A Genetic Algorithm for Low Resolution Protein Structure Determination

P. CHACÓN[1,2], F. MORÁN[2], J.F.DÍAZ[3], E. PANTOS[4,*] AND J.M. ANDREU[1]

[1]*Centro de Investigaciones Biologicas, C.S.I.C. Velzquez 144, 28006 Madrid, Spain*

[2]*Dpto.de Bioquimica y Biologia Molecular I, Facultad de CC.Qumicas. U.C.M. Ciudad Universitaria s/n, 28040 Madrid, Spain*

[3]*Laboratorium voor Chemische en Biologische Dynamica, Celestijnenlaan 200D Katholieke Universiteit Leuven, B-3001 Leuven, Belgie*

[4]*CLRC, Daresbury Laboratory, Keckwick Lane, Warrington WA4 4AD, UK.*

Summary: A genetic algorithm for iterative fitting of SAXS data is presented. The algorithm described produces fast convergence to a fittest model mass distribution compatible with experimental data. This method affords a dramatic reduction of processor time required by other SAXS fitting methods and can be applied to any kind of structure, the only requirement is the target profile. The effectiveness of the procedure is demonstrated with synthetic objects and by deriving the low resolution model of a known protein structure from their corresponding computed SAXS profile.

Introduction

It is well known that low resolution (>10Å) Small-Angle X-ray Scattering (SAXS) data of proteins or other macromolecules in solution can give valuable information on their overall shape. Simulation techniques have been developed for fitting experimental data by calculating the profiles of model structures. In principle, any structure can be approximated at any resolution by a set of spheres of small enough diameter and the solution scattering pattern of such a model structure can be calculated using the Debye formula [1].

The principles of the method have already been described [2,3]. It has been frequently used to compare experimental data with profiles obtained from models derived from crystallographic structures mainly for detecting changes between biological macromolecules in the crystalline state and in solution [4-9] and for the characterization of the low resolution structures of tubulin microtubules [10] and tubulin rings [11] in dilute solutions.

This is the direct scattering problem (Figure 1) which consists of the computation of the scattering function of a model structure. The quality of the results relies critically on the use of a-priori information from an expert user. In all these cases the solution of the inverse scattering problem is indirect, that is, it is based on the calculation of profiles of known structures which are compared to the experimental SAXS data or model structures built manually by the user, one at a time, or, in some cases [12] generated automatically by the computer program from a prescribed set of configurations. CPU-efficient algorithms exist now for speedy computation of the scattering profile of structures of even tens of thousands of scattering elements [13].

Despite the excellent results obtained in specific cases, the scope of wider applicability is limited. The number of iterations that can be performed to provide sufficient level of confidence in the uniqueness of the solution is limited by practical considerations. We focus this work on the inverse problem which consists in deducing the possible structure(s) or shape(s) or structural changes of a macromolecule given its X-ray scattering profile to a given resolution (Figure 1). Contrary to the direct scattering problem, the inverse problem can not be solved analytically; i.e., no "inverse" Debye formula can be constructed to yield 3D position coordinates from scattering profile data. Moreover, it is important to note that it is possible to find different models with the same profile to a given resolution i.e., the inverse

*To whom all correspondence should be addressed.

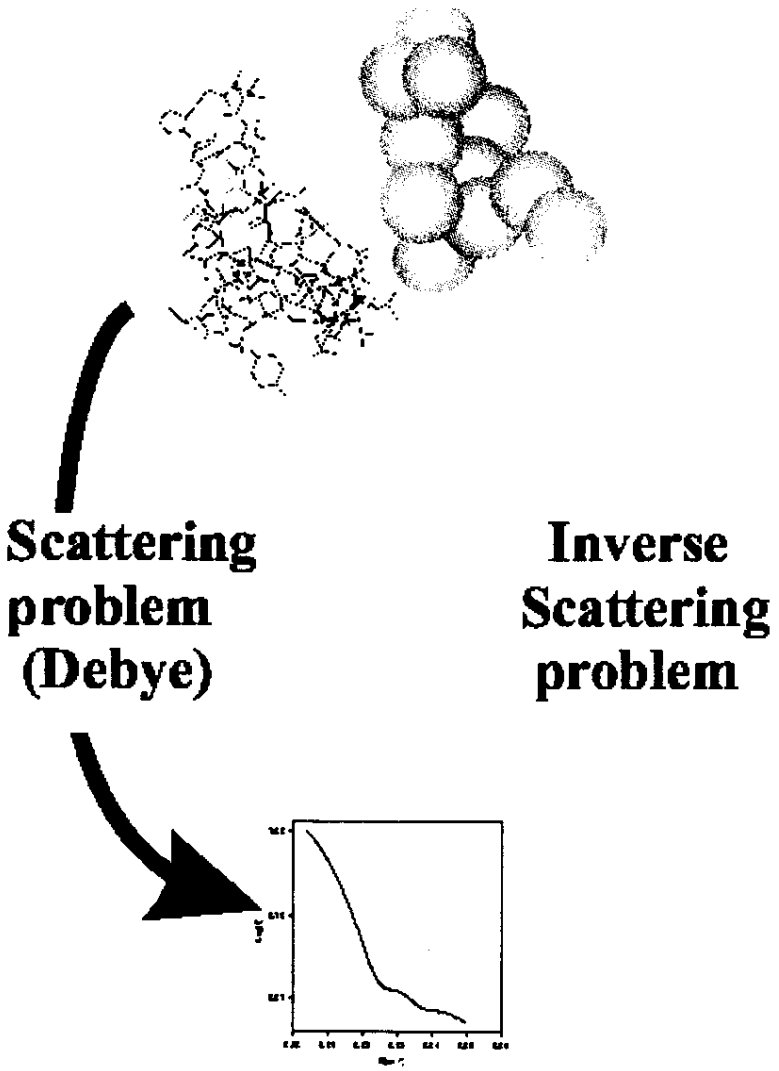

Fig. 1: Schematic representation of the inverse scattering problem in real space.

scattering problem has no unique solution. This intrinsic ambiguity adds great complexity but we will see below how this ambiguity can be reduced.

