighting function, leads to the most homogeneous classification and is able to seperate the crystal structures into four groups consisting of 1: only N, 2: only A, 3: four B's and one D, and 4: E, C, and F.

To illustrate the stability of the generalized similarity criterion using the simple triangle weighting function, similarity matrices were calculated for increasing values of the parameter *l*. The same clustering procedure was applied on the corresponding similarity matrices, and the resulting dendrograms are shown in Figure 6. For small values of *l* (smaller than 0.6), very inhomogeneous classification dendrograms are found. For very large values of *l* (larger than 3.0) also inhomogeneous classification dendrograms are found. However, in the range 0.6 to 3.0, similar dendrograms of comparable homogeniety are found.

Discussion and Conclusion

The generalized expression for the similarity of powder diffraction patterns shows that the criteria described in literature all refer to the correlation function. It also shows that the differences between the criteria can be explained by different choices of weighting functions for the auto- and crosscorrelation terms. The nature of the weighting functions used determines whether a pointwise or neighborhood approach is applied, and whether the resulting measure for similarity or dissimilarity is on an absolute scale or not.

The importance of a neighborhood approach has been demonstrated for the classification of crystal structures on the basis of their calculated powder diffraction patterns. Including the neighborhoods

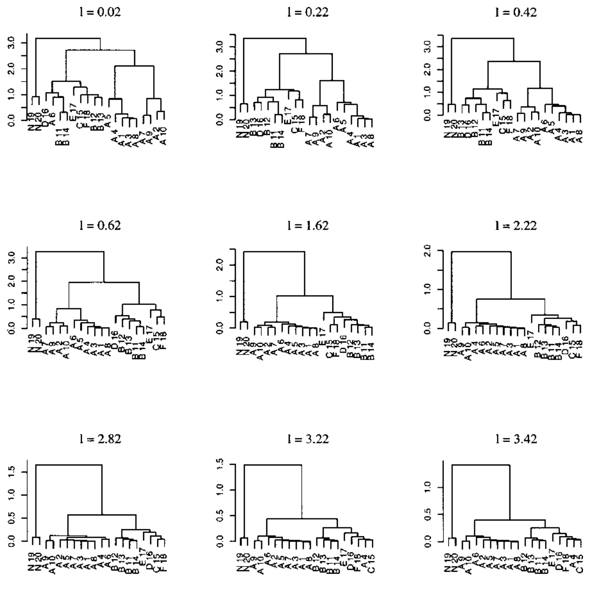

FIGURE 6. Clustering dendrograms found for increasing values of *l* (only unique dendrograms in the range *l* = 0.02 to 3.42 are shown).

leads to a significant improvement of discriminating power compared to pointwise approaches, and makes it possible to recognize closely related structures. The way in which the neighborhood is taken into account depends on the shape of the weighting function used. A simple triangle function leads to very useful results and performs even better than the more complex weighting function used by Karfunkel et al. It should, however, be emphasized that we did not try to adapt the parameters defining the fold weighting function to optimize the classification of our set of structures. We just took the function and parameters the authors optimized for their specific problem.

In our tests we used calculated powder diffraction patterns as the source of structural information. In practice, one should like to apply the classification procedure on experimentally obtained powder patterns. In that case, there might be factors like zero-point shifts, preferred orientation, peak broadening, nonzero background and experimental noise that may influence the outcome of the classification procedure. However, the classification of experimentally obtained patterns and its associated problems will be the subject of our further research, and will not be discussed any further here.

In the introduction it was mentioned that the generalized expression might also be used for database searching and optimization problems. For database searching it is of crucial importance that a measure for similarity or dissimilarity is defined on an absolute scale. It is necessary to know the numerical range of the measure to define limits for acceptance in the selection of subsets of patterns from a database. The newly proposed criterion for similarity can be used for the purpose of database searching, because its value always ranges from 0 to 1. The width of the triangle function can be adapted to change the search from a strict to a more generous one. It is up to the user to optimize the settings for his particular case.

For optimization problems the aspect of normalization, to obtain a measure on an absolute scale, is less important. The width of the neighborhood, however, may determine the overall success of an optimization procedure (see Hageman et al.[6]) in which parameters defining a theoretical pattern are optimized with respect to an experimental spectrum. The effect of the inclusion of the neighborhood is a more realistic assesment of the error, and this may guide the process to the global optimum. A possible application of the newly proposed similarity criterion could be the determination of cell and/or positional parameters of crystal structures from powder diffraction data only. In principle, the criterion allows for a gradual improvement of both peak positions and peak intensities as was shown by Hageman et al.

Although this article has focussed on powder diffraction patterns, which are one-dimensional patterns, it should be emphasized that the generalized expression for similarity is applicable to other types of spectra and is not limited to one-dimensional patterns only. If the variables x and r are replaced by vectors the expression can directly be used in multidimensional space. Then it can also be applied, for example, to express the similarity between 2D, 3D, or 4D-NMR spectra.

