*Pure & Appl. Chem.*, Vol. 66, No. 1, pp. 77–82, 1994.
Printed in Great Britain.

# Supercomputer simulation of small angle X-ray scattering, electron micrographs and X-ray diffraction patterns of macromolecular structures

**E. Pantos and J. Bordas**

*SERC, Daresbury Laboratory, Warrington WA4 A4D, Great Britain.*

***Abstract***

We describe two algorithms used in structural studies of proteins and macromolecular complexes. The first deals with simulation of Small Angle X-ray Scattering patterns. An approximate representation of the Debye formula based on the discretized histogram of pair distances allows the SAXS calculation of large macromolecules modelled by thousands of spheres to be performed in a few minutes of CPU time. The second algorithm concerns the computation of the mass projection of a structural model at a given projection angle. Such projections are used to reproduce structural features observed in electron micrographs of large biological structures. The accumulation of the Fourier transform intensity of mass projections of model structures for a series of angles is used to simulate X-ray diffraction patterns. Both algorithms have been parallelised and implemented on the Intel iPSC/860 hypercube.

## INTRODUCTION

It is a widely held belief that protein function is intimately related to protein structure. In the relatively few cases where a protein can be crystallised X-ray diffraction can be employed to determine the protein structure in its crystalline state to atomic resolution. In many more cases, however, only dilute solutions of a protein can be probed at moderate resolution (~3nm) through the technique of small angle X-ray scattering (SAXS) (ref.1-2). Advances in experimentation using synchrotron radiation (ref. 3) permit measurements at even higher resolution but the major constraint remains the one-dimentional nature of the data which is not in itself sufficient for unambiguous structure determination in three dimensions. On the other hand, the natural environment where most proteins exhibit their functional properties is in solution. Many proteins are of particular biological interest because of the macromolecular complexes they form either by themselves, e.g., microtubules, or in combination with other proteins. A prime example of a very large macromolecular structure is the muscle fibre (ref. 4). In such cases other structural techniques are employed, i.e., electron microscopy (ref. 5-6) and X-ray diffraction (ref. 7).

The use of SAXS has proved extremely valuable in structural studies both of individual molecules and of aggregates (ref. 8). Of specific interest are conformational changes upon change of environmental conditions (e.g., pH, counterion concentration, ligand attachment) or of the aggregation dynamics of macromolecular complexes. Time resolved SAXS and X-ray diffraction experiments bring additional information into play. Structural modifications provide the mechanism by which time evolution of a system develops. Modelling of protein structures, either in a static or in a time evolving form is crucial in visualising the processes that lead to specific functional behaviour. We outline below computer algorithms for the simulation of SAXS data, electron micrographs and X-ray diffraction patterns based on the modelling of structures by assemblies of spheres.

## SAXS SIMULATION

A frequently used approach to simulating SAXS patterns of large molecules is to build models of closely packed spheres and then use Debye's formula (ref. 9-10)

$$I(S)=\sum_{j=1}^{N} I_j(S)+2\sum_{j=1}^{N}\sum_{k=1}^{N} F_j(S)F_k(S)\sin(2\pi S r_{jk})/2\pi S r_{jk} \qquad j\neq k$$

to calculate the scattered intensity, I(S), for each value of the scattering vector modulus, S. The first sum gives the intensity for spheres in isolation, while the double sum gives the contributions from density-density correlations. $I_j$ is the scattered intensity by each sphere, $F_j(S)$ is the form factor for each sphere and $r_{jk}$ is the distance between pairs of spheres. The use of spheres is convenient both for the derivation of the formula and for model building purposes. It is equivalent to sampling the structure at the points where the spheres are positioned. The basic assumption is that the mass distribution is adequately and uniformly sampled for a given resolution.

