Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Implementation of cluster analysis for *ab initio* phasing using the molecular envelope from solution X-ray scattering

D. M. Ockwell,[a] M. A. Hough,[a,b] J. G. Grossmann,[b] S. S. Hasnain[a,b] and Q. Hao[a,c]*

[a]Department of Chemistry, De Montfort University, Leicester LE1 9BH, England, [b]CCLRC Daresbury Laboratory, Warrington WA4 4AD, England, and [c]Institute of Physics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China

Correspondence e-mail: qhao@dmu.ac.uk

Received 3 March 2000
Accepted 25 May 2000

Solution of the phase problem is central to crystallographic structure determination. The conventional methods of isomorphous replacement (MIR or SIR) and molecular replacement are ineffective in the absence of a suitable isomorphous heavy-atom derivative or knowledge of the structure of a homologous protein. A recent method utilizing the low-resolution molecular shape determined from solution X-ray scattering data has shown to be successful in locating the molecular shape within the crystallographic unit cell in the case of the trimer nitrite reductase (NiR, 105 kDa) [Hao *et al.* (1999), *Acta Cryst.* D**55**, 243–246]. This was achieved by performing a direct real-space search for orientation and translation using knowledge of the orientation of the polar angles of the non-crystallographic axis obtained by performing a self-rotation on crystallographic data. This effectively reduces the potential six-dimensional search to a four-dimensional one (Eulerian angle γ and three translational parameters). In the case of NiR, the direct four-dimensional search produced a clear solution that was in good agreement with the known structure. The program *FSEARCH* incorporating this method has been generalized to handle molecules from all space groups and in particular those in possession of non-crystallographic symmetry. However, the method employed was initially unsuccessful when applied to the small dimeric molecule superoxide dismutase (SOD, 32 kDa) owing to the absence of strong reflections at low resolution caused by saturation at the detector. The determined solution deviated greatly from that of the known structure [Hough & Hasnain (1999), *J. Mol. Biol.* **287**, 579–592]. It was found that once these absent reflections were replaced by a series of randomly generated intensity values and cluster analysis was performed on the output, the signal-to-noise ratio was improved and a most probable solution was found. The electron-density map of the stochastically determined solution agrees well with the known structure; the phase error calculated from this map was 67° within 14 Å resolution.

1. Introduction

Solution X-ray scattering data obtained using synchrotron radiation X-rays have proven to be very useful in providing low-resolution structural details of proteins and other macromolecules in solution. Owing to the recent and significant progress that has been made with the development of *ab initio* phasing methods for low-resolution shape restoration in terms of spherical harmonics (Svergun & Stuhrmann, 1991; Svergun *et al.*, 1996), the spatial parameters of a structure's molecular envelope can be determined in a model-independent manner which does not, for example, require the use of crystal structure coordinates for interpretation. This method has been used to analyse scattering data from a nitrogenase protein complex to provide a stable and unique shape restoration at ~15 Å resolution (Grossmann *et al.*, 1997). Recently, the low-resolution structure of nitrite reductase (105 kDa) from *Alcaligenes xylosoxidans* (AxNir) has been determined from scattering data (Grossmann & Hasnain, 1997). Although the crystal structure has previously been solved by the molecular-replacement method at 2.8 Å (Dodd *et al.*, 1997), the molecular shape of nitrite reductase (NiR) determined from solution scattering was successfully used as a candidate for *ab initio* phasing by locating the molecule within the crystallographic unit cell (Hao *et al.*, 1999). In order to test the generality of this method, the fairly small dimeric molecule superoxide dismutase (SOD) from bovine erythrocytes (32 kDa) was treated similarly.