Appendix

A.1. The Integral of the Autocorrelation Function $c_{ff}(r)$

$$c_{ff}(r) = \int f(x) f(x+r)\, dx$$

$$\begin{aligned}\int c_{ff}(r)\, dr &= \iint f(x) f(x+r)\, dx\, dr \\ &= \iint f(x) f(r')\, dx\, dr' \qquad (r = r' - x) \\ &= \int f(x)\, dx \int f(r')\, dr' = \left(\int f(x)\, dx \right)^2\end{aligned}$$

A.2. d_{fg} Written as Auto- and Crosscorrelation Functions

$$\begin{aligned}d_{fg} &= \int \left(f(x) - g(x) \right)^2 dx \\ &= \int f^2(x)\, dx + \int g^2(x)\, dx - 2 \int f(x) g(x)\, dx\end{aligned}$$

A pointwise comparison of patterns $f(x)$ and $g(x)$ (neglecting the neighborhoods of points in the patterns) would, speaking in terms of the crosscorrelation function $c_{fg}(r)$, be identical to calculating the crosscorrelation function for patterns $f(x)$ and $g(x)$ at $r = 0$:

$$c_{fg}(0) = \int f(x) g(x)\, dx$$

Therefore, d_{fg} can be written as:

$$d_{fg} = c_{ff}(0) + c_{gg}(0) - 2c_{fg}(0)$$

A.3. r_{fg} Written as Auto- and Crosscorrelation Functions

$$r_{fg} = \frac{\int (f(x) - \langle f(x)\rangle)(g(x) - \langle g(x)\rangle)\,dx}{\left(\int (f(x) - \langle f(x)\rangle)^2\,dx \int (g(x) - \langle g(x)\rangle)^2\,dx\right)^{1/2}}$$

Defining the new patterns $f'(x) = f(x) - \langle f(x)\rangle$ and $g'(x) = g(x) - \langle g(x)\rangle$ and considering these new patterns $f'(x)$ and $g'(x)$ as vectors $\mathbf{f}'$ and $\mathbf{g}'$ in n-dimensional space (n being the number of points x_i at which values for $f'(x_i)$ and $g'(x_i)$ are measured or calculated) this expression reduces to the cosine of the angle $\alpha_{f'g'}$ between vectors $\mathbf{f}'$ and $\mathbf{g}'$:

$$\begin{aligned} r_{fg} &= \cos(\alpha_{f'g'}) \\ &= \int f'(x)g'(x)\,dx \\ &\quad \Big/ \left(\int f'^2(x)\,dx \int g'^2(x)\,dx\right)^{1/2} \\ &= \mathbf{f}' \cdot \mathbf{g}' / |\mathbf{f}'||\mathbf{g}'| \end{aligned}$$

$$\begin{aligned} c_{f'f'}(0) &= \int f'^2(x)\,dx \\ c_{g'g'}(0) &= \int g'^2(x)\,dx \\ c_{f'g'}(0) &= \int f'(x)g'(x)\,dx \\ r_{fg} &= c_{f'g'}(0) \big/ \left(c_{f'f'}(0) c_{g'g'}(0)\right)^{1/2} \end{aligned}$$

A.4. The Overlap Integral of Lawton and Bartell

The overlap integral for peaks i from pattern α and peaks j from pattern β, defined by Lawton and Bartell, is expressed as follows:

$$\begin{aligned} S_{\alpha\beta} &= \sum\sum \left(I_\alpha(i) \big/ \left(\sigma_\alpha(i) a_\alpha^{1/2}\right)\right)\left(I_\beta(j) \big/ \left(\sigma_\beta(j) a_\beta^{1/2}\right)\right) \\ &\quad \times \exp\left[-\left(d_\alpha(i) - d_\beta(j)\right)^2 \big/ 4\left(w_\alpha(i) \cdot w_\beta(j)\right)\right] \end{aligned}$$

where

$$\begin{aligned} a_\alpha &= \sum\sum \left(I_\alpha(i)/\sigma_\alpha(i)\right)\left(I_\alpha(j)/\sigma_\alpha(j)\right) \\ &\quad \times \exp\left[-\left(d_\alpha(i) - d_\alpha(j)\right)^2 \big/ 4\left(w_\alpha(i) \cdot w_\alpha(j)\right)\right] \end{aligned}$$

$$\begin{aligned} a_\beta &= \sum\sum \left(I_\beta(i)/\sigma_\beta(i)\right)\left(I_\beta(j)/\sigma_\beta(j)\right) \\ &\quad \times \exp\left[-\left(d_\beta(i) - d_\beta(j)\right)^2 \big/ 4\left(w_\beta(i) \cdot w_\beta(j)\right)\right] \end{aligned}$$

I_α and I_β correspond to the diffracted relative intensities in patterns α and β, σ_α, and α_β are parameters representing the characteristic variations in these intensities, d_α and d_β correspond to the interplanar spacings in patterns α and β and w_α and w_β are window parameters describing the windows of acceptance.

