Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Macromolecular envelope determination and envelope-based phasing

Quan Hao

MacCHESS, Cornell High Energy Synchrotron Source, Cornell University, Ithaca, NY 14853-8001, USA

Correspondence e-mail: qh22@cornell.edu

Received 31 March 2006
Accepted 19 April 2006

Small-angle X-ray scattering (SAXS) or electron-microscopy (EM) data have proven to be very useful in providing low-resolution structural details of proteins and other macromolecules. To utilize the envelope information for crystallographic phasing, it is essential to develop a method for correctly positioning the known envelope in a crystallographic unit cell. The low-resolution phases calculated from the correctly positioned molecular envelope can be used as a good starting point for phase extension. This paper describes the development of the *FSEARCH* program for locating envelopes in the unit cell and possible ways to extend phases to crystallographic data resolution.

1. Introduction

Once crystals have been obtained, solving the phase problem of crystallography is the biggest hurdle (Shen *et al.*, 2006). Multi-wavelength anomalous dispersion (MAD) phasing helps tremendously, but usually requires the incorporation of selenium into the protein by using selenomethionine during the expression of the protein. However, this is only successful when the gene encoding the particular protein is known and expression has been established and when the methionine substitution does not affect the activity of the protein and also does not disrupt the crystalline order.

There are many proteins of interest that cannot be produced in bacteria, a process that is usually required to incorporate selenomethionine. This can arise because bacteria do not contain the accessory proteins (chaperones) that are often a fundamental prerequisite for correct protein folding or because they do not contain the machinery for making the post-translational modifications that are often essential for function. This is particularly true for proteins from higher organisms, including human proteins of crucial medical importance. In such cases, the proteins can often be successfully expressed in yeast, insect or eukaryotic cell lines, but incorporation of selenomethionine into these alternative expression systems can present serious technical difficulties. For this important class of problems, development of phasing methods that do not require the production of proteins with heavy atoms would provide an attractive alterative to structure solution. The same argument holds for those proteins that are isolated directly from natural sources rather than being produced by recombinant technology.

New phasing methods have been explored that do not require the presence of heavy atoms at low resolution (see reviews by Lunin *et al.*, 2002; Urzhumtsev & Podjarny, 1995). Rabinovich & Shakked (1984), Lunin *et al.* (1995) and Hauptman *et al.* (2002) proposed and successfully tested the

few-atoms-model methods based on *R*-factor calculations or a clusterization procedure to recognize the correct solution from randomly generated models. A method that utilizes the small-angle X-ray scattering (SAXS) technique for macromolecular phasing has also given promising results (Hao *et al.*, 1999; Ockwell *et al.*, 2000; Hao, 2001).

SAXS data have proven to be very useful in providing low-resolution structural details of proteins and other macromolecules in solution. The spatial parameters of a structure's molecular envelope can be determined in a model-independent manner (Svergun & Stuhrmann, 1991; Svergun *et al.*, 1996). Cryo-electron microscopy (EM) of single particles can reach subnanometer resolution (6–10 Å), at which long α-helices and large β-sheets of the protein components of a large macromolecular assembly can be discerned (Jiang & Ludtke, 2005). The SAXS or EM technique combined with crystallography may become a powerful biophysical tool as the trend in structural biology moves towards studying larger and more complex assemblies in the structural proteomics era.

To utilize the envelope information for crystallographic phasing, it is essential to develop a method for correctly positioning the known envelope in a crystallographic unit cell. We have implemented a computer program called *FSEARCH* (Hao *et al.*, 1999; Ockwell *et al.*, 2000; Hao, 2001) to do this by performing a simultaneous six-dimensional search on orientation and translation to find the best match between experimental structure factors, F_{obs}, and calculated structure factors, F_{calc}. The program can be used in general six-dimensional cases for a molecular-replacement solution given a predetermined envelope from any source, such as electron-microscopic images (EM), solution scattering (SAXS) or coordinates of a homologous structure, provided that the