2. X-ray scattering experiments and molecular-shape determination

Bovine SOD was prepared in 50 m*M* glycylglycine buffer at pH 6.5 according to standard procedures (Hough & Hasnain, 1999). X-ray solution-scattering experiments were performed with protein concentrations between 1 and 5 mg ml^{-1} on station 2.1 (Towns-Andrews *et al.*, 1989) at the SRS Daresbury Laboratory, England using a position-sensitive multiwire proportional counter (Lewis, 1994). The sample-to-detector distance of 1.8 m and the X-ray wavelength of λ = 1.54 Å allowed the coverage of a momentum-transfer interval of $0.005 \leq s \leq 0.06$ Å^{-1} (where the modulus of the momentum transfer is defined as $s = 2\sin\theta/\lambda$, where 2θ is the scattering angle). The *s* range was calibrated using an oriented specimen of wet rat tail collagen (based on a diffraction spacing of 670 Å). Reduction and analysis of scattering data was carried out as described previously (Grossmann *et al.*, 1998). The radius of gyration and the intraparticle distance distribution function $p(r)$ were calculated from the experimental scattering data using the indirect Fourier transform method as implemented in the program *GNOM* (Semenyuk & Svergun, 1991). The solution-scattering curves and intra-particle distance distributions for oxidized bovine SOD are shown in Fig. 1(*a*) and allowed the determination of the following structural parameters: R_g (radius of gyration) = 21.0 Å ± 1%, $D_{\max}$ (maximum particle dimension) = 64 Å ± 4% and V (particle volume) = 50 000 Å^3 ± 5%.

The multipole-expansion method proposed by Stuhrmann (1970) and developed by Svergun and coworkers (Svergun & Stuhrmann, 1991; Svergun *et al.*, 1996, 1997) was used to restore the molecular shape of SOD. The smoothed scattering profile was fitted *ab initio* by the scattering from an envelope function starting from an ellipsoidal initial approximation. The molecular shape was characterized with spherical harmonics up to sixth order assuming a twofold symmetry axis (see Table 1), which is acceptable considering the information content of the data. The restored envelope is displayed in Fig. 1(*b*) and its theoretical scattering curve is superimposed on the experimental data in Fig. 1(*a*). The shape of the protein

