

Automated matching of high- and low-resolution structural models

M. B. Kozin and D. I. Svergun

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site provided that this cover page is retained. Republication of this article or its storage in electronic databases or the like is not permitted without prior permission in writing from the IUCr.

Automated matching of high- and low-resolution structural models

M. B. Kozin^{a,b} and D. I. Svergun^{a,b*}Received 15 July 2000
Accepted 12 October 2000^aInstitute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia, and ^bEMBL, Hamburg Outstation, Notkestraße 85, D-22603 Hamburg, Germany.
Correspondence e-mail: svergun@embl-hamburg.de

A method is presented for automated best-matching alignment of three-dimensional models represented by ensembles of points. A normalized spatial discrepancy (NSD) is introduced as a proximity measure between three-dimensional objects. Starting from an inertia-axes alignment, the algorithm minimizes the NSD; the final value of the NSD provides a quantitative estimate of similarity between the objects. The method is implemented in a computer program. Simulations have been performed to test its performance on model structures with specified numbers of points ranging from a few to a few thousand. The method can be used for comparative analysis of structural models obtained by different methods, *e.g.* of high-resolution crystallographic atomic structures and low-resolution models from solution scattering or electron microscopy.

© 2001 International Union of Crystallography
Printed in Great Britain – all rights reserved

1. Introduction

The problem of finding the best-matching superposition of three-dimensional structures is central in pattern recognition and computer vision, and in docking and classification studies in structural biology. Considering only rigid-body transformations, the problem can be formulated as follows: given the two sets of points (*e.g.* representing two structural models), find the rotation and movement of the second set that minimizes a discrepancy between the two sets. A number of theoretical and applied approaches proposed recently for solving this computationally intensive problem can be subdivided into several groups (Loncaric, 1998; Veltkamp & Hagedoorn, 1999).

The methods of the first group aim at matching corresponding features of the two models. The simplest case implies that every point in one set has a mate in the other (Kabsch, 1978; Heffernan & Schirra, 1994; de Reezende & Lee, 1995). More generally, features like control points (Ton & Jain, 1989), shape segments, contours, *etc.* (Mount *et al.*, 1999) are matched. The latter approach has been most intensively used in protein–ligand docking (Peters *et al.*, 1996; Rarey *et al.*, 1997), recognition of topological similarities (Vriend & Sander, 1991; Diederichs, 1995), structure classification and database design (Gilbert *et al.*, 1999; Greaves *et al.*, 1999). Enhanced comparison tools for protein structures allow alignment of structures with uneven numbers of feature points (*LSQMAN*; Kleywegt & Jones, 1995) and freely permuted equivalenced segments (*DALI*; Holm & Sander, 1997).

The feature extraction process could be computationally expensive and in some cases it is hardly applicable at all, for example, if an atomic model is compared with a low-resolution

model. Performing comparisons between models of different nature and resolution permits the cross-validation of structural results obtained by different techniques. Such comparisons are further important for understanding the structure–function relationship, as the crystal structure of a macromolecule may differ from the structure in solution (Svergun *et al.*, 1997, 2000). Optimal superposition of heterogeneous models is not a simple task, especially for larger molecules, because of principal differences in the geometric nature of the objects to be compared. Low-resolution models are represented, for example, as smooth surfaces defined by an angular envelope function (Svergun, 1994), as sets of densely packed dummy atoms (Chacon *et al.*, 1998; Svergun, 1999) or as stacks of contoured layers (Frank *et al.*, 1995). These models display no domains, C_α backbones or other characteristic features of atomic structures. There is usually no *a priori* information about the correspondence between specific elements of high- and low-resolution models. Feature points could be extracted by wavelet decomposition of the images (Mount *et al.*, 1999), but this approach is not sufficiently accurate. The methods employing affine invariants (Mundy & Zisserman, 1992) are usually limited to planar objects, and they provide only a measure of similarity without constructing the transformation itself.

The superposition methods dealing directly with data sets rather than with feature points are more useful in this case. The well known alignment of the principal axes of the inertia tensor (Holupka & Kooy, 1992; Galvez & Canton, 1993) may yield ambiguous results, especially for objects with twofold symmetry axes (Galvez & Canton, 1993). Further, this method provides no quantitative estimate of the similarity of the objects. A similar technique employing second-order moment

invariants on the spherical harmonics basis (Burel & Hugues, 1995) is applicable to a limited class of objects only. A neural-network approach implemented in the *SITUS* program package by Wriggers *et al.* (1999) has a profound theoretical background, consists of several independent program modules and is well suited to high-resolution structures docking into low-resolution maps, especially those obtained by electron microscopy.

Our aim is to create a simple, reliable and time-efficient algorithm not only for best matching the models, but also for providing a quantitative estimate of their similarity. For this, we introduce a proximity measure between objects represented as ensembles of points in three-dimensional space. This measure is stable with respect to noise and is normalized to be independent of size and geometric nature of the models. The inertia-axes alignment is used as a starting approximation for minimizing the proximity measure, the value of which at the end of minimization quantifies the similarity between the objects. The efficiency and limitations of the method are analysed by its application to several low- and high-resolution structural models.

2. Normalized spatial discrepancy between three-dimensional point sets

A proper choice of quantitative similarity measure between sets of three-dimensional points is a necessary prerequisite for a reliable best-matching algorithm. To be an analogue to a standard Euclidian distance between points, this measure should obey the three distance axioms (see Appendix A). In practice, it is more important for the similarity measure to be smooth and stable than to obey the formal metric axioms (Hagedoorn & Veltkamp, 1999). The *directed Hausdorff distance* $h(S_1, S_2)$ between two point sets S_1 and S_2 is defined as the maximum over the distances of each point in S_1 to the set S_2 . The latter is defined as the minimal distance to the points in S_2 :

$$h(S_1, S_2) = \max_{a \in S_1} \min_{b \in S_2} \|a - b\|. \quad (1)$$

The *standard Hausdorff distance*, defined by Huttenlocher *et al.* (1993) as a maximum of the two directed distances, is frequently used. The latter measure obeys the metric axioms but is not stable to noise, being sensitive to outlying points (Veltkamp & Hagedoorn, 1999). A *directed partial Hausdorff distance* $H_k(S_1, S_2)$, equal to the k th maximal value in (1) rather than to the maximal value itself, is of more practical value. Although being non-symmetric [$H_k(S_1, S_2) \neq H_k(S_2, S_1)$], it is more stable to noise and distortions (Mount *et al.*, 1999). A *partial Hausdorff distance* defined as the maximum of the two directed partial Hausdorff distances $H_k(S_1, S_2)$ and $H_k(S_2, S_1)$ is used in matching images under homothetic transformations (translation and scaling), although this distance is not a metric because it fails the triangle inequality test (Huttenlocher *et al.*, 1993). A *symmetric difference* measure of similarity of convex shapes (Alt *et al.*, 1996) defines a noise-stable metric. Hagedoorn &

