Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Objective comparison of protein structures: error-scaled difference distance matrices

Thomas R. Schneider

Department of Structural Chemistry, University of Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany

Correspondence e-mail: trs@shelx.uni-ac.gwdg.de

Received 17 December 1999
Accepted 7 March 2000

Understanding of macromolecular function in many cases relies on the comparison of related structural models. Commonly used least-squares superposition methods suffer from bias introduced into the comparison process by the subjective choice of atoms employed for the superposition. Difference distance matrices are a more objective means of comparing structures as they do not depend on a particular superposition scheme. However, they suffer from very high noise originating from coordinate errors. Modern refinement programs allow the rigorous estimation of standard uncertainties for individual atomic positions. These errors can be propagated through the calculation of a difference distance matrix allowing one to assess the significance level of structural differences. An algorithm is presented which produces an intuitive graphical representation of difference distance matrices after normalization to their error levels. Two examples where its application was revealing are described. Alternatives are suggested for cases where rigorous estimation of individual errors by the inversion of the full least-squares matrix is not feasible. The method offers an unbiased way to detect significant similarities and differences between related structures, as encountered in studies of complexes and mutants or when multiple models are obtained from experiments such as crystal structures involving non-crystallographic symmetry or different crystal modifications, or ensembles derived from NMR spectroscopy.

1. Introduction

With the growing speed of modern macromolecular structure-determination techniques, methods for rapid and objective comparison of structures become ever more important. On one hand, structural homologies in seemingly unrelated proteins have to be detected when novel structures are determined; on the other hand, small conformational differences in closely related structures have to be interpreted when molecular function is elucidated by comparison of complexes and mutants. Furthermore, in the case of NMR investigations or in the presence of non-crystallographic symmetry in a crystal structure determination, the result of the experiment is an ensemble of molecules and valuable information can be derived about the conformational flexibility and rigidity by analysis of this ensemble.

Currently, the majority of structure comparisons are based on the least-squares superposition of specific atoms and a number of implementations are available (*e.g.* Kabsch, 1978; Jones *et al.*, 1991). This approach has been successful in many cases, but suffers from the fact that the results are strongly influenced by the choice of atoms to be superimposed: a wrong

choice of the superposition set may obscure significant and important differences. An iterative procedure that automatically identifies the optimal set of atoms to be used for superposition taking into account the different precision of different atoms in a structure has been presented (Peters-Libeu & Adman, 1997), but was found to still face difficulties when motions of large domains occur. Several approaches that are based on the analysis of backbone torsion angles have been suggested to compare structures of, for example, molecules related by non-crystallographic symmetry (Kleywegt, 1996). Very recently, a method which employs experimental data to calculate electron-density maps whose local density correlation is then used to assess the significance of structural differences has been presented (Kleywegt, 1999). In the present paper, a method for the comparison of models for protein structures is described where the different levels of precision in different models and in different parts of the models are explicitly included in the assessment of whether or not regions in the different models are similar.

When two conformations of a molecule are compared, the central question is whether the *relative positions* of atoms are different. As all structural models contain errors, different in this context means significantly different with respect to the precision of the atomic coordinates in the structural models being compared. For this purpose, the representation of atomic positions in Cartesian coordinates is clearly inappropriate and the distance matrix, *i.e.* the matrix of the distances between all pairs of atoms within a molecule, represents a suitable alternative. The distance matrix of a molecule contains all the information about the geometry of a molecule except its handedness and, if all interatomic distances are known, can be used to reconstruct its three-dimensional structure (Crippen & Havel, 1988). In the case of proteins, the most commonly used distance matrix is that of the pairwise distances between C^{α} atoms (Philips, 1970). $C^{\alpha}-C^{\alpha}$ distance matrices have been used to great effect for the recognition of secondary and tertiary structure (Rossmann & Liljas, 1974; Kuntz, 1975) and for fast scoring of fragments used in model building (Jones & Thirup, 1986).

Matrices constructed as the differences between distance matrices, the so-called difference distance matrices, represent a sensitive and objective measure of differences (and similarities) between related structures. First proposed by Nishikawa & Ooi (1974), difference distance matrices have been used in a variety of contexts. One class of applications concerns the identification of secondary structure and substructures (Richards & Kundrot, 1988) and the detection of structural homologies (Padlan & Davies, 1975; Holm & Sander, 1993) in different molecules. The other class includes the use of difference distance matrices in the comparison of identical molecules under different circumstances, as in the discovery of the thermal expansion of myoglobin (Frauenfelder *et al.*, 1987), the investigation of lysozyme at different hydrostatic pressures (Kundrot & Richards, 1987) or the analysis of molecular-dynamics simulations (Elber & Karplus, 1987). Recently, Nichols and coworkers employed difference distance matrices for the identification of rigid domains (Nichols *et al.*, 1995), forming the basis for the subsequent description of conformational changes (Nichols *et al.*, 1997). There is, however, a serious drawback to difference distance matrices: as their elements represent small differences between relatively large numbers, they are intrinsically noisy.

Modern crystallographic refinement programs deliver, albeit at substantial computational cost, accurate estimates for standard deviations of refined parameters by inversion of the full least-squares matrix (Sheldrick & Schneider, 1997; Tickle *et al.*, 1998). In this paper, we describe how these estimated standard uncertainties (s.u.s) on refined parameters can be propagated through the calculation of a difference distance matrix so that the significance of the differences can be assessed more precisely.