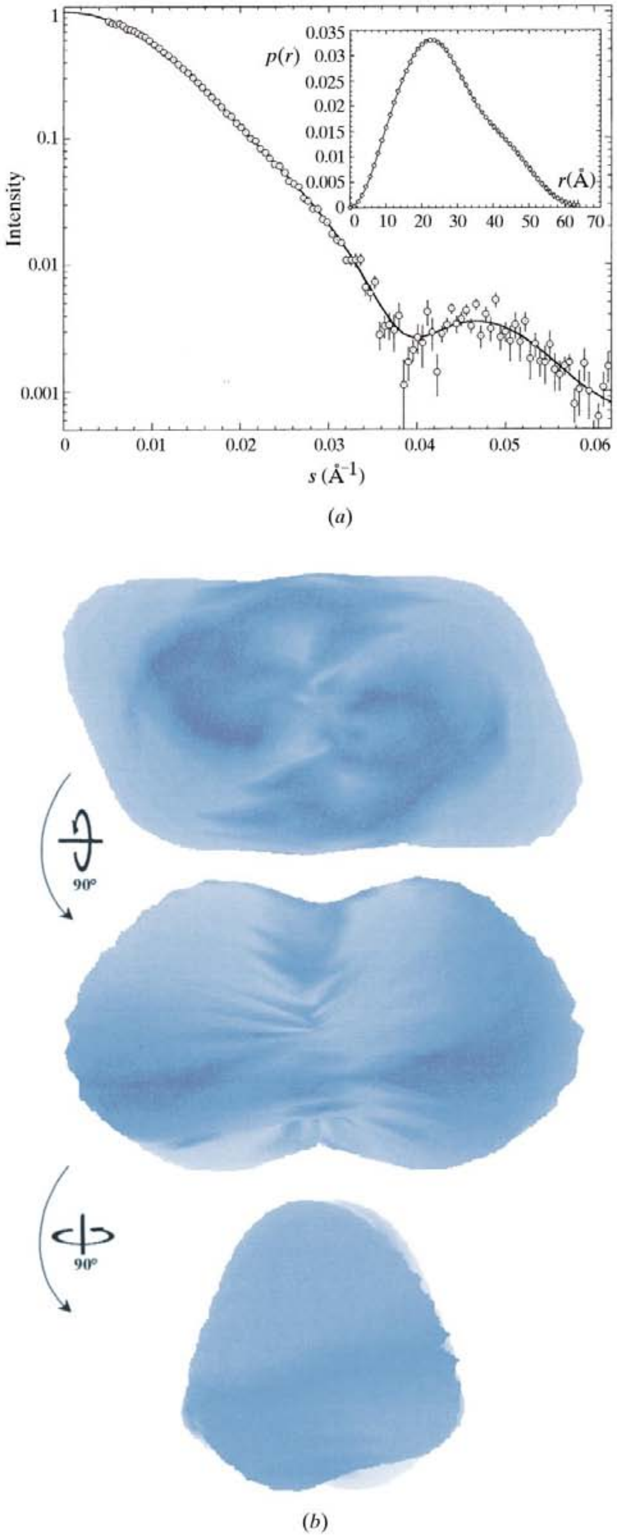

Figure 1
Experimental SAXS results for the solution of oxidized bovine SOD. (*a*) Representation of the scattering curve with error bars based on counting statistics. The calculated distance-distribution function is shown as an inset. The fit (solid line) to the experimental curve corresponds to the scattering from the restored shape. (*b*) Molecular shape of SOD deduced from X-ray solution scattering data for up to the sixth order of harmonics (displayed in three different orientations).

Table 1
Multipole coefficients (real and imaginary components of f_{lm} values) evaluated for the shape restoration of SOD up to the sixth order of harmonics (including a scale factor R_0 = 22.85 Å).

For further explanations see Hao *et al.* (1999).

l	m	Re	Im
0	0	3.214	–
1	0	0.005	–
2	0	−0.354	–
2	2	−0.338	−0.541
3	0	−0.265	–
3	2	0.097	0.100
4	0	−0.039	–
4	2	−0.052	0.013
4	4	−0.095	−0.109
5	0	0.152	–
5	2	−0.071	−0.036
5	5	0.051	0.036
6	0	0.051	–
6	2	−0.048	−0.034
6	4	0.039	0.054
6	6	−0.079	−0.016

neatly reproduces the molecular features when compared with the overall details from the crystal structure.

3. Seeking the molecular shape in the crystallographic unit cell

SOD crystallizes in space group $P2_12_12_1$, with unit-cell parameters a = 47.8, b = 51.4, c = 148.0 Å. There is a single SOD homodimer in the asymmetric unit. X-ray diffraction data extending to 2.5 Å were collected at station BL6A-2 of the Photon Factory, KEK, Japan. The data has an overall completeness of 88%. The data completeness was 85% to 10 Å, 78% to 12 Å and 69% to 14 Å. The crystal structure was originally solved (Hough & Hasnain, 1999) by the molecular-replacement method at 1.65 Å using *AMoRe* (Navaza, 1994) with the 2.0 Å structure of cobalt-substituted SOD (PDB code 1cob) as the search model.

The conventional method for correctly positioning a known search molecule in a crystallographic unit cell – an important first step in solving macromolecular structures by the molecular-replacement method – is by the use of the cross-rotation function (Rossmann & Blow, 1962). An attempt was made to locate the molecular shape determined by solution scattering by performing a Patterson search at different resolutions using *AMoRe* (Navaza, 1994). This was unsuccessful as the intra-envelope vectors are uniformly distributed and do not match the intramolecular (atom-to-atom) vectors represented by the Patterson function. In the case of a macromolecule lacking non-crystallographic symmetry, a full six-dimensional search (three orientational and three translational) in real space appears necessary. However, for a molecule in possession of non-crystallographic symmetry the search may be performed in two separate stages.