Veltkamp (1999) proposed a variation of the symmetrical difference for arbitrary point sets. Walther *et al.* (2000) used a discrete analog of this measure to resolve the ambiguity in inertia-axes superposition of low-resolution models and atomic structures taken from the Protein Data Bank (PDB; Bernstein *et al.*, 1977).

The above measures are not convenient for our purposes for various different reasons (instability, asymmetry, dependence on the nature and scale of the objects). We introduce a *normalized spatial discrepancy* (NSD) as follows. For every point s_{1i} from the set $S_1 = \{s_{1i}, i = 1, \dots, N_1\}$, the minimum value among the distances between s_{1i} and all points in the set $S_2 = \{s_{2i}, i = 1, \dots, N_2\}$ is denoted as $\rho(s_{1i}, S_2)$. The NSD between the sets S_1 and S_2 is defined as a normalized average,

$$\rho(S_1, S_2) = \left\{ (1/2) \left[(1/N_1 d_2^2) \sum_{i=1}^{N_1} \rho^2(s_{1i}, S_2) + (1/N_2 d_1^2) \sum_{i=1}^{N_2} \rho^2(s_{2i}, S_1) \right] \right\}^{1/2}, \quad (2)$$

where N_i is the number of points in S_i and the *fitness* d_i is the average distance between the neighbouring points in S_i . The NSD is a modification of the distance employed by Bloch *et al.* (1993) for matching three-dimensional convex polyhedra. Provided that the fitness parameters are computed in advance, the NSD is evaluated in $O(N_1 N_2)$ time. This measure is symmetric, independent of the size of the objects as a result of the normalization, and is stable to the outlying points as a result of the averaging. For ideally superimposed similar objects, NSD tends to 0; it exceeds 1 if the objects systematically differ from one another. NSD is not a metric, because only two of the three metric axioms are fulfilled. It is obvious that $\rho(S_1, S_2) = 0$ if and only if $S_1 = S_2$, and that $\rho(S_1, S_2) = \rho(S_2, S_1)$, but the triangle inequality is not always true (a counter example is presented in Appendix A). The following property is more important in practice: denote by $D(\mathbf{T}, \alpha, \beta, \gamma)$ the matrix of a rigid-body transformation ($\det D = 1$), determined by the translation vector $\mathbf{T} = (T_x, T_y, T_z)$ and the Euler rotation angles α, β, γ . Then it can be proved (Bloch, 1990; Bloch *et al.*, 1993) that the function $\rho[S_1, D(S_2)]$ behaves smoothly with respect to the six parameters $T_x, T_y, T_z, \alpha, \beta$ and γ .

3. The superposition algorithm

Numerical minimization of any proximity measure between three-dimensional objects with respect to the positional and especially rotational parameters is known to be a difficult task. Function (2) displays multiple local minima. Minimization algorithms tend to converge to the local minimum near the starting point. As the use of a global minimization algorithm or of an exhaustive six-dimensional space search would require too much computing time, we employ a local minimization starting from the position provided by the inertia-axes matching.

The principal axes of inertia are found for both objects as the eigenvectors of the inertia tensor

$$I = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{xy} & I_{yy} & -I_{yz} \\ -I_{xz} & -I_{yz} & I_{zz} \end{bmatrix} \quad \begin{array}{l} I_{xx} = \mu_{020} + \mu_{002} \\ I_{yy} = \mu_{200} + \mu_{002} \\ I_{zz} = \mu_{200} + \mu_{200} \end{array} \quad \begin{array}{l} I_{xy} = \mu_{110} \\ I_{xz} = \mu_{101} \\ I_{yz} = \mu_{011} \end{array} \quad (3)$$

where

$$\mu_{ijl} = (1/N_k) \sum_{q=1}^{N_k} (x_{kq} - x_k^0)^i (y_{kq} - y_k^0)^j (z_{kq} - z_k^0)^l, \quad (4)$$

$$i + j + l = 2, \quad k = 1, 2,$$

are the second central moments of distribution around the centroid specified by

$$x_k^0 = (1/N_k) \sum_{q=1}^{N_k} x_{kq}, \quad y_k^0 = (1/N_k) \sum_{q=1}^{N_k} y_{kq}, \quad (5)$$

$$z_k^0 = (1/N_k) \sum_{q=1}^{N_k} z_{kq}, \quad k = 1, 2.$$

Here, I_k is a symmetric matrix with real eigenvalues $\lambda_{k1} \geq \lambda_{k2} \geq \lambda_{k3}$ and corresponding orthonormal eigenvectors \mathbf{v}_{k1} , \mathbf{v}_{k2} , \mathbf{v}_{k3} . An object is said to be in a canonical position if it is origin-centred (that is, shifted by the vector $\mathbf{T}_k^0 = \{-x_k^0, -y_k^0, -z_k^0\}$) and rotated so that its principal inertia axes taken in ascending order of eigenvalues are aligned along the X , Y and Z axes, respectively. The rotation matrix is $[\mathbf{M}_k]^T$, where \mathbf{M}_k is composed by columns from the eigenvectors in ascending order of eigenvalues. Up to possible column permutations, \mathbf{M}_k is equal to the matrix of diagonalizing transformation. A negative determinant $\det(\mathbf{M}_k)$ corresponds to an enantiomorph transformation, which can be either allowed or prohibited [in the latter case the sign of the last column of \mathbf{M}_k is changed to ensure that $\det(\mathbf{M}_k) = 1$].