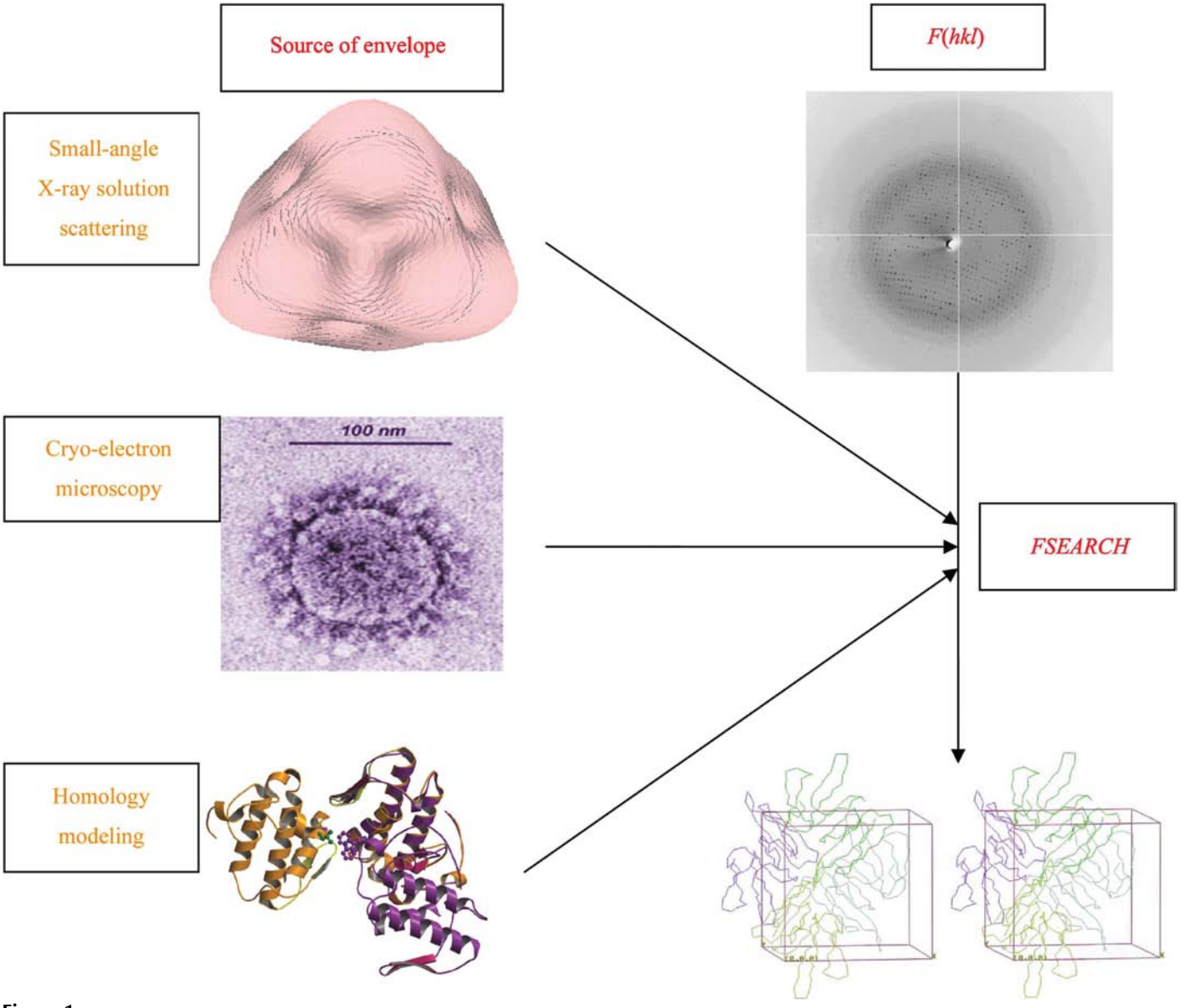

Figure 1
Schematic diagram of the envelope-phasing method.