Utilizing a self-rotation function using *ALMN* from the *CCP*4 suit (Collaborative Computational Project, Number 4, 1994) with 2.5 Å crystallographic data yielded Eulerian angles for the non-crystallographic twofold axis (at α = 97.3°, β = 96.1°) of the molecular shape. Once the orientation of this twofold axis is known, the potential six-dimension search is reduced to four (Eulerian angle γ and three translational parameters), resulting in a significant reduction in calculation time when locating the molecular shape within the crystallographic unit cell.

The program *FSEARCH*, which has been generalized to handle molecules from all space groups and in particular those in possession of non-crystallographic symmetry, was used to conduct a simultaneous rotational and translational search to find the best match between $F_{\rm obs}$ and F_c. However, a four-dimensional search within the unit cell using crystallographic data in the resolution range ∞–10 Å initially failed to find solutions that matched those of the known structure. This failure was attributed to a significant number of strong low-resolution reflections being absent from the experimental X-ray diffraction data owing to saturation at the detector (the re-run with error-free values for the missing terms computed from the final PDB coordinates produced the correct solution in a straightforward manner). The effect of this was to lower the signal-to-noise ratio and thus produce results that fail to correctly position the molecular shape within the unit cell. If the signal-to-noise ratio could be improved sufficiently then it should be possible to find a solution which corresponds more closely to the correct one. This has been achieved in the case of experimental X-ray diffraction data from the SOD structure through cluster analysis.

4. Cluster analysis

Cluster analysis has been successfully employed in order to determine the molecular envelope of proteins using low-resolution complete crystallographic data (Lunin *et al.*, 1990). The rationale behind cluster analysis is that provided a large enough number of random trials are performed, a substantial set of solutions close to the ideal will appear with greater frequency than solutions which are randomly distributed and significantly different from one another.

When X-ray crystallographic data has a sufficiently high level of completeness, the computer program *FSEARCH* has been shown to be a very effective tool in locating low-resolution molecular shapes within the unit cell (Hao *et al.*, 1999). However, if a significant number of strong low-resolution reflections are absent owing to saturation at the detector, then *FSEARCH* is found to yield erroneous shape-location coordinates when compared with the known structure. Stochastic methods such as cluster analysis provide a means whereby correct solutions can be ascertained even when the level of low-resolution data completeness is poor.

5. Method

A set of random F values were generated using the Monte Carlo method with values in the range 0–1000. The range was chosen so that its mean value was approximately equal to the average of the reflection F values present in the experimental

Table 2
Cluster analysis of the real-space search results using the molecular shape from solution scattering and crystallographic data at 10, 12 and 14 Å resolution in descending order of frequency *f*.

The chosen solution and its shoulder peaks are indicated by bold values. As a comparison, the solution from the known structure (Hough & Hasnain, 1999) was calculated to be γ = 60°, X = 17, Y = 2, Z = 58 Å.

10 Å					12 Å					14 Å				
γ	X	Y	Z	f	γ	X	Y	Z	f	γ	X	Y	Z	f
69	17.6	10.9	46.8	20	3	10.4	13.9	21.5	19	**60**	**14.3**	**5.6**	**58.8**	**17**
174	17.3	11.0	46.8	12	6	2.8	9.2	31.4	13	0	18.8	19.9	17.7	11
39	14.3	12.3	26.0	9	102	18.9	2.4	47.7	10	3	18.5	21.8	13.6	11
60	13.1	10.5	26.3	8	105	18.7	0.0	18.7	9	180	18.8	19.9	17.7	11
78	7.9	7.9	23.6	8	0	2.6	2.6	18.4	8	**57**	**14.1**	**9.0**	**56.7**	**10**
171	18.4	11.3	44.6	8	0	18.4	21.0	13.1	8	63	7.2	3.0	43.5	10
93	15.5	0.0	10.3	7	3	18.4	18.0	46.9	8	**66**	**11.3**	**7.5**	**55.1**	**8**
R-factor range														
0.615–0.672					0.613–0.684					0.530–0.621				