When the two objects are in canonical positions, the inertia ellipsoids are optimally superposed. This does not mean yet that the principal inertia vectors are superposed as the signs of the eigenvectors cannot be determined from the inertia tensor only. In order to resolve this ambiguity, Galvez & Canton (1993) suggested selecting as positive the directions to the most distant points from the centroid along the first two eigenvectors. This noise-sensitive approach does not guarantee the best initial superposition. We shall select the axes using the proximity measure: if both S_1 and S_2 are in the canonical position, the best orientation of S_2 should minimize $\rho(S_1, S_2)$ [a similar approach was used by Walther *et al.* (2000)]. Depending on whether enantiomorph transformations are allowed or not, there are eight or four sign combinations of the eigenvectors, respectively. After the matrices \mathbf{M}_1 and \mathbf{M}_2 have been determined, the rotation $\mathbf{M}_1 \mathbf{M}_2^T$ and the shift by the vector $\mathbf{T}_2^0 - \mathbf{T}_1^0$ provide the inertia-axes superposition of S_2 onto S_1 .

A more complicated ambiguity is observed if two eigenvalues are equal as the orientation of the object in the plane containing these eigenvectors is not defined. In this case, the

best orientation can be determined by minimizing the NSD against a number of discretely sampled in-plane rotations.

Summarizing, the algorithm for best-matching superposition of a three-dimensional point set S_2 onto S_1 is sketched out as follows.

(i) Inertia tensors and their eigenvectors are computed for both S_1 and S_2 .

(ii) With the transformation \mathbf{M}_1 being fixed, the signs of the columns of the matrix \mathbf{M}_2 are selected out of four sign combinations (or eight, if the enantiomorphs are allowed).

(iii) S_2 is rotated by $\mathbf{M}_1 \mathbf{M}_2^T$ and shifted by $\mathbf{T}_2^0 - \mathbf{T}_1^0$ to align its principal axes of inertia with those of S_1 .

(iv) The position of S_2 is refined by minimizing NSD (2). The minimum value of the latter provides the estimate of dissimilarity between the objects.

The program *SUPCOMB*, implementing the above algorithm, was written in Fortran. The *QL* algorithm with implicit shifts was employed to calculate the eigenvectors of inertia matrix, reduced to tridiagonal form by the Householder method (Press *et al.*, 1992). The variable metric Brojden–Fletcher nonlinear minimization algorithm with simple bounds (Gill *et al.*, 1981) was used to minimize the function (2) starting from the canonical position.

4. Results and discussion

Numerical simulations were performed to study the performance of the algorithm and its stability to noisy and incomplete data; the results are summarized in Table 1. A template object (S_1) and the object to be superposed onto the template (S_2) were high- or low-resolution models of biological macromolecules or their fragments. In the cases in which the best-matching position of S_2 was known in advance, the corresponding NSD ρ_0 and the root mean square deviation RMS_0 (if S_1 and S_2 had a one-to-one correspondence of points) were calculated for this position. *SUPCOMB* was then applied to ‘best match’ an arbitrarily rotated and shifted model S_2 with S_1 . When the best-matching position was unknown, the program was started from an arbitrary position of S_2 . The NSD between the two objects after the inertia-axes alignment (ρ_i), the final values of ρ_f and RMS_f (when applicable), and the CPU time used by the algorithm on a 180 MHz Silicon Graphics workstation are presented in Table 1.

4.1. Aligning atomic models

A fragment of the atomic model of rat kinesin (Kozielski *et al.*, 1997) (20 C_a atoms corresponding to residues 12–31) displayed in Fig. 1 was used as a template (c20) in the first series of simulations (No. 1–12). In order to test the noise robustness of the algorithm, the positions of the C_a atoms in the template were randomized by a uniformly distributed noise with magnitude ranging from 0.1 to 0.5 nm (structures c20n1 to c20n5). To test the stability of the method towards incomplete structures, the template was elongated from both ends with the C_a chain atoms from the same structure, composing the segments of 22 (c22), 24 (c24), 26 (c26) and 28 (c28) atoms (Table 1, rows 6–9). To test the algorithm against

Table 1

Results of computer simulation tests using *SUBCOMB*.

The number of atoms in the object is either included in its name or given in brackets.

No.	Template object S_1	Matched object S_2	ρ_0	ρ_i	ρ_f	RMS ₀	RMS _f	CPU time (s)
1	c20	c20n1	0.28	0.27	0.26	1.05	0.99	0.018
2		c20n2	0.56	0.55	0.54	2.05	1.97	0.019
3		c20n3	0.72	0.68	0.65	3.33	3.33	0.021
4		c20n4	0.75	0.75	0.70	3.95	4.00	0.026
5		c20n5	1.01	0.96	0.93	5.24	9.28	0.024
6		c22	0.21	0.40	0.20	–	–	0.025
7		c24	0.41	0.63	0.39	–	–	0.030
8		c26	0.65	0.76	0.61	–	–	0.031
9		c28	0.89	0.86	0.83	–	–	0.026
10		c24n2	0.70	0.74	0.65	–	–	0.027
11		c24n3	0.84	0.88	0.82	–	–	0.021
12		c24n4	1.09	1.03	1.01	–	–	0.024
13	dk675	dk700	0.36	1.40	0.35	–	–	6.92
14	lyz_pdb (123)	lyz_sld (611)	1.36	1.36	1.33	–	–	4.07
15	lyz_pdb (123)	lyz_dam (3018)	0.76	0.79	0.75	–	–	10.8
16	lyz_sld4 (611)	lyz_sld7 (611)	0.58	0.71	0.57	1.72	1.73	6.46
17	1got (338)	lyz_sld (611)	–	2.44	2.24	–	–	5.79
18	hCP (1046)	hCP10 (1046)	0.65	0.70	0.59	4.81	3.86	14.95
19	hCP (1046)	hCP20 (1046)	1.48	1.81	0.92	11.43	6.87	18.8
20	hCPdam1 (1091)	hCP10 (1046)	0.63	0.86	0.63	–	–	20.7
21	hCPdam (1071)	hCP10 (1046)	0.83	0.87	0.81	–	–	21.5
22	1got_pdb (338)	1got_dam (410)	–	1.33	1.24	–	–	4.4
23	pvd_pdb (1074)	pvd_dam (582)	–	1.14	1.10	–	–	13.2

noisy and incomplete data, the object c24 was further smeared with noises of 0.2 nm (c24n2), 0.3 nm (c24n3) and 0.4 nm (c24n4) width (Table 1, rows 10–12).