envelope can be converted to the standard PDB format or expressed in terms of spherical harmonics. The *FSEARCH* program is now supported by *CCP*4 (Collaborative Computational Project, Number 4, 1994). A parallel-aware version, *MPI_FSEARCH*, has been used successfully to perform an exhaustive six-dimensional search to phase very low resolution X-ray data using a molecular envelope (Liu *et al.*, 2003). The *FSEARCH* program has also been used to find a molecular-replacement solution with data from the 420 kDa lobster clottable protein crystals; the search model was a 17 Å resolution structure determined by single-particle EM (Kollman & Quispe, 2005).

The low-resolution phases calculated from the correctly positioned molecular envelope can be used as a good starting point for phase extension with the genetic algorithm and/or the iterative-projections method (Elser, 2003*a*), in which the envelope would be used as the arena for ascertaining a macromolecule's internal structure. Once the resolution of the structure has been improved to ∼5 Å using these methods, phase extension to higher resolutions may be achieved by density-modification methods (*e.g.* solvent flattening, histogram matching or non-crystallographic symmetry averaging).

The phasing procedure described in this paper requires three steps. One is to obtain an SAXS pattern and to determine the molecular envelope. The second is to collect a standard crystallographic data set and use the envelope information for low-resolution phases. The third is to extend phases to higher resolution. The second and third steps can also be applied to envelopes determined from other source, such as electron microscopy (EM) or homology modeling. A schematic diagram of this procedure is shown in Fig. 1.

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

Once the solution-scattering data from a protein sample have been obtained, the next step is to recover the three-dimensional envelope from the one-dimensional scattering pattern. Two methods developed by Svergun and colleagues are commonly used.

2.1. Spherical harmonics method

In the first general *ab initio* approach (Stuhrmann, 1970; Svergun & Stuhrmann, 1991), an angular envelope function of the particle is described by a series of spherical harmonics. The low-resolution shape is thus defined by a few parameters (the coefficients of this series) that fit the scattering data. This approach was implemented in the computer program *SASHA* (Svergun *et al.*, 1997). It was demonstrated that under certain circumstances a unique envelope could be extracted from the scattering data (up to an enantiomorphic shape; this ambiguity holds for all *ab initio* methods in SAXS; Svergun *et al.*, 1996). Both 'left' and 'right' hands should be tested and the ambiguity will be resolved in a later stage of the structure determination when the hand of the helix turns could be determined. This method has been used to analyze scattering data from the nitrite reductase protein from *Alcaligenes xylosoxidans* (NiR; 105 kDa; Grossmann & Hasnain, 1997) and from the dimeric molecule superoxide dismutase (SOD; 32 kDa; Ockwell *et al.*, 2000).

2.2. Monte Carlo method

The use of such envelopes as defined by spherical harmonics is limited to globular particles with relatively simple shapes and without significant internal cavities. More detailed models can be constructed *ab initio* using different types of Monte Carlo searches and the utilization of a simulated-annealing approach in which the shape of the complex is modeled by a large number of close-packed beads which are moved around so as to match the observed scattering profile as best as possible (Svergun *et al.*, 2001). The Monte-Carlo-based models contain hundreds or thousands of parameters and caution is required to avoid over-interpretation. A common approach is to align a set of models resulting from independent shape-reconstruction runs to obtain an average model that retains the most persistent and conceivably also the most reliable features, *e.g.* using the program *SUPCOMB* (Kozin & Svergun, 2001). Particle symmetry, if known, provides very useful constraints, which can be imposed in the programs *SASHA* and *DAMMIN* and in the program *GASBOR* (Svergun *et al.*, 2001).