Two examples are described for which the use of error-scaled difference distance matrices was instrumental in defining functional properties of the investigated molecules. In the first example (mersacidin, a 20 amino-acid antibiotic described in the accompanying paper; Schneider *et al.*, 2000), error estimates were available from full-matrix inversion and flexible and rigid parts of the molecule were identified by comparison of six molecules related by non-crystallographic symmetry. In the second example (tryptophan synthase, an enzyme with 660 amino acids), individual positional errors for the atoms were not available. In such a case, in principle, a variety of methods can be used to derive approximate values (*e.g.* Cox & Cruickshank, 1948; Cruickshank, 1949; Murshudov & Dodson, 1997). We have used an approximation based on the recently proposed diffraction precision indicator (DPI; Cruickshank, 1999) to place errors in different structures onto a common scale and have exploited the experimentally observed correlation between s.u.s and B values in order to derive estimates for individual coordinate errors. Inspection of the resulting error-scaled difference distance matrices between different complexes of tryptophan synthase allowed the identification of a rigid, but moveable, domain whose motion plays a central role in the function of the enzyme.

In both cases, the interpretation of the distance matrices in terms of conformationally invariant regions to be used for subsequent least-squares superposition was greatly facilitated by an intuitive graphical representation, which is described in §2.5 of this paper.

2. The method

2.1. Definition of the difference distance matrix

For a given conformation a, the elements of the distance matrix D_{ij}^{a} are the distances between atoms i and j in a molecule,

$$D_{ij}^{a} = |\mathbf{r}_i - \mathbf{r}_j|, \tag{1}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the Cartesian coordinate vectors of atoms i and j. The elements Δ_{ij}^{ab} of the difference-distance matrix for two conformations a and b are

$$\Delta_{ij}^{ab} = D_{ij}^{a} - D_{ij}^{b}. \quad (2)$$

If Δ_{ij}^{ab} is positive, the interatomic vector between i and j in conformation b is contracted with respect to a. Conversely, negative elements of the difference distance matrix indicate an expansion of the interatomic vector. If Δ_{ij}^{ab} is zero for a group of atoms, this group can be considered as conformationally invariant with respect to conformations a and b. Owing to the presence of errors, this condition has to be loosened in practice and conformational invariance is assumed if $|\Delta_{ij}^{ab}|$ is smaller than a certain threshold (*e.g.* Nichols *et al.*, 1995).

2.2. Estimated standard deviations for the elements of a difference distance matrix

Upon convergence of a least-squares refinement of a structural model against crystallographic data, variances and covariances of the refined parameters (typically coordinates and B factors) can be estimated by inversion of the full least-squares matrix and used to calculate estimated standard uncertainties for the refined parameters themselves and derived properties (such as bond lengths and angles; Sands, 1966; Rollett, 1970; Huml, 1980). This requires massive calculations but is feasible with currently available computer programs such as *SHELXL*97 (Sheldrick & Schneider, 1997) or *RESTRAIN* (Tickle *et al.*, 1998) for proteins of modest size. The largest system for which a full-matrix inversion has been reported so far is the 1.25 Å structure of the cholera toxin B-pentamer containing 5 × 103 amino acids corresponding to 6310 atomic sites (Merritt *et al.*, 1998).

To describe the uncertainty in atomic coordinates of individual atoms, the program *SHELXL*97 calculates a 'radial positional error' $\sigma_{r,i}$ for every atom i, taking into account the variance of the coordinate along the crystallographic axes and the covariances (*i.e.* the off-diagonal elements of the inverse of the LS matrix) between them (G. M. Sheldrick, personal communication). Neglecting the covariance terms for atoms i and j, this error estimate can be used to obtain a first-order approximation of the error for the element D_{ij}^{a} of a distance matrix,

$$\sigma(D_{ij}^{a}) = [(\sigma_{r,i}^{a})^2 + (\sigma_{r,j}^{a})^2]^{1/2}. \quad (3)$$

In principle, the expression for $\sigma(D_{ij}^{a})$ can be evaluated rigorously employing the covariance between all refined parameters contributing to the calculation of $\sigma(D_{ij}^{a})$. This would, however, require the full variance/covariance matrix to be accessible at the time when the $\sigma(D_{ij}^{a})$s are calculated.

Given the s.u.s of the elements of the underlying distance matrices $\sigma(D_{ij}^{a})$ and $\sigma(D_{ij}^{b})$, the s.u.s of the elements of a difference distance matrix can be estimated as

$$\sigma(\Delta_{ij}^{ab}) = [\sigma^2(D_{ij}^{a}) + \sigma^2(D_{ij}^{b})]^{1/2}, \quad (4)$$

again neglecting the covariances between the contributions.

Combining (3) and (4), a simple expression for the estimated standard deviation of an element of the difference distance matrix is obtained,

$$\sigma(\Delta_{ij}^{ab}) = [(\sigma_{r,i}^{a})^2 + (\sigma_{r,j}^{a})^2 + (\sigma_{r,i}^{b})^2 + (\sigma_{r,i}^{b})^2]^{1/2}. \quad (5)$$

This expression can be evaluated without the full matrix being available at the time of its calculation.

2.3. Error estimates in the absence of calculated s.u.s

In many cases, the inversion of the full least-squares matrix is not feasible owing to its sheer size, and estimates of the coordinate error of individual atoms have to be derived by other means. Commonly, estimates for the mean coordinate error in a crystal structure are determined using the methods of Luzzati (1952); for a critical discussion of this practice, see Cruickshank (1999) or Read (1986).