In all these tests, the objects were recognized as similar ($\rho_f < 1$) and the initial position of the template was neatly restored (see examples in Fig. 1) except for the three cases with the largest distortions (c20n5, c28 and c24n4). In the examples

c20n1 to c20n4, *SUBCOMB* yielded results that virtually coincided with those provided by the method of Kabsch (1978), explicitly utilizing the one-to-one correspondence of the atoms. For the noise level of 0.5 nm (c20n5), the template was so severely distorted that *SUBCOMB* found the solution with a lower NSD by flipping the chain with respect to the YZ plane in Fig. 1(a) (upper panel).

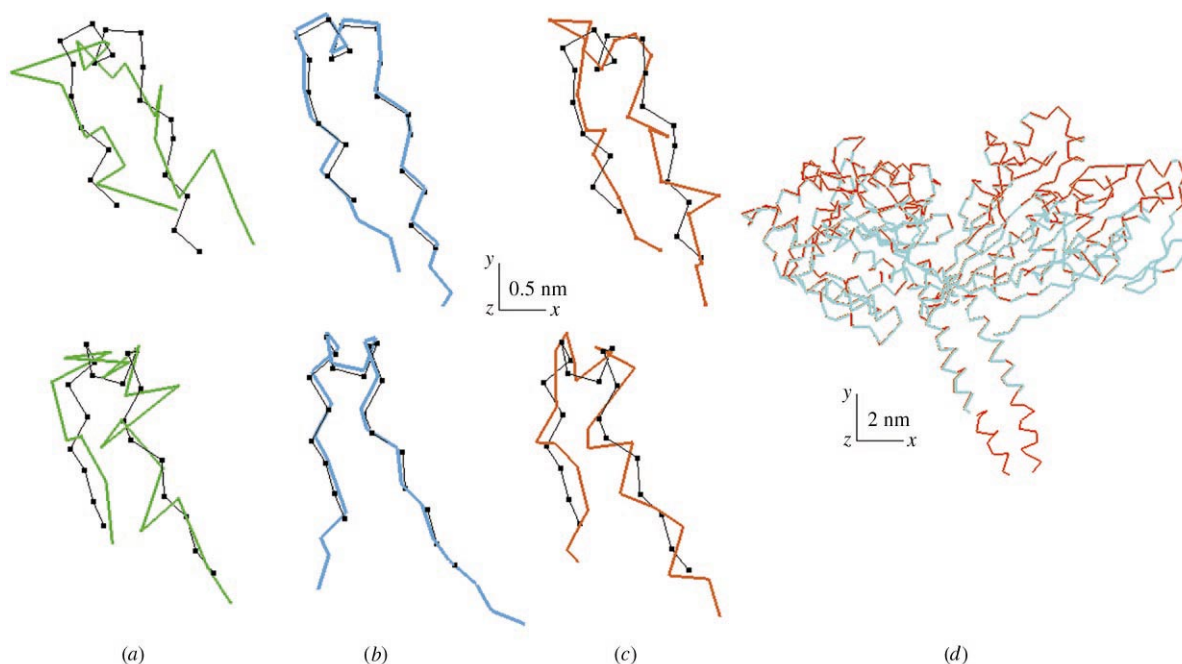
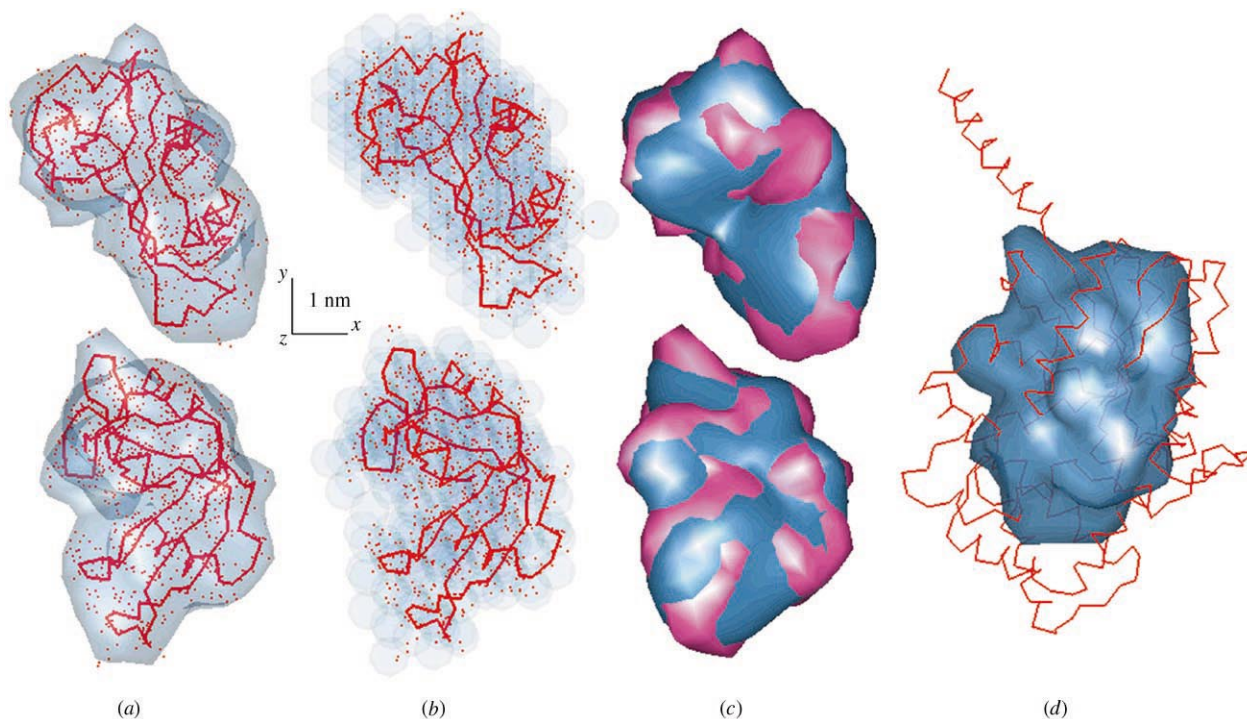
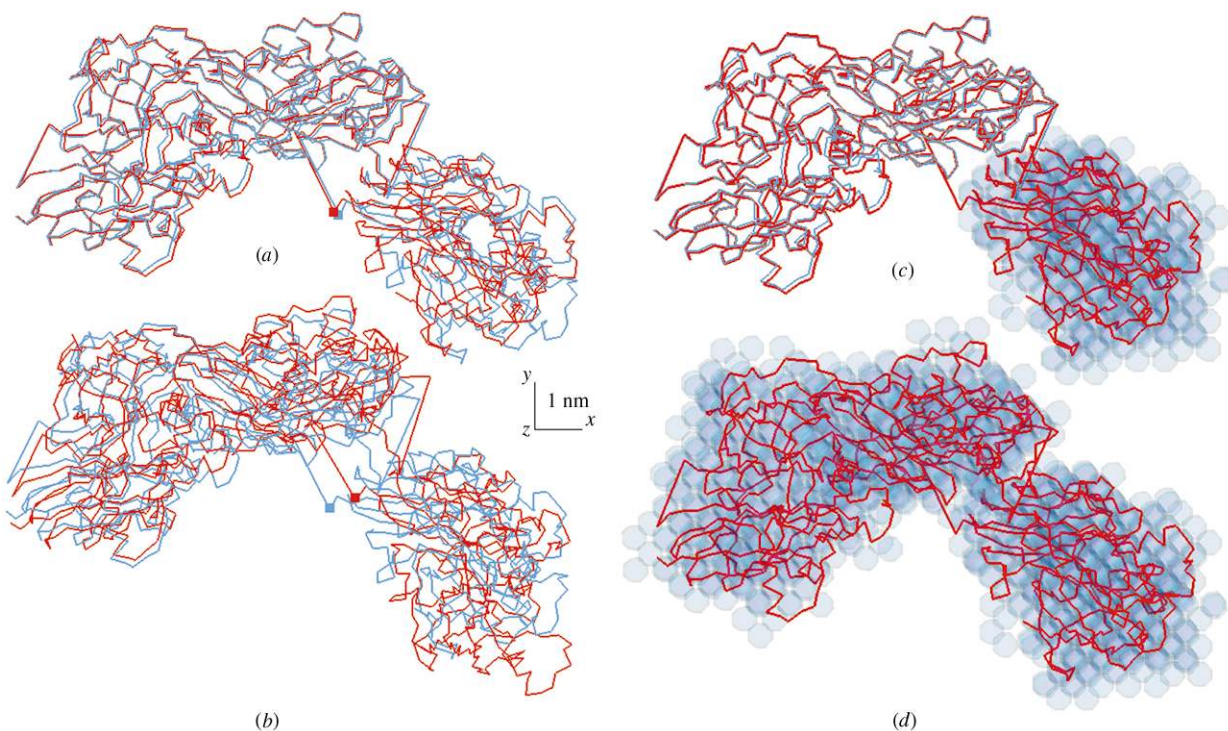