Initial models for a starting overall shape are usually based on prior information from biophysical data, e.g. electron microscopy. Subsequent refinement is carried out by adding or repositioning spheres and recalculating the Debye formula. This manual procedure relies on the expertise of the modeller and is limited by the very large number of configurations that need to be tried. An elegant procedure described by Svergun and Sturhman (ref. 11) using a method based on spherical harmonics rather than groups of spheres is very powerful for fitting SAXS data of globular structures but it is rather limiting when more complex shapes are to be fitted. We have developed an algorithm for iterative calculation of successive configurations of assemblies of spheres on a predetermined grid capable of selecting good fits among millions of configurations but this will be detailed elsewhere (program DALAI, E.Pantos,D.Holden, J.West and J.Bordas, unpublished). We concentrate here on the implementation of a simplified form of the Debye formula and an approximation which makes possible the SAXS simulation for structures modelled by thousands of spheres in a few minutes of CPU time.

The computational task in the double summation of the Debye equation can be much reduced in size if all spheres are given the same radius and mass density. This implies that the mass density of the structure is uniform over the sampling grid, a reasonable assumption for protein molecules over the resolution range they are probed by SAXS. The form factor product $F_j(S)F_k(S)$ is now a constant for each value of S. The Debye formula takes now the form

$$I(S)=\sum_{j=1}^{N} I_j(S)+2F^2(S)\sum_{j=1}^{N}\sum_{k=1}^{N} \mathrm{sinc}\,(2\pi S r_{jk}) \qquad j\neq k$$

If we calculate the sinc function in advance for all the possible sphere pairs at each value of the scattering vector then the double sum is implemented as a simple vector summation of precalculated terms in a double loop, which can be very efficiently optimised by vectorising and parallelising compilers.

Thus far, we have not compromised the accuracy of the calculation. We have simply shifted the bulk of the computation to the initial stage of the algorithm. The number of spheres in the structure that can be treated is limited by memory requirements of the array for the sinc function (required to aid vectorisation and parallelisation) which scales as $K(n^2/2-n)$, where K is the number of scattering vector values and n is the number of spheres. To adequately model large structures requiring very large number of spheres a further approximation is called for. We use the distance histogram approach suggested by Glatter (ref. 10). We detail the procedure here because of the drastic effect it has both on memory allocation and on calculation time.

Pair distances are discretized in a histogram of bin size commensurate with the resolution of the data ( ~ 100 times smaller than the high resolution limit). This effectively "*fuzzes out*" the sampling grid by an amount that would not be detected in the resolution range of the simulation. The pair distance matrix of $r_{jk}$ values now becomes a one-dimensional array of distances weighted by the number of distances occurring in an interval of binsize, the bin population. The number of terms in the inner sum is now the number populated bins. The scattering formula becomes

$$I(S)=\sum_{j=1}^{N} I_j(S) + 2F^2(S)\sum_{k=1}^{Nbins} m(r_k)\mathrm{sinc}(2\pi S r_k)$$

where $m(r_k)$ is the bin population at pair distance $r_k$ and the limits of the sum are the number of distance bins. This method was used very effectively to interpret changes in the SAXS data of the iron carrying protein transferrin where domain movement has been shown to be involved in iron intake and release (ref. 12).

We are now well equipped to be more ambitious in the application of the code. The opportunity arose in connection with the interpretation of SAXS and electron microscopy data of microtubules. Microtubules are organelles which are found in all eukaryotic cells (ref. 13). They are involved in intracellular transport processes and in cell division. They can be described as hollow cylinders of some 24nm mean diameter and length of several microns. The basic building block is the tubulin heterodimer which is formed from the $\alpha$ and $\beta$ subunits. The axial sequence of $\alpha$–$\beta$ heterodimers are called protofilaments.