Recently, Cruickshank (1999) has reviewed the problem of assigning accurate uncertainties to atomic coordinates in macromolecular crystal structures. He presented an empirical formula that describes the influence of the number of fully occupied atomic sites N_i, the number of reflections $n_{\rm obs}$ employed in refinement, the completeness C and the maximum resolution $d_{\rm min}$ of the diffraction data and the final value of $R_{\rm free}$ on the *positional* error $\sigma_r^{\rm DPI}(B_{\rm avg})$ of an atom with the average B factor $B_{\rm avg}$ of the model,

$$\sigma_r^{\rm DPI}(B_{\rm avg}) = 3^{1/2}\sigma_x^{\rm DPI}(B_{\rm avg}) = 3^{1/2}(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\rm min}, \quad (6)$$

where DPI stands for 'diffraction-component precision index' and $\sigma_x^{\rm DPI}(B_{\rm avg})$ stands for the corresponding *coordinate* error. Note that this formula does not give estimates for the absolute errors of individual atomic coordinates, but allows the errors in different models to be put onto a common absolute scale to which the individual positional errors can be related.

Studies of coordinate precision in crystal structures have invariably identified the atomic B factor to be highly correlated with the coordinate error (Chambers & Stroud, 1979; Daopin *et al.*, 1994; Stroud & Fauman, 1995; Tickle *et al.*, 1998; Cruickshank, 1999; Parisini *et al.*, 1999). Different parametrizations have been described, the general result being that the higher the B value, the larger the coordinate uncertainty. Therefore, in a first-order approximation, we can assume a linear relation between the estimated positional error $\tilde{\sigma}_{r,i}$ and the B factor B_i of an atom i of the form

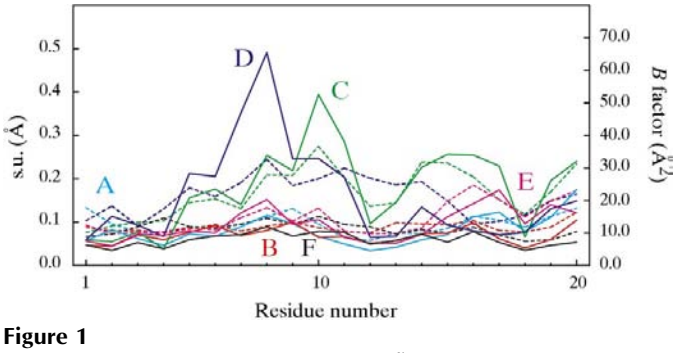

Figure 1
Radial positional error (full lines) for C^{α} atoms for six molecules of mersacidin as determined by *SHELXL*97 after inversion of the full least-squares matrix. Values for different molecules are shown in different colours; assignment is given in the figure. The broken lines represent the B values of the corresponding atoms.

$$\tilde{\sigma}_{r,i} = \frac{\sigma_r^{\mathrm{DPI}}(B_{\mathrm{avg}})}{B_{\mathrm{avg}}} B_i, \qquad (7)$$

where correct normalization is achieved through division by B_{avg}. The values obtained for $\tilde{\sigma}_{r,i}$ can then be used to substitute the s.u.s, $\sigma^a_{r,i}$ *etc.* in (5).

2.4. Error-scaling of difference distance matrices

Prior to display, each element of the difference distance matrix Δ^{ab}_{ij} is normalized by dividing it by its error $\sigma(\Delta^{ab}_{ij})$ to obtain the elements of the error-scaled difference distance matrix,

$$E^{ab}_{ij} = \Delta^{ab}_{ij} / \sigma(\Delta^{ab}_{ij}). \qquad (8)$$

The elements of this matrix are a measure of the significance of a change in a distance between two atoms i and j for two structural models a and b.

2.5. Representation of distance matrices

Difference distance matrices can be immense in size and of overwhelming complexity, necessitating an intuitive representation to expedite their interpretation. We have devised (Schneider, 1996) a scheme in which the matrix is shown as a two-dimensional plot. Each element of the matrix is displayed as a colour-coded square to represent the change in distance between two atoms: blue stands for a positive (contraction) and red for a negative (expansion) change in the length of the corresponding interatomic vector. Absolute values are indicated by the intensity of the colour, which is ramped from light to full, full colours representing larger values. The lowest and the highest values to be displayed are user-defined in order to suppress noise and use the dynamic range of the colour-ramping scheme optimally.

The maximum size of a matrix that can be displayed on an A4 sheet of paper while still allowing clear visual inspection is about 150 × 150 elements. Therefore, for proteins containing more than a user-defined number of of amino acids, the matrix undergoes a binning procedure before being displayed: a binning factor N is chosen such that the dimension of the resulting matrix is less than the user-defined limit; then $N \times N$ submatrices are collapsed to the element with the largest absolute value; finally, this value is stored as an element of the binned matrix. Using this procedure, all relevant information about differences between structures is maintained while presenting the information in a more digestible format. Subsequently, selected regions of the original matrix can be inspected at full resolution.