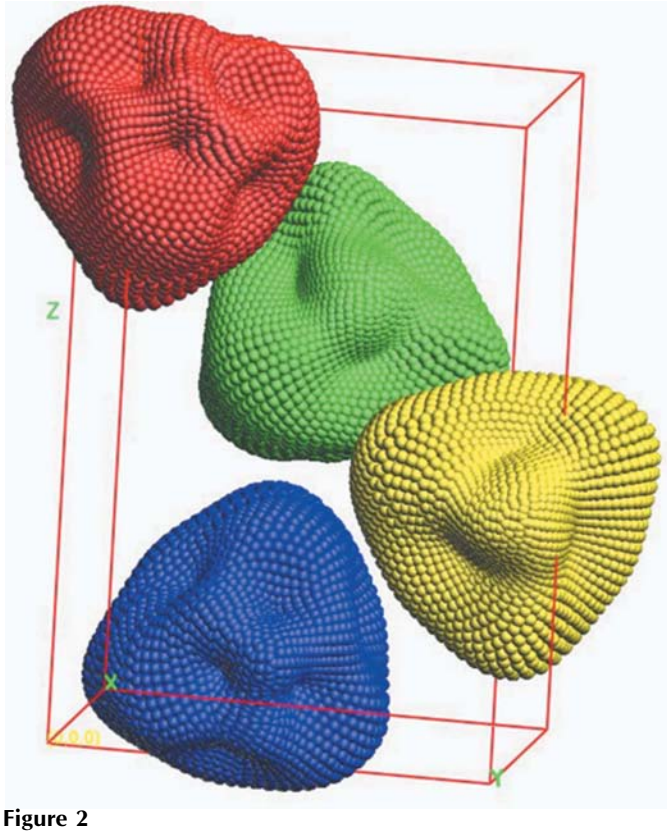

Figure 2
Packing of the SAXS envelopes in the NiR crystal unit cell. The molecular-replacement solution was found with *FSEARCH*. This figure was prepared with *O* (Jones *et al.*, 1991).

3. Locating the molecular envelope in the crystallographic unit cell

Once the molecular envelope has been determined from the SAXS pattern and crystallographic data have been collected, it is natural to use a molecular-replacement method to locate the envelope in the crystallographic unit cell, thus providing low-resolution phases. All methods based on the Patterson function require discrimination between intermolecular vectors and intramolecular vectors. They often work well for molecular replacement with high-resolution diffraction data, but may be less successful when only low-resolution diffraction data or low-resolution search models are available. In cases where the search model comes from electron microscopy or small-angle solution X-ray scattering, all the Patterson-based methods have difficulties; as the density inside the envelope is uniform, there are no discrete intra-envelope Patterson vectors from the model to be matched with observed peaks. We have therefore developed the *FSEARCH* program (Hao *et al.*, 1999; Ockwell *et al.*, 2000; Hao, 2001) to locate an envelope in a crystallographic unit cell by performing a simultaneous six-dimensional search on orientation and translation to find the best match between experimental structure factors and calculated ones. The *R* factor

$$R = \sum_{hkl} |F_{\rm obs}(hkl) - F_{\rm calc}(hkl)| / \sum_{hkl} F_{\rm obs}(hkl),$$

is used to identify the solution, where $F_{\rm obs}(hkl)$ is the observed structure factor of reflection *hkl* and $F_{\rm calc}(hkl)$ is the magnitude of the calculated structure factor.

The *FSEARCH* program can accept either of the two forms of envelope (spherical harmonics or Monte-Carlo-based models) as an input search model. *FSEARCH* was first tested in the location of an SAXS molecular shape within the crystallographic unit cell for the trimeric nitrite reductase (NiR; 105 kDa). Detailed results have previously been published (Hao *et al.*, 1999). Briefly, a self-rotation function using the 2.8 Å crystallographic data yielded the polar angles of the non-crystallographic symmetry (NCS) threefold axis. Knowledge of the orientation of this axis reduced the potential six-dimensional search to four dimensions (the Eulerian angle γ and three translational parameters). The direct four-dimensional search within the crystallographic unit cell produced a clear solution. The packing of the SAXS envelopes in the unit cell based on this solution is shown in Fig. 2. The electron-density map agrees well with the known structure and the phase error calculated from the map was 61° within 20 Å resolution. A bulk-solvent correction (Fokine & Urzhumtsev, 2002) was subsequently attempted, but did not improve the solution.