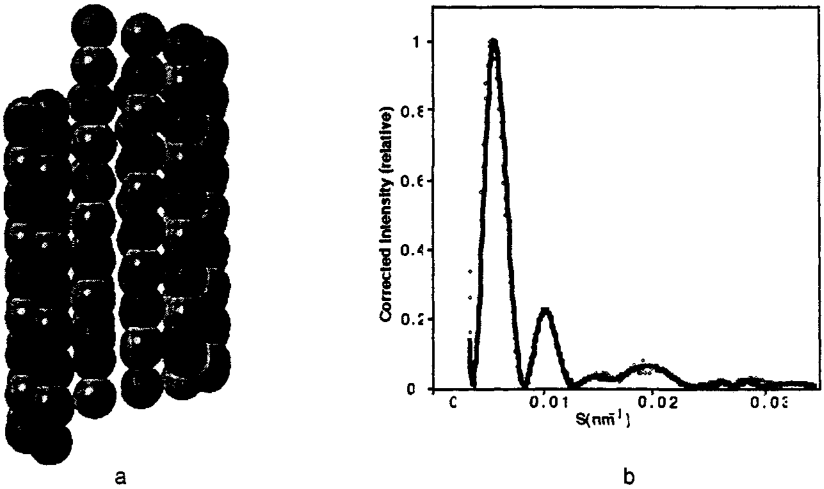
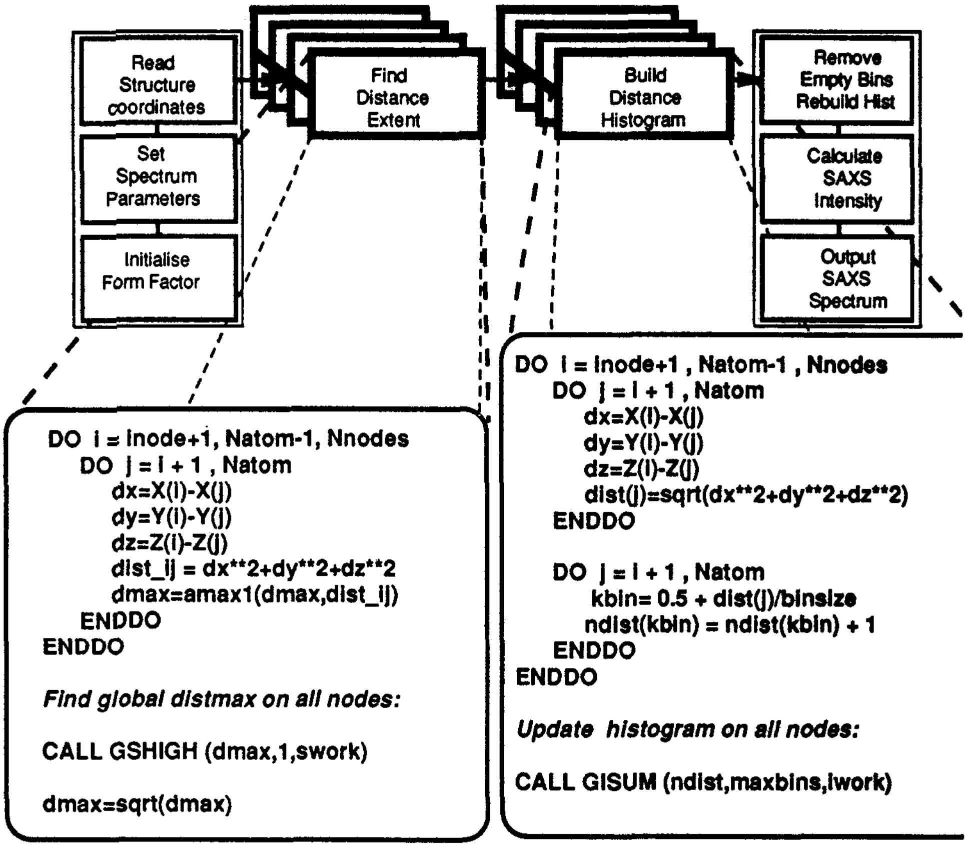


Figure 1. a) Solid sphere model of 3 turns of the 3-start helix of a 12-protofilament microtubule. The different shade spheres represent the $\alpha$ and $\beta$ tubulin monomers. b) The experimental (circles) and fitted (line) SAXS patterns of taxol microtubule. Each monomer has been modelled by a group of spheres in an ellipsoidal envelope with the major axis at an angle to the cylinder axis (ref. 5).