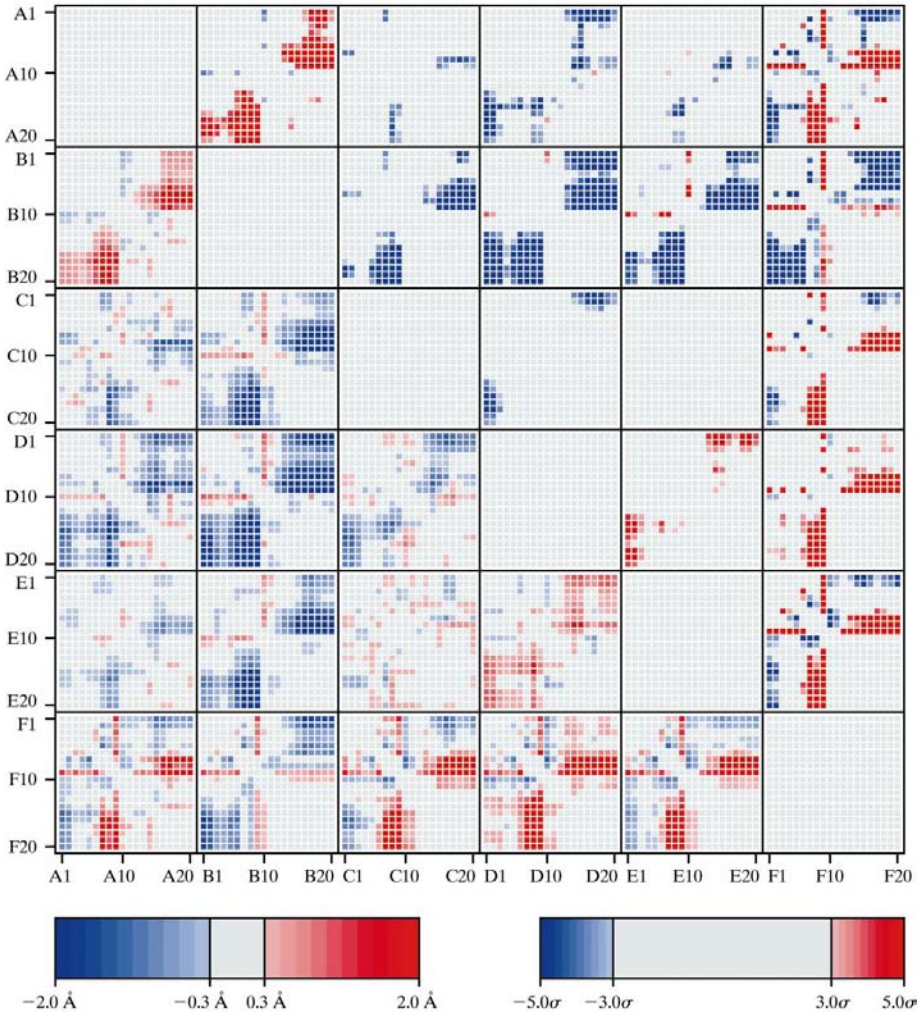

Figure 2
Difference distance matrices and error-scaled difference distance matrices for the C^{α} atoms of the six molecules of mersacidin. In the lower left triangle, ordinary difference distance matrices for all pairs of NCS copies are shown. The colour coding is according to the bar on the lower left: all changes in distances smaller than 0.3 Å are shown in grey; differences in distances between 0.3 and 2.0 are shown using a colour gradient, where red stands for expansion and blue for contraction, light colours represent small changes and dark colours large changes; all differences larger than 2.0 Å are shown as full blue and full red, respectively. The blocks in the upper right triangle show the error-scaled difference distance matrices for all pairs of molecules. Here, all differences lower than $3.0\sigma(\Delta^{ab}_{ij})$ are mapped to grey. Changes greater than $3.0\sigma(\Delta^{ab}_{ij})$ and smaller than $5.0\sigma(\Delta^{ab}_{ij})$ are colour coded using a scheme analogous to that used for ordinary difference distance matrices.

3. First example: flexibility and rigidity in mersacidin

Mersacidin is a polypeptide antibiotic containing 20 amino acids that crystallizes with six molecules in the asymmetric unit. The structure was solved and refined against merohedrally twinned data to 1.06 Å resolution

as described in the accompanying paper (Schneider *et al.*, 2000). The conformations of the six molecules are similar to each other, with a mean r.m.s.d. for all 15 possible pairwise least-squares superpositions (using all C^{α} atoms) of 0.83 Å as calculated by *LSQKAB* (Kabsch, 1978). Analysis of the superimposed molecules to identify rigid and flexible regions was inconclusive.

With 826 non-H atoms corresponding to 7438 parameters refined against 33 449 diffraction intensities, the inversion of the full least-squares matrix for this structure is feasible using currently available computers (the matrix inversion takes about 5.8 h of CPU time on a Pentium II CPU running at 450 MHz and requires a total of 112 MB of memory). The mean positional error for C^{α} atoms is 0.11 ± 0.08 Å, with a minimum of 0.034 Å for C^{α}(A12) and a maximum of 0.49 Å for C^{α}(D8). Most of the s.u.s fall within the range 0.05–0.15 Å (Fig. 1). These values are high for a structure refined at atomic resolution and reflect the loss in precision owing to merohedral twinning and the intrinsic flexibility of parts of the structure. Elevated estimated standard uncertainties higher than 0.2 Å and ranging up to 0.5 Å are found for parts of molecules *C* and *D*. Fig. 1 also shows a remarkably good linear correlation between *B* values and coordinate errors, which however breaks down for very high *B* values.

Difference distance matrices displayed with a lower cutoff of 0.3 Å gave rise to the suspicion that the C-terminal half of the molecule was more rigid than the N-terminal half (Fig. 2, lower left half): the matrices between molecules *C* and *B*, between molecules *F* and *B* and between *C* and *E* molecules show differences smaller than 0.3 Å for the interatomic distances within the C-terminal half of the molecules. Qualitatively, this behaviour can also be found in the other matrices, but they are too noisy to allow a firm conclusion.