In the second test case (for more details, see Ockwell *et al.*, 2000), however, the same 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. 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. This example has demonstrated that the absence of strong reflections at low resolution can degrade *FSEARCH* solutions greatly. Therefore, particular attention should be paid in crystallographic data-collection experiments to ensure that the low-resolution data (100–10 Å) are near-complete (by using a small beam stop) and not saturated (by reducing exposure time).

If no NCS axis exists in the molecule, a six-dimensional search would be necessary and a time frame of the order of 100 CPU hours on a single-processor computer would be expected. At MacCHESS, there has been a large effort put into the development of fast computing, including the construction of a parallel processing cluster ('Feynman'). Feynman is a 128-processor Linux-based cluster comprising 63 dual-processor 'client' nodes and one dual-processor master or 'server' node. Feynman has already been used for rapid phasing of macromolecular structures. A parallel-aware version of the *FSEARCH* program, *MPI_FSEARCH*, has been implemented on Feynman. The message-passing interface (MPI) is one of the most favored packages for parallel computation. It was designed for high performance on both

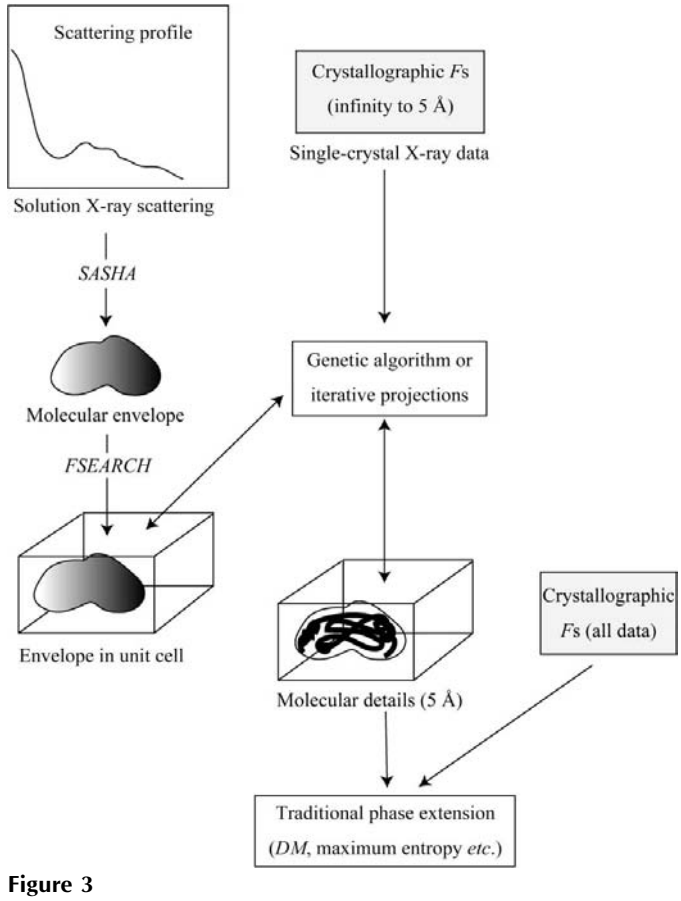

Figure 3
A flow-chart of the genetic algorithm and the iterative-projections method.

parallel machines and on network clusters. By calling a series of MPI library routines, it is straightforward to split the computing task into approximately equal pieces and distribute them to all CPUs. *MPI_FSEARCH* has been used successfully to perform an exhaustive six-dimensional search to phase very low resolution X-ray data using a molecular envelope (Liu *et al.*, 2003). The entire six-dimensional computation for a moderate-size (~100 kDa) protein usually takes about 2–3 h on this 128-processor cluster.