The tubulin monomers are arranged in a lattice of helical symmetry. Fig. 1a shows three turns of a 3-start helical model of a 12-protofilament microtubule with single spheres for the monomers. It can easily be observed that several pair distances repeat, e.g., along protofilaments, along the helical paths, and so on. If the single sphere monomer is substituted by the group of spheres used to fit the monomer, pair distances within each monomer repeat in every other monomer. The number of identical pair distances can be precalculated from the symmetry parameters of the model structure and the distance histogram population can be updated accordingly without having to compute every single pair distance.

The details of the structural parameters obtained from simulations of the SAXS data of taxol and W-tubulin microtubules are given in ref. 5. To construct a model of appropriate extent and resolution some 40000 spheres were used. The fitting procedure involved iterative tilting and scaling of the tubulin monomer in the lattice. The drastic improvement in execution time brought about by utilising symmetry, made possible interactive runs of hundreds of conformations in refining the pitch, diameter, number of protofilaments, helical start-number and orientation of the monomer (Fig. 1b).



```
DO i = inode+1, Natom-1, Nnodes
   DO j = i + 1 , Natom
      dx=X(i)-X(j)
      dy=Y(i)-Y(j)
      dz=Z(i)-Z(j)
      dist_ij = dx**2+dy**2+dz**2
      dmax=amax1(dmax,dist_ij)
   ENDDO
ENDDO

Find global distmax on all nodes:

CALL GSHIGH (dmax,1,swork)

dmax=sqrt(dmax)
```

```
DO i = inode+1 , Natom-1 , Nnodes
   DO j = i + 1 , Natom
      dx=X(i)-X(j)
      dy=Y(i)-Y(j)
      dz=Z(i)-Z(j)
      dist(j)=sqrt(dx**2+dy**2+dz**2)
   ENDDO

   DO j = i + 1 , Natom
      kbin= 0.5 + dist(j)/binsize
      ndist(kbin) = ndist(kbin) + 1
   ENDDO
ENDDO

Update histogram on all nodes:

CALL GISUM (ndist,maxbins,iwork)
```

Figure 2. Schematic breakdown of the SAXS simulation code highlighting the parts which are executed in parallel. The distance extent is calculated first to set the minimum binsize permitted. This step can be omitted if the maximum distance extent is known and the binsize fixed at the start or for subsequent iterations. The modules on the right and left run sequentially on the root node.

Fig. 2 gives a schematic breakdown of the algorithm which utilises the distance histogram approximation. It turns out that most of the processing time is spent on a rather simple task, the calculation of pair distances and the construction of the distance histogram. The parallelisation of this code is a straight forward case of coarse grain parallelisation in SIMD (Single Instruction Multiple Data) mode. Identical copies of the code are executed on different processors while the rest of the program is run on the root node. The array of sphere coordinates is apportioned to the different nodes for processing and the histogram values are brought together in the distance histogram array. Details of implementation of the DALAI code on the Intel iPSC/860 hypercube will be given elsewhere (ref. 14) together with a description of a graphics user interface which allows submission of the compute intensive part on a parallel machine while user interaction and graphics display of data is performed on a graphics workstation. The largest system processed so far is for 500000 particles used to simulate aggregation and aging phenomena in silicate solutions (ref. 15). The computation time on 32 nodes of the Intel iPSC/860 was ~4 hours.

## ELECTRON MICROGRAPH AND X-RAY DIFFRACTION PATTERN SIMULATION

A transmission electron micrograph or indeed an X-ray radiograph gives the projection of the mass of the object under investigation. To compute the mass projection all that is required is to calculate the intersections of rays parallel to the projection axis with the object volume. The length of the intersecting cords multiplied by the mass density corresponds to the mass penetrated by the incident radiation (electron or x-ray). To mass project a model structure at a given projection angle, coordinates and sphere radii are first normalised for the extent of the image dimensions, typically 1024x1024 pixels. It is a simple matter of trigonometry to calculate the length of the intersecting cords for each sphere. Pixels within a sphere "footprint" are weighted by the thickness (mass) of the sphere at that pixel. The mass value of that pixel is added to the previous value generated for pixels of any other spheres that are intersected by the same projection ray.