Calculation of error-scaled difference distance matrices considerably improved the situation (Fig. 2, upper right half). If a lower cutoff of $3.0\sigma(\Delta_{ij}^{ab})$ is employed for display, the noise is dramatically reduced while maintaining sufficient signal, now supporting the hypothesis of conformational invariance for the C-terminus with respect to all pairs of molecules except the combinations $A \leftrightarrow B$, $A \leftrightarrow D$ and $A \leftrightarrow F$. It should be noted that with regard to the pairs found to define conformationally invariant regions already from ordinary difference distance matrices ($C \leftrightarrow B$, $F \leftrightarrow B$, $F \leftrightarrow C$, $F \leftrightarrow E$) the situation does not change after error scaling. The most pronounced effect of error scaling is seen for the comparison of molecules *C* and *E*: ordinary difference distance matrices had shown a large number of differences ranging up to 1 Å, but after error scaling it becomes apparent that both molecules are in fact identical within errors. This may in part be because of the particularly large errors observed for molecule *C*, which require the conformational difference to be large to be significant. Interestingly, the error-scaled matrix between the two least well defined molecules, *C* and *D*, still shows significant features. These features involve the N-termini of both molecules, which are in fact the relatively best defined (Fig. 1) regions of molecules *C* and *D*, owing to the monosulfide bridge present between residues Cys1 and Aba2. For such well defined parts, smaller differences in coordinates are sufficient to be significant.

Based on the error-scaled difference distance matrices, residues 12–20 of mersacidin can be regarded as forming a rigid domain. Least-squares superposition, employing C^{α} atoms of residues 12–20, gave r.m.s.d.s between superimposed C^{α} atoms ranging from 0.11 to 0.36 Å, with a mean value of 0.21 Å. The superimposed molecules provide an interesting view, clearly dividing the molecule into a rigid and a flexible part (Schneider *et al.*, 2000). In fact, the C-terminal region has also been found to be rigid in an NMR study (Prasch *et al.*, 1997) and a highly homologous region is found in the related molecule actagardine (Zimmermann & Jung, 1997), suggesting a functional role.

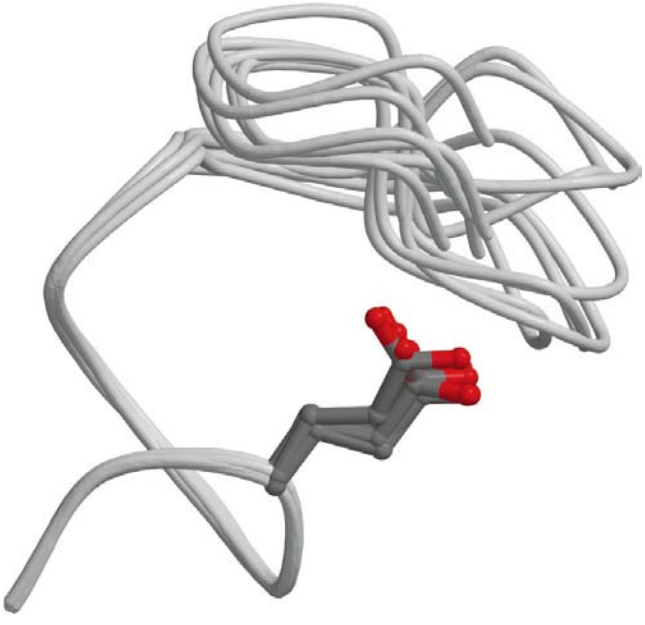

Figure 3
Least-squares superposition of six molecules of mersacidin based on the C^{α} atoms of residues 12–20. The backbone is shown in light grey and the side chain of Glu17 in dark grey and red. The figure was drawn with *MOLSCRIPT* (Kraulis, 1991) and *Raster3D* (Bacon & Anderson, 1988; Merritt & Murphy, 1994).

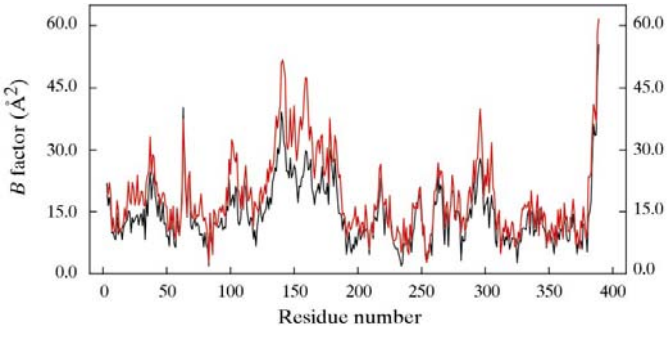

Figure 4
B values for C^{α} atoms in the crystal structure of tryptophan synthase in complex with F-IPP ($TRPS^{F\text{-}IPP}$) in black and in complex with F-IPP and amino-acrylate ($TRPS^{F\text{-}IPP}_{A\text{-}A}$) in red.

4. Second example: domain motion in tryptophan synthase

Tryptophan synthase catalyses the last two reactions in the biosynthesis of tryptophan, the cleavage of indole 3-glycerol phosphate (IGP) to indole and glyceraldehyde 3-phosphate (α-reaction) and the subsequent condensation of indole with serine to form tryptophan (β-reaction) (Hyde & Miles, 1990). The reactions take place at two active centres which are separated by a distance of more than 25 Å, but nevertheless are precisely synchronized (Anderson *et al.*, 1991). In a study aimed at the understanding of the interaction between the two active sites, crystal structures of the enzyme in complex with the substrate analogue 5-fluoroindole propanol phosphate ($TRPS^{F\text{-}IPP}$; PDB entry 1a50) and in complex with both F-IPP and L-serine ($TRPS^{F\text{-}IPP}_{A\text{-}A}$, where 'A-A' stands for the amino-acrylate that is formed at the β-site under the experimental conditions chosen; PDB entry 1a5s) were determined (Schneider *et al.*, 1998).