Figure 1

Superposition of atomic models. (a), (b), (c) Aligning a template chain containing 20 atoms (c20, black C_α trace) with the distorted templates c20n3 (a), c26 (b) and c24n2 (c) (see Table 1 and text for more details). The bottom row is rotated clockwise by 90° about the Y axis. (d) Superposition of two models of rat kinesin differing in the length of the coil-coil fragment. All figures were prepared on a Silicon Graphics workstation using the program ASSA (Kozin *et al.*, 1997).


Figure 2

(a), (b) Superposition of the atomic model of lysozyme with its low-resolution envelope and with the space-filling dummy-atom representation, respectively; (c) superposition of two envelope models of lysozyme at different resolution levels. The bottom row is rotated clockwise by 90° about the Y axis. (d) The envelope model of lysozyme superposed with the atomic model of the GTP-binding transducer. Blue: templates. Red: search models.


Figure 3

Superposition of multidomain human ceruloplasmin models. (a), (b) Alignment of the template high-resolution atomic model (hCP) and the same structure with one of the domains rotated about residue No. 338 (marked by a square) by the Euler angles $\{10^\circ, 10^\circ, 10^\circ\}$ (hCP10) and $\{20^\circ, 20^\circ, 20^\circ\}$ (hCP20), respectively. (c), (d) Superposition of hCP10 onto the template with the space-filling dummy-atom model replacing the structure of the rotated domain (hCPdam1) and the whole template (hCPdam), respectively. The bottom row is rotated clockwise by 90° about the Y axis. Black: templates. Red: search models.

In simulation No 13, the two copies of the entire model of rat kinesin (Kozielski *et al.*, 1997) differing by the length of the coil-coil fragment were superimposed. As seen from Fig. 1(d), in the orientation found by *SUPCOMB*, common portions of the molecule are neatly matched. The ρ_f value does not exceed unity (Table 1, row 13) indicating that the two objects are similar.

4.2. Aligning structural models of different resolution

Simulations No. 14–16 dealt with different models of the same protein (lysozyme). The method has found a correct matching of the hollow envelope model lyz_sld computed by the program *CRY SOL* (Svergun *et al.*, 1995) from the PDB code 6lyz (Diamond, 1974) with the atomic model lyz_pdb (Fig. 2a) although the NSD exceeds 1 because the points describing lyz_sld are located only on the surface of the object. The NSD decreased when the envelope model was uniformly filled by densely packed dummy atoms (lyz_dam) (Fig. 2b). Fig. 2(c) presents the superposition of two envelope models of lysozyme, differing by the resolution defined by the maximum order of spherical harmonics used (Svergun, 1994): $L = 4$ for lyz_sld4 and $L = 7$ for lyz_sld7. According to *SUPCOMB*, these models of the same protein are indeed similar ($\rho_f = 0.57$). By contrast, an attempt to superpose rather different objects, namely the globular hollow envelope model of lysozyme (lyz_sld) and a prolate atomic structure of the β subunit of the GTP-binding transducer (Sondek *et al.*, 1996), PDB code 1got, yields a poor $\rho_f = 2.24$ (Fig. 2d and Table 1, row 17).