X-ray diffraction data set. In the resolution range ∞–10 Å, the mean $F_{\rm obs}$ value for the experimental data was ~500 units. If the random set is scaled much higher than this range, for instance equal to the expected average of the saturated reflections, then the cluster analysis produces biased results owing to the dominance of the random data. These randomly generated intensity values were then substituted for the missing reflection intensity values directly into the X-ray diffraction data set. The modified data set was then fed as input into the program *FSEARCH*. In each of the resolution ranges ∞–10, ∞–12 and ∞–14 Å, the process of random intensity value generation was performed ten times so that at each resolution range ten sets of solutions were obtained. Note that the solutions obtained at each resolution were independent of each other, so that cluster analysis was performed at each resolution separately. The solutions obtained were sorted using the *R* factor as a figure of merit; the solutions with the top 20 lowest *R* factors were then collected and sorted by Eulerian angle γ, the rest being discarded. The *R*-factor range for the top 20 lowest *R* factors collected for 10, 12 and 14 Å resolution data were 0.615–0.672, 0.613–0.684 and 0.530–0.621, respectively.

Cluster analysis was then performed whereby all solutions separated by an absolute radial distance of 8 Å (8 Å was found to give the best distribution), where the radial distance is simply the distance between the calculated centre of gravity of the mask, were summed and averaged to yield the best mean translation solution for a group, a group in this case being defined as the set of translational parameters associated with a common Eulerian angle γ. The Eulerian angle γ was kept fixed throughout, as grouping solutions around a change in this angle resulted in a dramatic increase in the disparity of the solutions. The most frequently occurring solution is deemed to be the most probable solution and is ranked accordingly, as shown in Table 2.

As can be observed, the results at 14 Å resolution can be argued to be superior to those at 12 and 10 Å. The average *R* factors for the 14 Å resolution results are significantly lower than one would expect for a random solution (*R* = 0.67), unlike those at 10 and 12 Å. Furthermore, statistical significance may be attached to the presence of large shoulder peaks (peaks 5 and 7) straddling the average solution at γ = 60°. It should be noted that cluster analysis was performed with a larger number of random trials (20) for each of the three resolution ranges and similar results to those above were obtained. However, using larger number of trials was time-consuming.

Phases were then calculated from the top solution at 14 Å resolution, γ = 60°, X = 14.3, Y = 5.6, Z = 58.8 Å, and compared with the phases of the known structure (Table 3). The 69 reflections within 14 Å have an average phase error of 67° and can be used as good starting phases for further phase extension to higher resolutions. The quality of fit between the molecular shape found here and the molecular structure as found by Hough & Hasnain (1999) is illustrated in Fig. 2.

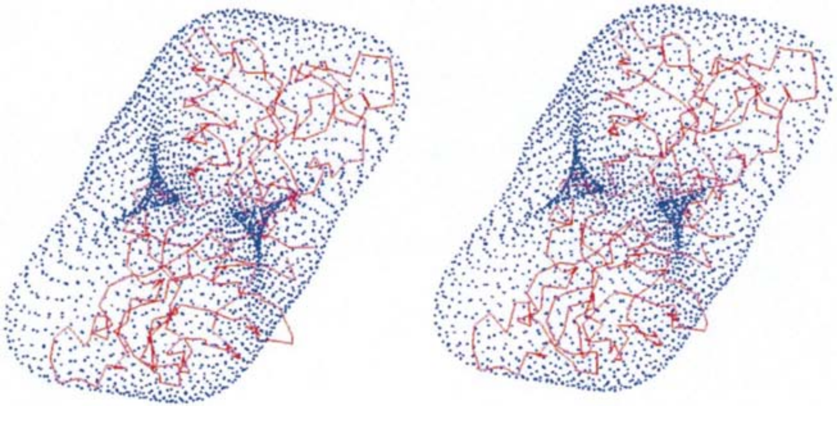

Figure 2
Stereo pair showing the molecular shape found from solution scattering and located in the unit cell by a real-space search, superimposed on the 1.65 Å crystal structure model (Hough & Hasnain, 1999). The molecular shape is represented by dots uniformly distributed within its outline and the crystal structure model by red chains.