Both crystal structures have similar statistics (see caption for Fig. 5) and a least-squares superposition based on C^{α} atoms gave an r.m.s.d. of 0.28 Å, which is very close to the mean positional errors as determined by the σ_A method (Read, 1986), 0.21 Å for $TRPS^{F\text{-}IPP}$ and 0.27 Å for $TRPS^{F\text{-}IPP}_{A\text{-}A}$ (Schneider *et al.*, 1998). The least-squares superposition did not reveal any specific features. Ordinary difference distance matrices (Fig. 5, lower left half) were decidedly noisy.

As no individual s.u.s for atomic coordinates were available from the refinement, estimates for the radial error of an atom with an average B factor, $\sigma_r^{DPI}(B_{avg})$, were determined using Cruickshank's formalism. The values obtained for $\sigma_r^{DPI}(B_{avg})$ were 0.21 and 0.24 Å for $TRPS^{F\text{-}IPP}$ and $TRPS^{F\text{-}IPP}_{A\text{-}A}$, respectively, which is in very good agreement with error estimates derived by the σ_A method (see above). Standard uncertainties for individual atoms were then estimated following (7) after replacing B factors of lower than 10 $Å^2$ by a value of 10 $Å^2$ in order to avoid unrealistically low error estimates stemming from underestimated B factors. The individual s.u.s have mean values of 0.22 and 0.24 Å, respectively. They range from 0.14 to 0.80 Å and from 0.24 to 0.79 Å for $TRPS^{F\text{-}IPP}$ and $TRPS^{F\text{-}IPP}_{A\text{-}A}$, respectively. The resulting error-scaled difference matrix for residues 3–389 of the β-subunit of tryptophan synthase was displayed after 3×3 binning and revealed a much clearer picture (Fig. 5, upper right half). The region corresponding to residues 125–180 is of particular interest, as this region of the molecule has elevated B values in both crystal structures (Fig. 4). Consequently, the corresponding part or the normal difference distance matrix is quite noisy. Translating the B values of the C^{α} atoms into approximate standard uncertainties (7) and employing these for the calculation of the error-scaled difference distance matrix flattens out this part while preserving the signal in the rest of the matrix. The presence of large 'empty' blocks along the diagonal proves the presence of three conformationally invariant regions I, II and III, comprising residues 3–101, 102–189 and 190–389, respectively. Furthermore, regions I and III do not move relative to one another, whereas region II moves