4.3. Aligning multidomain models

In the following four simulations (18–21), the models of the same multidomain particle in different conformations and at different resolutions were aligned. The structure of apo human ceruloplasmin (hCP) (Zaitsev *et al.*, 1999), containing 1046 residues, was taken as a template in simulation No 18. The apo-hCP contains three domains linked by flexible loops freely moving in solution (Vachette, 1999). One of the side domains was rotated around the connecting residue (No. 338) by the Euler angles $\{10^\circ, 10^\circ, 10^\circ\}$ to obtain the search model hCP10. It took less than 15 s of CPU time for *SUPCOMB* to recover the superposition of the two structures in Fig. 3(a), identical to the superposition provided by the method of Kabsch (1978). The position of the undistorted domain, including the rotation point indicated as a square, is restored correctly. The same experiment was repeated with the object hCP20 obtained from the template by rotation around the same residue with the Euler angles $\{20^\circ, 20^\circ, 20^\circ\}$. In this case the restored position of undistorted domains differs more significantly from the initial one (Fig. 3b), although the final value of $\rho_f = 0.93$ indicates that *SUPCOMB* still regards the objects as similar. The differences in the ρ_f and RMS_f values (Table 1, rows 18 and 19) for these two simulations are clearly visualized in the residue-distance distributions in Figs. 4(a) and 4(b). The averaged displacements between the corresponding residues for the two domains are shown by the dashed lines in Fig. 4. The ratio between the averaged

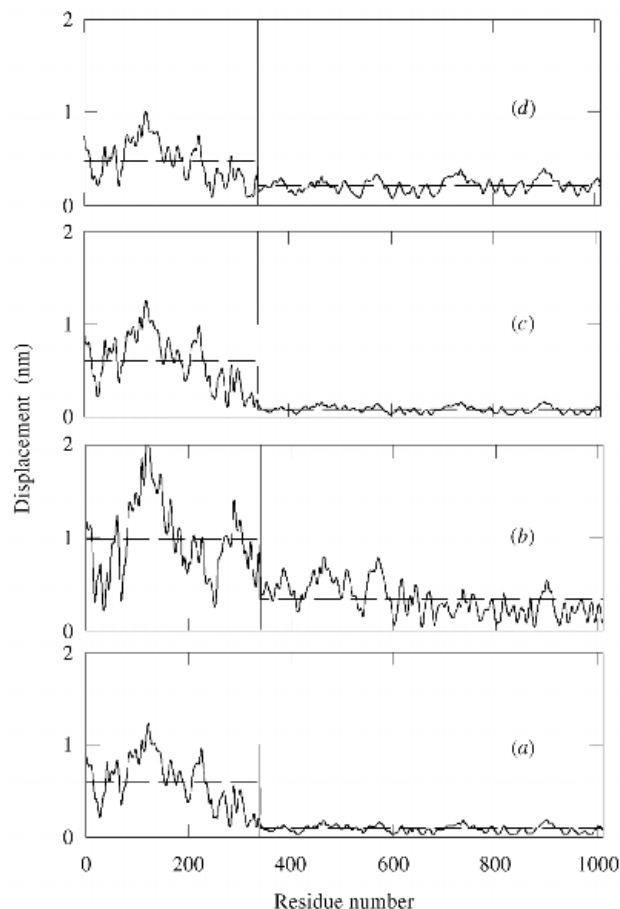


Figure 4

Displacements between the corresponding residues in the structures of human ceruloplasmin hCP (template) and the search models in the position found by *SUPCOMB*. The search models are: (a) hCP10 superposed on hCP; (b) hCP20 superposed on hCP; (c) hCP10 superposed on hCPdam1; (d) hCP10 superposed on hCPdam. The vertical line indicates the rotation point. Dotted horizontal lines indicate the averaged displacements for corresponding domains.

displacements for fixed and rotated domains (AD ratio) is 6.0/0.9 for hCP10 and 9.8/3.3 for hCP20. As seen from Fig. 4, the residue separating the two domains can be approximately located for hCP10 (Fig. 4a) but not for hCP20 (Fig. 4b). It should be stressed that the alignment is obtained without prior knowledge about the residue–residue correspondence so that the domains may contain different numbers of atoms. In this example, we have taken the same number of atoms for illustrative purposes only (Fig. 4).

Moreover, it is possible to use this technique if the template (or part of it) is known only at low resolution. In the experiment No. 20, the structure of the domain to be rotated was replaced in the template by its shape represented by dummy atoms (hCPdam1), and hCP10 was taken as a search model. The position restored by *SUPCOMB* (Fig. 3c) is nearly identical to that obtained in the experiment No. 18. In the experiment No. 21, a space-filling representation of the whole template was used (hCPdam). *SUPCOMB* still detects the similarity of the structures ($\rho_f = 0.81$) and restores the initial

position of the ‘undistorted’ portion of the template reasonably well (Fig. 3*d*). To characterize the obtained alignments, the RMS and residue displacement distribution were evaluated for the superimposed models with respect to the high-resolution template (hCP). The AD ratio for the experiment 20 (6.1/0.8) and the residue–residue distribution (Fig. 4*c*) are virtually the same as those obtained for the high-resolution template. For the experiment No. 21, the AD ratio is diminished to 4.73/2.08, but the border between the domains can still be seen on the displacement plot (Fig. 4*d*). Thus, the method is able to identify structural domains in the models of complex particles against both high- and low-resolution templates.