A.5. $d_{fg}(r)$ Written as Auto- and Crosscorrelation Functions

$$\begin{aligned} d_{fg}(r) &= \int \left(f(x) - g(x+r)\right)^2 dx \\ &= \int f^2(x)\,dx + \int g^2(x+r)\,dx \\ &\quad - 2\int f(x)g(x+r)\,dx \\ &= c_{ff}(0) + c_{gg}(0) - 2c_{fg}(r) \end{aligned}$$

$$\int d_{fg}(r)\,dr = c_{ff}(0) + c_{gg}(0) - 2\int c_{fg}(r)\,dr$$

A.6. The Fold Written as Auto- and Crosscorrelation Integrals

$$\text{fold} = \mathbf{d}^T\mathbf{F}\mathbf{d}$$

where an element of the vector $\mathbf{d}$, $d(x) = f(x) - g(x)$.

The elements of the matrix $\mathbf{F}$ are initially set to 1:

$$F_{ij} = 1$$

$$\begin{aligned} \mathbf{d}^T\mathbf{F}\mathbf{d} &= \iint \left(f(x) - g(x)\right)\left(f(x+r) - g(x+r)\right) dx\,dr \\ &= \iint f(x)f(x+r)\,dx\,dr \\ &\quad + \iint g(x)g(x+r)\,dx\,dr \\ &\quad - \iint f(x)g(x+r)\,dx\,dr \\ &\quad - \iint f(x+r)g(x)\,dx\,dr \\ &= \int c_{ff}(r)\,dr + \int c_{gg}(r)\,dr - 2\int c_{fg}(r)\,dr \end{aligned}$$

where $r = i - j$.

Introducing the original matrix $\mathbf{F}$ into the expression for the fold leads to:

$$F_{ij} = 1 \big/ \left(1 + \alpha|i-j|^\beta\right)$$

$$\begin{aligned} \mathbf{d}^T\mathbf{F}\mathbf{d} &= \iint \left(f(x) - g(x)\right)w(r) \\ &\quad \times \left(f(x+r) - g(x+r)\right) dx\,dr \\ &= \iint w(r)f(x)f(x+r)\,dx\,dr \\ &\quad + \iint w(r)g(x)g(x+r)\,dx\,dr \end{aligned}$$

$$
\begin{aligned}
&-\iint w(r)f(x)g(x+r)\,dx\,dr \\
&-\iint w(r)f(x+r)g(x)\,dx\,dr \\
=&\int w(r)c_{ff}(r)\,dr+\int w(r)c_{gg}(r)\,dr \\
&-2\int w(r)c_{fg}(r)\,dr
\end{aligned}
$$

where $w(r) = 1/(1 + \alpha|r|^{\beta})$ $(r = i - j)$.

References

1. Young, R. A. The Rietveld Method; International Union of Crystallography; Oxford University Press: Oxford, 1993, p. 21.
2. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers–Verbeke, J. Handbook of Chemometrics and Qualimetrics: Part B, Data Handling in Science and Technology; Elsevier Science Publishers: New York, 1998, p. 20B.
3. Lawton, S. L.; Bartell, L. S. Powder Diffraction 1994, 9, 124.
4. Karfunkel, H. R.; Rohde, B.; Leusen, F. J. J.; Gdanitz, R. J.; Rihs, G. J Comp Chem 1993, 14, 1125.
5. Stephenson, D. S.; Binsch, G. J Magn Res 1980, 37, 409.
6. Hageman, J. A.; Wehrens, R.; De Gelder, R.; Meerts, W. L.; Buydens, L. M. C. J Chem Phys 2000, in press.
7. Harris, K. D. M.; Johnston, R. L.; Kariuki, B. M. Acta Crystallogr 1998, A54, 632.
8. Dods, J.; Gruner, D.; Brumer, P. J Chem Phys Lett 1996, 261, 612.
9. Kemperman, G. J.; De Gelder, R.; Dommerholt, F. J.; Raemakers–Franken, P. C.; Klunder, A. J. H.; Zwanenburg, B. Chem Eur J 1999, 5, 2163.
10. Kemperman, G. J.; De Gelder, R.; Dommerholt, F. J.; Raemakers–Franken, P. C.; Klunder, A. J. H.; Zwanenburg, B. Perkins Trans II 2000, 7, 1425.
11. Kaufman, L.; Rousseeuw, P. J. Finding Groups in Data, An Introduction to Cluster Analysis; Wiley: New York, 1989.