This simple algorithm has been employed to generate mass projections of microtubule models where, for a given number of protofilaments and other helical structure parameters, it was used to demonstrate the appearance of density fringes of characteristic repeat in cryo-electron micrographs (ref. 5,16). Fig. 3a shows the mass projection of the microtubule model of Fig. 1a. The number of spheres used in the microtubule EM simulations was relatively small (a few thousand). In a different application, however, the model making algorithm produces several millions of spheres. It was used to create structural models of the muscle fibre (ref. 17) consisting of three main structural components, the actin thin filaments, the myosin backbone and the myosin heads, some 20 million spheres in total. The objective was to simulate the X-ray diffraction pattern that would result from such a structure and compare it with experiment (ref. 7). This can be achieved by generating mass projections of the structure at a series of angles about the long axis, then Fourier transforming each projection and accumulating the Fourier intensities. The computational task is quite formidable as each projection requires several minutes of CPU time on a Convex-220 mainframe and some 60 projections at symmetry related angles are needed to match the experimental resolution. In addition, the procedure needs to be repeated for different structure conformations representing the time evolution from the rest state to the fully activated, tension producing state of the fibre. Fig. 3b shows the resultant difference diffraction pattern for two end states.

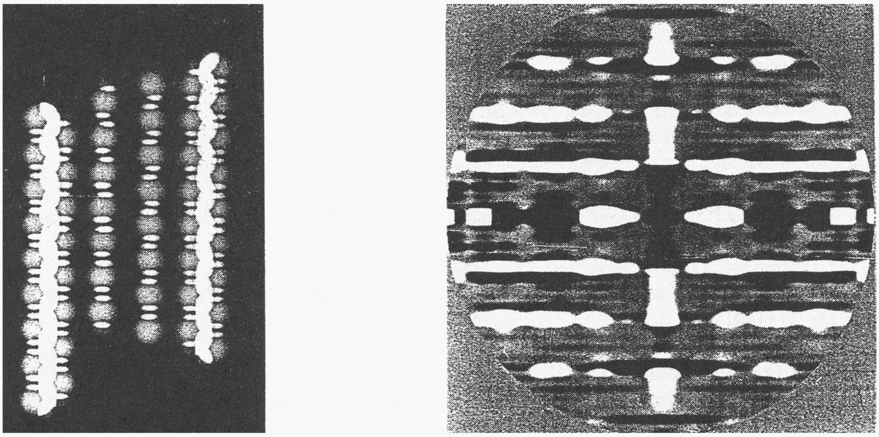


Figure 3. a) The mass projection of the microtubule model shown in Fig. 1a. Notice the density variation across a single unobstructed sphere and its overlap with its neighbour (top) and the higher density of line-of-site overlap at the centre and the edges of the cylinder. The slight off-axis tilt of the protofilaments results in the appearance of density fringes repeated at an interval of several cylinder diameters (ref. 5,14). b) Calculated difference diffraction pattern of muscle fibre. The structure extent was ~ 1 cubic micron and some 20000000 spheres were used to model the three components, actin filaments, myosin backbone and myosin heads.

An important feature of the mass projection algorithm is that it has no preference for the order in which the projection of individual spheres is computed. Sphere coordinates produced by the structure modelling program can be organised in any convenient way, e.g., in terms of the structural component they form or in terms of a slab of space they occupy. The code then reads as many coordinates as can be handled within the available memory, processes them, and then updates the two dimensional array containing the whole projection. A further time saving step can be taken: Since there are only a few different types of spheres, precalculated templates for a given sphere type and image resolution are simply copied onto the projection array at the appropriate projection centre.

Details of the parallel implementation have been described elsewhere (ref. 18). Apart from the initialisation and final output stage, the parallel mass projection code is a simple multiple copy of the sequential code. The load is balanced evenly and all processors are kept busy equally at all times. Processing time is now of the order of a few minutes for each projection. Just as for the parallel implementation of the SAXS simulation code it is a classic case of coarse-grain SIMD parallelisation with virtually no interprocessor communication. Apart from machine specific code for distributing data to each processor and for summing the partial results, the rest of the code is identical to the sequential one.