4.4. Aligning atomic models with models from solution scattering

The two final examples display superposition of crystallographic atomic models of proteins with low-resolution shapes restored *ab initio* from experimental solution-scattering data using the dummy-atoms method of Svergun (1999). The *ab initio* solution-scattering models are not only different in resolution, but they also include a hydration shell around the macromolecule. Such differences may cause ambiguities in superposition, as illustrated for the β subunit of the GTP-binding transducer (Sondek *et al.*, 1996). Using *SUPCOMB*, the superposition of the atomic structure (1got_pdb) with the *ab initio* dummy-atom model of the particle (1got_dam) in Fig. 5(*a*) yielded $\rho_f = 1.24$. A manual search for the initial approximation enables a visually more appropriate super-

position to be obtained, as shown in Fig. 5(*b*), with a slightly better NSD ($\rho_f = 1.22$). However, in all other cases analysed up to now (more than three dozen), it was not possible to improve the quality of the *SUBCOMB* superposition by the manual search. A typical result obtained for a dimeric yeast pyruvate decarboxylase (Arjunan *et al.*, 1996) is illustrated in Fig. 5(*c*) (pvd_pdb is the atomic structure and pvd_dam is the dummy-atom model). It is important that the enantiomorph option is allowed in such superpositions as the handedness of the *ab initio* models from solution scattering can be selected arbitrarily.

5. Conclusions

The proposed algorithm enables the rapid superposition of structural models of different nature and provides a quantitative estimate of similarity between the models. The above numerical simulations demonstrate the reliability of the method, but also reveal its limitations. A significant improvement over the initial inertia-axes alignment is observed especially for similar structures with different numbers of atoms (Table 1, rows 8 and 13) and for models differing in resolution (Table 1, rows 16 and 20). The normalized spatial discrepancy (2) is analogous to the error-weighted discrepancy χ characterizing deviations between one-dimensional data sets. The fineness d_i plays the role of a standard deviation, so that the value $\rho_f > 1$ points to systematic deviations between the objects.

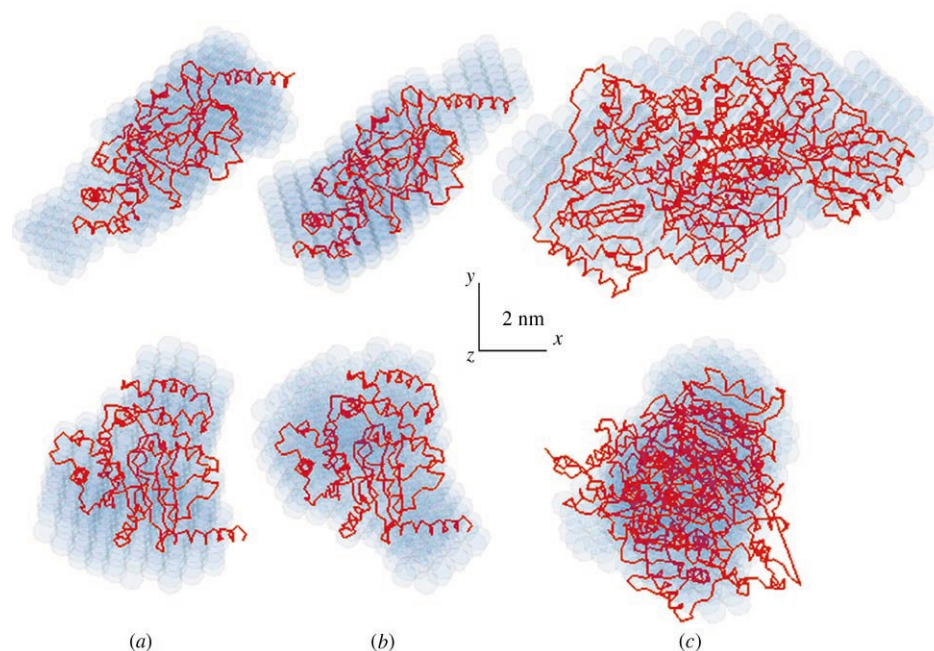


Figure 5

Superposition of *ab initio* dummy-atom low-resolution models (blue semi-transparent spheres) with the atomic models (red): (*a*), (*b*) automated and manual superposition of the β subunit of the GDP-binding transducer; (*c*) pyruvate decarboxylase. The bottom row is rotated clockwise by 90° about the Y axis.

The CPU time used by *SUPCOMB* is proportional to the product of the numbers of points in the two objects. As seen from Table 1, the program can handle about 40 noisy 20-atom motifs per second. The program can thus be used for automated density-map interpretation in protein crystallography, where the need for fast and reasonably accurate identification methods is rather high. Specifically, the method could be applied for detecting certain structure templates, such as α -helices in pseudo-atomic structures obtained from electron density maps by the program *aRP/wARP* (Perrakis *et al.*, 1999).

The program can read standard PDB files, envelope functions specified by spherical harmonics and the major formats used in electron microscopy. The executable module of *SUPCOMB* for Windows 9x/NT and major Unix platforms is available via www.embl-hamburg.de/ExternalInfo/Research/Sax/index.html.

APPENDIX A

Normalized spatial discrepancy and triangle inequality

A metric on a set of objects \mathbf{X} is a function $p: \mathbf{X} \times \mathbf{X} \rightarrow \mathbf{R}$, satisfying the following conditions for all $x, y, z \in \mathbf{X}$ (Copson, 1968):

$$p(x, y) = 0 \Leftrightarrow x = y, \quad (6)$$

$$p(x, y) = p(y, x) \quad (\text{symmetry}), \quad (7)$$

$$p(x, y) + p(y, z) \geq p(x, z) \quad (\text{triangle inequality}). \quad (8)$$

Let us construct an example when the measure (2) fails the triangle inequality. First, we extend the definition of the structure fineness parameter to the case of singleton structure, assuming that $d_i = 1$ if S_i contains only one point. Consider a rhomb $ABCD$ of unit side with the diagonal $AC = a$. Let the set S_1 consist of the vertex A , the set S_2 consist of the vertices A, B, C and D , and the set S_3 consist of the vertex C . Then $d_1 = d_2 = d_3 = 1$, and according to (2),

$$\begin{aligned} \rho(S_1, S_2) + \rho(S_2, S_3) &= 2\rho(S_1, S_2) \\ &= 2[(1/2)(a^2 + 2)/4 + 0]^{1/2} \\ &= [(a^2 + 2)/2]^{1/2} \end{aligned} \quad (9)$$

and

$$\rho(S_1, S_3) = a, \quad (10)$$

so that

$$\rho(S_1, S_3) > \rho(S_1, S_2) + \rho(S_2, S_3), \quad (11)$$

if $a > 2^{1/2}$, *i.e.* if the angle B is obtuse. Now assume that both S_1 and S_3 consist of two or more points separated by the same distance $d \ll 1$. Then the distances (9) and (10) must be divided by $d_1 = d_3 = d$. Since the other changes in these values are negligibly small, the inequality (11) holds true.