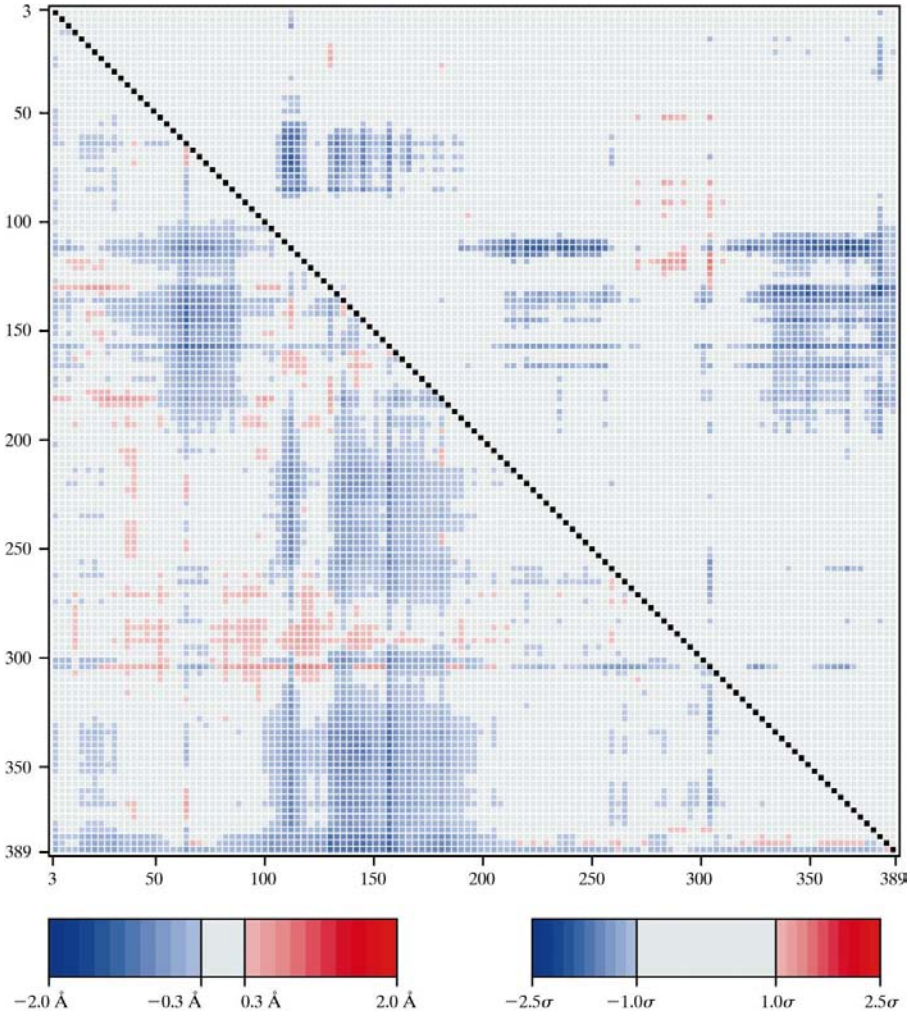

Figure 5
Difference distance matrix between the structures of tryptophan synthase in complex with F-IPP ($TRPS^{F\text{-}IPP}$) and in complex with F-IPP and amino-acrylate ($TRPS^{F\text{-}IPP}_{A\text{-}A}$). In the lower left half, the ordinary difference matrix is displayed using a lower cutoff of 0.25 Å, approximately corresponding to the 1σ level as determined by the σ_A method. For scaling, an upper cutoff of 2.0 Å has been employed. In the upper right half, the error-scaled difference distance matrix is displayed using upper and lower cutoffs of 1.0 and 2.5σ, respectively. Both matrices underwent 3×3 binning prior to being displayed. Individual coordinate errors were estimated using (7) employing the following values for the parameters: $TRPS^{F\text{-}IPP}$: $N_i = 5191$, $n_{obs} = 31627$, $C = 0.965$, $R_{free} = 0.221$, $d_{min} = 2.29$ Å; $TRPS^{F\text{-}IPP}_{A\text{-}A}$: $N_i = 5148$, $n_{obs} = 30327$, $C = 0.938$, $R_{free} = 0.247$, $d_{min} = 2.30$ Å.

closer to both region I and III in $TRPS^{F\text{-}IPP}_{A\text{-}A}$ with respect to $TRPS^{F\text{-}IPP}$, indicated by the blue blocks in Fig. 3. Interestingly, both residues at the border between regions I and II and II and III are glycines that act as hinges to support a rigid-body motion of domain II upon formation of the reactive intermediate at the β-site. Superposition of $TRPS^{F\text{-}IPP}$ and $TRPS^{F\text{-}IPP}_{A\text{-}A}$ based on 557 C^{α} atoms from all residues except those belonging to region II results in a slightly reduced r.m.s.d. of 0.24 Å; the graphical representation clearly shows the motion of region II relative to the rest of the protein (Fig. 6). This domain movement triggered by structural changes at the β-site facilitates the transmission of information to the α-site across a distance of more than 25 Å (Schneider *et al.*, 1998).

5. Conclusions and perspectives

Difference distance matrices are a useful tool for deriving conformationally invariant regions by comparison of structures of related molecules. If estimates of individual coordinate errors are available, these errors can be propagated through the calculation and used to significantly reduce the noise in difference distance matrices.

In principle, the error of any interatomic distance in a structural model derived from crystallographic data can be rigorously determined by inversion of the full least-squares matrix and taking into account the variances and covariances of all refined parameters contributing to the calculation of that particular distance (Sands, 1966; Huml, 1987). However, in practice, there are two complications, the first being that inversion of the full-matrix is not feasible in many cases owing to the sheer size of the computational task. The second complication concerns the use of covariances between refined parameters: if covariances are included in the estimation of uncertainties, the full variance/covariance matrix has to be available when the calculations are performed. This requires computer memory of the same size as is necessary for the matrix inversion itself. Both problems will be alleviated with increasing computer power. In the method presented, the covariances between refined parameters are neglected. Inclusion of covariances in the future will provide more accurate error estimates, which in turn will lead to a higher accuracy of the resulting error-scaled difference matrices.

Meanwhile, in cases where matrix inversion is too computationally expensive, approximations can be made to derive approximate positional errors for individual atoms. Several approaches based on analysis of the refinement process have been described (Cox & Cruickshank, 1948; Cruickshank, 1949, 1999; Murshudov & Dodson, 1997) and measures of relative precision of atomic coordinates based on the analysis of electron densities in terms of a real-space correlation coefficient (Zhou *et al.*, 1998) may represent useful alternatives.

Once the uncertainties in the coordinates of individual atoms have been estimated, they can, independently of their source and size, be propagated through the calculation in a consistent manner, finally yielding an error-scaled difference distance matrix. In particular, error propagation allows the straightforward comparison of models with very different levels of precision: for example, a structure determined at atomic resolution with s.u.s obtained by full-matrix inversion can be compared with a structure where the experimental data are not available and the level of precision of individual atomic coordinates has to be derived by some rough approximation.

To rationalize the enormous amount of information, an intuitive graphical interpretation plays a central role. The present implementation allows on-line variation of error models and scaling parameters to clarify the picture. In fact, the interpretation of difference distance matrices can be regarded as a process similar to the interpretation of electron densities in crystallographic model building: even if the scale used to present the information is not correct in an absolute sense, correct relative scaling allows to enhance important features, hence facilitating interpretation.

The applications discussed in the present paper pertain to comparison of C^{α} positions in protein molecules with identical sequences where the correspondence of pairs of residues is clearly defined. In cases where the sequences are not 100% identical but still closely related, for example in the presence of point mutations or short insertions or deletions, non-equivalent regions can be manually excluded from the comparison process to allow an assignment of corresponding residues.

In addition to comparisons of the polypeptide backbone, the algorithm is equally applicable to any groups of atoms, for example the atoms surrounding an active site, where the discrimination between significant and insignificant changes upon binding of a ligand or the mutation of an amino acid can be of importance in understanding functional aspects.