## CONCLUSIONS

We have shown that computer simulations of large macromolecular structures studied experimentally by three different but complementary structural techniques, small angle x-ray scattering, electron microscopy and x-ray diffraction, can now be realised in affordable time scales. By using simple coarse-grain parallelisation techniques we can take advantage of the raw computing power of parallel computers. The size of the structure or the structural detail can be increased by scaling up the number of processing elements. Both algorithms and their parallel implementations are applicable to any structural system which is adequately represented by a sphere model. We can now concentrate on developing structure modelling schemes. For the case of SAXS of proteins, of particular interest is the modelling of domain movements and real time interaction with the model being manipulated on a graphics workstation. For the case of the muscle fibre, the challenge is to include additional components, such as troponin and tropomyosin which have not been taken into account so far, and to improve the resolution of representing the actin and myosin molecules. The combination of high spatial or temporal resolution experimental data from synchrotron radiation sources and computer modelling of the kind we have described is a very promising one.

## REFERENCES

1. E.Madelkow, E.M.Madelkow and J.Bordas, *J.Mol.Biol.*, **167**: 179-186, (1983).
2. P.Matsudaira, J.Bordas and M.H.J.Koch, *Proc.NatlAccad.Sci.*USA, **84**: 3151-3155, (1987).
3. T.Ueki, *Nucl.Instr.Meth.Phys.Res.*, **A303**: 464-475, (1991).
4. J. M.Squire, *The structural basis of muscular contraction,* Plenum Press, New York, (1981).
5. J.M.Andreu, J.Bordas, J.F.Diaz, J.Garcia de Ancos, R.Gil, F.J.Medrano, E.Nogalez, E.Pantos and E.Towns-Andrews, *J.Mol.Biol.*, **226**: 169-184, (1992)
6. R.H.Wade, D.Chretien and D.Job, *J.Mol.Biol.*, **212**: 775-786, (1990) .
7. J.Bordas, G.P.Diakun, J.E.Harries, R.Lewis, J.Lowey, G.R.Mant, M.L.Martin-Fernandez and E.Towns-Andrews, *Adv. Biophys.* **15**: 27 (1991).
8. P.B.Moore, *J.Appl.Cryst.*, **21**: 675-680, (1988).
9. P.Debye, *Ann.Phys. (Leipsig),* **46**: 2927, (1915).
10. O.Glatter and O.Kratky, *Small Angle X-ray Scattering,* Academic Press, London, (1982).
11. D.I.Svergun and H.B.Stuhrmann, *Acta Cryst.*, **A47**: 736-744, (1991).
12. J.G.Grossman, M.Neu, E.Pantos, F.J.Schwab, R.W.Evans, E.Towns-Andrews, P.F.Lindley, H.Appel, W.G.Theis and S.S.Hasnain, *J.Mol.Biol.*, **225**: 811-819 (1992).
13. A.Amos and P.M.A.Eagles, *Microtubules in Fibrous Proteins,* pp. 215-246, J.M.Squire and P.J.Vibert, eds, Academic Press, London, (1987).
14. E.Pantos, C.E. Dean, P.C. Stephenson, G. J. Milne and H.F. van Garderen, Daresbury Laboratory Preprint DL/CSE/P27E, October 1993.
15. H.F. van Garderen, E. Pantos, W.H. Dokter, T.P.M. Beelen and R.A. van Santen, Daresbury Laboratory Preprint DL/CSE/P22, May 1993, and Model. Simul. Mater. Sci. Eng., in press.
16. R.H.Wade, D.Chretien and E.Pantos, *Electron Crystallography, Proc. NATO Workshop,* pp. 317-325, Kluwer, Dordrecht, The Netherlands, (1991) .
17. F.G.Diaz, E.Pantos, and J.Bordas, *Rev.Sci.Instrum.*, **63-68**: 859, (1992).
18. F.G.Diaz, H.K.F.Yeung, E.Pantos and J.Bordas, *Parallel Computing and Transputer Applications,* M.Valero, E.Onate, M.Jane and B.Suarez (Eds), pp.1070-1079, IOS Press/CIMNE, Barcelona, (1992).