Table 3
Average phase errors of the cluster analysis determined molecular-shape replacement result against the refined model at 1.65 Å (Hough & Hasnain, 1999) in descending order of resolution.

Resolution (Å)	Data completeness (%)	Number of experimental reflections	Mean phase error (°)
30	0	0	–
25	26	5	25
20	49	19	40
14	69	69	67
12	78	115	73
10	85	209	83

6. Concluding remarks

We have shown that cluster analysis is a useful stochastic method that enables the correct positioning of a macromolecule within the unit cell from knowledge of the molecular shape determined from solution scattering, even when strong reflections are absent from crystallographic data owing to saturation at the detector. Once the orientation of a non-crystallographic symmetry axis has been determined by a self-rotation function, the potential six-dimensional search is reduced to four dimensions (Eulerian angle γ and three translational parameters), resulting in a great reduction in computational time. However, for cases where non-crystallographic symmetry does not exist, a time-consuming six-dimensional search would be necessary.

It is anticipated that the low-resolution phases calculated from the correctly positioned molecular shape can be used as a good starting point for phase extension through the use of genetic algorithms whereby the mask would be used as the arena for ascertaining a macromolecule's internal mass distribution. Once the resolution of the mask has been improved to ~5 Å using this method, phase extension to higher resolutions may be achieved by maximum-entropy and density-modification methods (*e.g.* solvent flattening, histogram matching, NCS averaging). Thus, it is hoped that this method will greatly facilitate the *ab initio* structure determination of proteins and provide a good foundation for further structure refinement.

We would like to thank EPSRC for a studentship to support DMO.

References

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Dodd, F. E., Hasnain, S. S., Abraham, Z. H. L., Eady, R. R. & Smith, B. E. (1997). *Acta Cryst.* D**53**, 406–418.

Grossmann, J. G., Crawley, J. B., Strange, R. W., Patel, K. J., Murphy, L. M., Neu, M., Evans, R. W. & Hasnain, S. S. (1998). *J. Mol. Biol.* **279**, 461–472.

Grossmann, J. G. & Hasnain, S. S. (1997). *J. Appl. Cryst.* **30**, 770–775.

Grossmann, J. G., Hasnain, S. S., Yousafzai, F. K., Smith, B. E. & Eady, R. R. (1997). *J. Mol. Biol.* **266**, 642–648.

Hao, Q., Dodd, F. E., Grossmann, J. G. & Hasnain, S. S. (1999). *Acta Cryst.* D**55**, 243–246.

Hough, M. & Hasnain, S. S. (1999). *J. Mol. Biol.* **287**, 579–592.

Lewis, R. (1994). *J. Synchrotron. Rad.* **1**, 43–53.

Lunin, V. Yu., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Semenyuk, A. V. & Svergun, D. I. (1991). *J. Appl. Cryst.* **24**, 537–540.

Stuhrmann, H. B. (1970). *Acta Cryst.* A**26**, 297–306.

Svergun, D. I. & Stuhrmann, H. B. (1991). *Acta Cryst.* A**47**, 736–744.

Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). *Acta Cryst.* A**52**, 419–426.

Svergun, D. I., Volkov, V. V., Kozin, M. B., Stuhrmann, H. B., Barberatop, C. & Koch, M. H. J. (1997). *J. Appl. Cryst.* **30**, 798–802.

Towns-Andrews, E., Berry, A., Bordas, J., Mant, P. K., Murray, K., Roberts, K., Sumner, I., Worgan, J. S. & Lewis, R. (1989). *Rev. Sci. Instrum.* **60**, 2346–2349.