Algorithms for automatic identification of conformationally invariant parts of molecules based on difference distance matrices have been described (Nichols *et al.*, 1995; Perry *et al.*, 1990) and clearly profit from a more rigorous treatment of errors. Once identified, conformationally invariant regions can be used for automatic superposition of molecules, which in

Figure 6
Backbone traces of $TRPS^{F\text{-}IPP}$ and $TRPS^{F\text{-}IPP}_{A\text{-}A}$ after least-squares superposition based on all C^{α} atoms except those of residues Gly102–Gly189 of the β chain. Residues Gly102–Gly189 of the β-chain are shown as thick lines. The α-site inhibitor F-IPP and the β-site cofactor pyridoxal 5′-phosphate are shown in grey to indicate the location of the active sites. Backbone traces were drawn with *MOLSCRIPT* (Kraulis, 1991).

turn enables the detection of significant differences. Eventually, an algorithmic definition of a superposition procedure taking experimental errors into account rigorously both in the selection of the atoms used for superposition and in the superposition process itself could allow an automatic and thus objective comparison of macromolecular structures.

I am grateful to George M. Sheldrick for helpful discussions and encouragement. A beta test version of the program *ESCET* to calculate and display error-scaled difference distance matrices is available from the author upon request.

References

Anderson, K. S., Miles, E. W. & Johnson, K. A. (1991). *J. Biol. Chem.* **266**, 8020–8033.

Bacon, D. J. & Anderson, W. F. (1988). *J. Mol. Graph.* **6**, 219–220.

Chambers, J. L. & Stroud, R. M. (1979). *Acta Cryst.* B**35**, 1861–1874.

Cox, E. G. & Cruickshank, D. W. J. (1948). *Acta Cryst.* **1**, 92–93.

Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation.* New York: John Wiley & Sons Inc.

Cruickshank, D. W. J. (1949). *Acta Cryst.* **2**, 154–157.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Daopin, S., Davies, D., Schlunegger, M. & Grütter, M. (1994). *Acta Cryst.* D**50**, 85–92.

Elber, R. & Karplus, M. (1987). *Science*, **235**, 318–321.

Frauenfelder, H., Hartmann, H., Karplus, M., Kuntz, I. D. Jr, Kuriyan, J., Parak, F., Petsko, G. A., Ringe, D., Tilton, R. F. Jr, Conolly, M. L. & Max, N. (1987). *Biochemistry*, **26**, 254–261.

Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123–138.

Huml, K. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 12.01–12.22. Bangalore, India: The Indian Academy of Sciences.

Hyde, C. C. & Miles, E. W. (1990). *Biotechnology*, **8**, 27–32.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.

Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.

Kleywegt, G. J. (1999). *Acta Cryst.* D**55**, 1878–1884.

Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.

Kundrot, C. E. & Richards, F. M. (1987). *J. Mol. Biol.* **193**, 157–170.

Kuntz, I. D. (1975). *J. Am. Chem. Soc.* **97**, 4362–4365.

Luzzati, P. V. (1952). *Acta Cryst.* **5**, 802–810.

Merritt, E. A., Kuhn, P., Sarfaty, S., Erbe, J. L., Holmes, R. K. & Hol, W. G. J. (1998). *J. Mol. Biol.* **282**, 1043–1059.

Merritt, E. A. & Murphy, M. E. P. (1994). *Acta Cryst.* D**50**, 869–873.

Murshudov, G. N. & Dodson, E. J. (1997). *CCP4 Newslett. Protein Crystallogr.* **33**, 31–39.

Nichols, W. L., Rose, G. D., Ten Eyck, L. & Zimm, B. H. (1995). *Proteins Struct. Funct. Genet.* **23**, 38–48.

Nichols, W. L., Zimm, B. H. & Ten Eyck, L. (1997). *J. Mol. Biol.* **270**, 598–615.

Nishikawa, K. & Ooi, T. (1974). *J. Theor. Biol.* **43**, 351–374.

Padlan, E. A. & Davies, D. R. (1975). *Proc. Natl Acad. Sci. USA*, **72**, 819–823.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.

Perry, K. M., Faumann, E. B., Finer-Moore, J. S., Montfort, W. R., Maley, G. F., Maley, F. & Stroud, R. M. (1990). *Proteins Struct. Funct. Genet.* **8**, 315–333.

Peters-Libeu, C. & Adman, E. T. (1997). *Acta Cryst.* D**53**, 56–76.

Philips, D. C. (1970). *Biochem. Soc. Symp.* **30**, 11–28.

Prasch, T., Naumann, T., Markert, R. L., Sattler, M., Schubert, W., Schaal, S., Bauch, M., Kogler, H. & Griesinger, C. (1997). *Eur. J. Biochem.* **244**, 501–512.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Richards, F. M. & Kundrot, C. (1988). *Proteins*, **3**, 71–84.

Rollett, J. S. (1970). *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 167–181. Copenhagen: Munksgaard.

Rossman, M. G. & Liljas, A. (1974). *J. Mol. Biol.* **85**, 177–181.

Sands, D. E. (1966). *Acta Cryst.* **21**, 868–872.

Schneider, T. R. (1996). *Röntgenkristallographische Untersuchung der Struktur und Dynamik einer Serinprotease.* PhD thesis, Technical University of Munich, Germany.

Schneider, T. R., Gerhardt, E., Lee, M., Lian, P., Anderson, K. S. & Schlichting, I. (1998). *Biochemistry*, **37**, 5394–5406.

Schneider, T. R., Kärcher, J., Pohl, E., Lubini, P. & Sheldrick, G. M. (2000). *Acta Cryst.* D**56**, 705–713.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.

Stroud, R. M. & Fauman, E. B. (1995). *Protein Sci.* **4**, 2392–2404.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 243–252.

Zhou, G., Wang, J., Blanc, E. & Chapman, M. S. (1998). *Acta Cryst.* D**54**, 391–399.

Zimmermann, N. & Jung, G. (1997). *Eur. J. Biochem.* **246**, 809–819.