The authors thank V. Volkov for useful discussions and P. Vachette and F. Kozielski for providing the test examples. The

work was supported by the EU Grant BIO4-CT97-2143 to DIS and by the EMBL fellowship to MBK.

References

- Alt, H., Fuchs, U., Rote, G. & Weber, G. (1996). Technical Report B 92-03. Freie Universitaet Berlin.
- Arjunan, P., Umland, T., Dyda, F., Swaminathan, S., Furey, W., Sax, M., Farrenkopf, B., Gao, Y., Zhang, D. & Jordan, F. (1996). *J. Mol. Biol.* **256**, 590–600.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
- Bloch, I. (1990). PhD thesis, Telecom Paris 90E018.
- Bloch, I., Maitre, H. & Minoux, M. (1993). *Pattern Recogn. Image Anal.* **3**, 137–149.
- Burel, G. & Hugues, H. (1995). *Graph. Models Image Proc.* **57**, 400–408.
- Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
- Copson, E. T. (1968). *Metric Spaces*. Cambridge University Press.
- Diamond, R. (1974). *J. Mol. Biol.* **112**, 535–542.
- Diederichs, K. (1995). *Proteins*, **23**, 187–195.
- Frank, J., Zhu, J., Penczek, P., Li, Y., Srivastava, S., Verschoor, A., Rademacher, M., Grassucci, R., Lata, R. K. & Agrawal, R. K. (1995). *Nature (London)*, **376**, 441–444.
- Galvez, J. M. & Canton, M. (1993). *Pattern Recogn.* **26**, 667–681.
- Gilbert, D., Westhead, D., Nagano, N. & Thornton, J. (1999). *Bioinformatics*, **15**, 317–329.
- Gill, P. E., Murray, W. & Wright, M. H. (1981). *Practical Optimization*. London: Academic Press.
- Greaves, R. B., Vagin, A. A. & Dodson, E. J. (1999). *Acta Cryst.* **D55**, 1335–1339.
- Hagedoorn, M. & Veltkamp, R. C. (1999). Technical Report UU-CS-1999-03. Utrecht University.
- Heffernan, P. J. & Schirra, S. (1994). *Comput. Geom. Theory Appl.* **4**, 137–156.
- Holm, L. & Sander, C. (1997). *Nucl. Acids Res.* **25**, 231–234.
- Holupka, E. J. & Kooy, H. M. (1992). *Med. Phys.* **19**, 433–438.
- Huttenlocher, D. P., Klanderermann, G. A. & Rucklidge, W. J. (1993). *IEEE Trans. Pattern Recogn. Mach. Intel.* **15**, 850–863.
- Kabsch, W. A. (1978). *Acta Cryst.* **A34**, 827–828.
- Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.
- Kozielski, F., Sack, S., Marx, A., Thormahlen, M., Schoenbrunn, E., Biou, V., Thompson, A., Mandelkow, E.-M. & Mandelkow, E. (1997). *Cell*, **91**, 985–994.
- Kozin, M. B., Volkov, V. V. & Svergun, D. I. (1997). *J. Appl. Cryst.* **30**, 811–815.
- Loncaric, S. (1998). *Pattern Recogn.* **31**, 983–1001.
- Mount, D. M., Netanyahu, N. S. & Le Moigne, J. (1999). *Pattern Recogn.* **32**, 17–38.
- Mundy, J. L. & Zisserman, A. (1992). *Geometric Invariants in Computer Vision*. Cambridge: MIT Press.
- Perrakis, A., Morris, R. & Lamzin, V. (1999). *Nature Struct. Biol.* **6**, 458–463.
- Peters, K. P., Fauck, J. & Froemmel, C. (1996). *J. Mol. Biol.* **256**, 201–213.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes*. Cambridge University Press.
- Rarey, M., Kramer, B. & Lengauer, T. (1997). *J. Comput.-Aided Mol. Des.* **11**, 369–384.
- Reezende, P. J. de & Lee, D. T. (1995). *Algorithmica*, **13**, 387–404.
- Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E. & Sigler, P. B. (1996). *Nature (London)*, **379**, 369–374.
- Svergun, D. I. (1994). *Acta Cryst.* **A50**, 391–402.
- Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
- Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

- Svergun, D. I., Barberato, C., Koch, M. H. J., Fetler, L. & Vachette, P. (1997). *Proteins*, **27**, 110–117.
- Svergun, D. I., Petoukhov, M. V., Koch, M. H. J. & Koenig, S. (2000). *J. Biol. Chem.* **275**, 297–302.
- Ton, J. & Jain, A. K. (1989). *IEEE Trans. Geosci. Remote Sensing*, **27**, 642–651.
- Vachette, P. (1999). *Abstracts of the Workshop 'Shape Determination of Biological Macromolecules in Solution and Related Topics'*, pp. 14–15. Hamburg: EMBL.
- Veltkamp, R. C. & Hagedoorn, M. (1999). Technical Report UU-CS-1999-27. Utrecht University.
- Vriend, G. & Sander, C. (1991). *Proteins*, **11**, 52–58.
- Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.
- Wriggers, W., Milligan, R. A. & McCammon, J. A. (1999). *J. Struct. Biol.* **125**, 185–195.
- Zaitsev, V. N., Zaitseva, I., Papiz, M. & Lindley, P. F. (1999). *J. Biol. Inorg. Chem.* **4**, 579–587.