4. Phase extension

The low-resolution phases calculated from the correctly positioned molecular envelope need to be extended to higher (crystallographic) resolution. The phase extension is a very challenging problem and requires a substantial amount of effort in the development of new methods. The standard density-modification methods such as solvent flattening, histogram matching, non-crystallographic averaging and maximum entropy are known to be most effective for phase extension in the resolution range 5 Å or higher (see, for example, Ma & Chang, 2004). To bridge the gap between the envelope resolution (usually in the range 10–20 Å) and 5 Å, new methods such as the genetic algorithm (GA) and the iterative-projections method are proposed. They are possible but not exclusive tools. A flowchart of these methods is shown in Fig. 3.

4.1. Genetic algorithm

The work of Chacon *et al.* (1998) on protein structures retrieved from X-ray scattering with a genetic algorithm (GA) demonstrates that the GA method may play an important role in finding detailed structures within a low-resolution envelope. This approach might be extended so that phases could be determined to 5 Å resolution. We define an appropriate initial object within the envelope formed by spheres with reasonable dimensions (in accordance with the desired resolution, which is extended gradually in the process) and codify it into a binary array (*i.e.* 0s and 1s) forming the 'chromosomes'. The initial binary array is filled with uniform distribution (*i.e.* 1s) with the envelope. Some of the initial 'chromosome' population (binary array) will be gradually replaced with 0s within the envelope to ascertain the internal structure, thus extending the resolution. In terms of the GA, the objective will be to find the molecular structure with the structure factors ($F_{\rm calc}$) closest to the experimental structure factors ($F_{\rm obs}$). In the initial stage, the scattering data not originally used in the envelope calculation, *i.e.* from 20 to 10 Å, will also be included as a target profile. This would then enable us to build details of the molecular structure within the envelope. Once a good agreement is achieved at 10 Å, the procedure will be extended to crystallographic *F*s at 5 Å where traditional phase-extension methods apply. Each 'chromosome' is decoded into the corresponding spatial coordinate set and processed by an FFT procedure to obtain a theoretical $F_{\rm calc}$. These $F_{\rm calc}$ values are compared with the experimental data to determine the fitness value: in this case, the *R* factor. The 'chromosomes' are combined using genetic operators (crossover and mutation) in such a way that the structures with better fits have a higher probability of reproducing (selection pressure). The repeated application of genetic operators to the fittest 'chromosomes' increases the average fitness of the population with time and accordingly generates better models. As a GA starts with randomly generated 'chromosomes', its success rate can be improved by a multi-solution process, *i.e.* a large number of 'chromosome' sets will be generated and the resulting models after a GA will be selected by cluster analysis (Lunin *et al.*, 1990; Ockwell *et al.*, 2000).

4.2. Iterative-projections method

Another way to extend phases is by the iterative-projections method (Elser, 2003*a*). The problem of reconstructing the electron density within a known molecular envelope given a set of structure factors shares many features of the problem of reconstructing an isolated object (non-crystal) from its diffraction pattern. For the latter problem, the phases associated with the diffraction intensities can be completely recovered if the data is sampled on a sufficiently fine grid. This criterion is called 'oversampling' and roughly corresponds to there being some correlation between intensities at adjacent grid points. In the crystallographic setting, the oversampling criterion can be stated in terms of the fraction of the asymmetric unit-cell volume occupied by the envelope. In the case that this fraction is significantly less than one-half, the phases can be extended simply by imposing the standard support constraint used in the reconstruction of non-crystalline objects. In macromolecular crystallography this procedure is known as 'solvent flattening'. More typical of large unit-cell protein crystals are envelope fractions that are close to one half and even larger. The support constraint, even with perfect knowledge of the envelope, is then insufficient to extend the phases. A refinement of the standard support constraint to extend phases may be used even in this situation. The method would be to use the iterative-projections scheme developed by Elser (Elser, 2003*a*,*b*,*c*) which allows arbitrary constraints. In particular, one could supplement the support constraint with a projection onto the correct electron-density histogram within the envelope. This procedure has been tested on simulated data, with good results (Elser, 2003*a*). Clearly, the first step in developing this method will be to study the electron-density histogram at available resolutions; this can be performed by means of numerical experiments using data derived from solved structures.

5. Conclusion

In many cases, incorporation of heavy atoms or anomalous scatterers into the protein molecules can present serious technical difficulties. The procedure described in this paper does not require structural modifications of the protein or the presence of heavy atoms. Small-angle X-ray scattering (SAXS) or electron-microscopy (EM) data can be very useful

in providing a low-resolution envelope of the macromolecule. A method has been developed by performing a simultaneous six-dimensional search on orientation and translation to place the envelope in the crystallographic unit cell. The low-resolution phases calculated from the correctly positioned molecular envelope can be used as a good starting point for phase extension. Possible ways to extend phases to crystallographic data resolution have been proposed, but a substantial amount of work is needed to implement and test these ideas. The envelope-based technique may provide crystallographers with a new tool in facilitating the *ab initio* structure determination of macromolecules.

I thank Veit Elser and Darren Ockwell for useful discussion, Gunter Grossman and Samar Hasnain for providing the test data, and Qun Liu and Eddie Snell for help in producing the figures. This work is supported by the NIH grant RR01646.

References

Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Elser, V. (2003*a*). *J. Opt. Soc. Am. A*, **20**, 40–55.
Elser, V. (2003*b*). *J. Phys. A*, **36**, 2995–3007.
Elser, V. (2003*c*). *Acta Cryst.* A**59**, 201–209.
Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst.* A**58**, 72–74.
Grossmann, J. G. & Hasnain, S. S. (1997). *J. Appl. Cryst.* **30**, 770–775.
Hao, Q. (2001). *Acta Cryst.* D**57**, 1410–1414.
Hao, Q., Dodd, F. E., Grossmann, J. G. & Hasnain, S. S. (1999). *Acta Cryst.* D**55**, 243–246.
Hauptman, H. A., Guo, D. Y., Xu, H. & Blessing, R. H. (2002). *Acta Cryst.* A**58**, 361–369.
Jiang, W. & Ludtke, S. J. (2005). *Curr. Opin. Struct. Biol.* **15**, 571–577.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kollman, J. M. & Quispe, J. (2005). *J. Struct. Biol.* **151**, 306–314.
Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Cryst.* **34**, 33–41.
Liu, Q., Weaver, A. J., Xiang, T., Thiel, D. J. & Hao, Q. (2003). *Acta Cryst.* D**59**, 1016–1019.
Lunin, V. Y, Lunina, N. L. Petrova, T. E. & Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 896–903.
Lunin, V. Y, Lunina, N. L., Podjarny, A., Bockmayr, A. & Urzhumtsev, A. G. (2002). *Z. Kristallogr.* **217**, 668–685.
Lunin, V. Y., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.
Ma, C. & Chang, G. (2004). *Acta Cryst.* D**60**, 2399–2402.
Ockwell, D. M., Hough, M, Grossmann, J. G., Hasnain, S. S. & Hao, Q. (2000). *Acta Cryst.* D**56**, 1002–1006.
Rabinovich, D. & Shakked, Z. (1984). *Acta Cryst.* A**40**, 195–200.
Shen, Q., Hao, Q. & Gruner, S. M. (2006). *Phys. Today*, March, pp. 46–52.
Stuhrmann, H. B. (1970). *Acta Cryst.* A**26**, 297–306.
Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
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., Barberato, C. & Koch, M. H. J. (1997). *J. Appl. Cryst.* **30**, 798–802.
Urzhumtsev, A. G. & Podjarny, A. (1995). *Acta Cryst.* D**51**, 888–895.