To this end, we have combined the Debye formula algorithm with an optimization tool, a genetic algorithm (GA). Genetic algorithms are search and optimization tools based on the natural evolution and genetics mechanism which have been applied very successfully to problems as diverse as robot trajectory generation to timetable and schedule organization [14,15]. Fruitful results in molecular modeling problems have been obtained recently using these techniques, such as in protein folding [16-20] RNA folding and secondary structure prediction [21] and docking and molecular recognition [22,23]. In other words, the genetic algorithm is a "black box" function optimization procedure with a wide scope of application.

The most important advantage of the GA approach is that the algorithm and its implementation into code is intrinsically very simple. There are no complicated mathematical formulae to be coded and no CPU-expensive functions to be computed. No problem-specific information about the solution needs to be predefined or identified although a-priori knowledge of maximum shape extent can be used to great advantage. The only requirements are

a) to be able to codify (map) the problem (in our case of spheres in a predetermined grid, it is codified into bit strings, bit on/off signifies the presence/absence of a sphere),

b) to define an objective goodness-of-fit function (in our case the sum of differences between experimental and calculated profiles).

Nothing else needs to be known about the problem. The aim of this work is to present and test this method in order to show its applicability to solving the inverse scattering problem. We will deal here with the essentials of the genetic algorithm implementation using a simple two-dimensional example for clarity of presentation. Further details are given elsewhere [24].

Algorithm description

1. Mapping of the configuration space into a bit string.

Figure 2a gives an example of a simple configurational space, a 7x9 regular grid of spheres masked by an ellipsoid, and the SAXS profile corresponding to one of the structures within it. The bit-string representation of this structure, the so-called chromosome, is

000 00100 0001000 0111110 0001000 0111110 0001000 00100 000

where the counting of bits (genes) is from left to right, top row to bottom row. Each one of the ON bits, corresponds to a unique coordinate position. The number of possible configurations is given by $N(m)=n!(n-m)!/m!$ where n is the length of the chromosome and m is the bit-sum (the number of spheres). It is shown graphically in Figure 2b.

Let us now consider the inverse problem: If the profile shown in Figure 2 is the "experimental data" how do we find the configuration that corresponds to it? An exhaustive search approach has been employed in a variant of program DALAI (Pantos unpublished) to model the SAXS profile of the tubulin dimer used to build tubulin superstructures [10,11]. This procedure is limited in scope because the exploration of all configurations becomes successively more difficult as the resolution, i.e., the number of spheres in a configurational space of given spatial extent, increases beyond reasonable limits. Searches have been made with configuration spaces not exceeding 31 spheres in total (2^{31}-1 =32-bit integer limit = 2,147,483,647 configurations). This limit can only be exceeded by including a "hard-core" of spheres always present in the model and consequently not counting towards the number of configurations generated but allowing the use of models with larger volume than that of just 31 spheres. The task becomes impossible for searches with no constraints on the minimum volume/shape and grid sizes commensurate with the resolution of the experimental data.

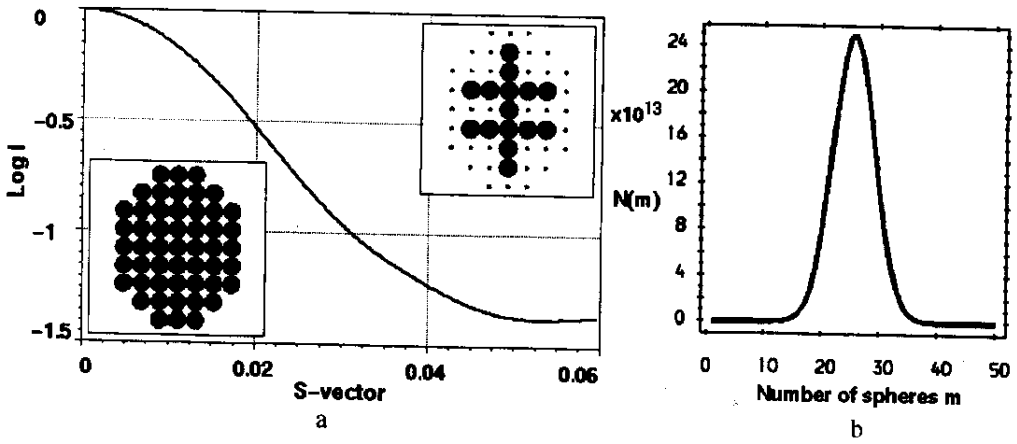

Fig. 2. a) The SAXS profile of the structure shown in the top-right insert within the configuration space of 51 spheres on a 7x9 grid with the corners of the rectangle masked-off (bottom-left insert). This structure can be found in 13 other symmetry equivalent positions. b) The number of configurations for mass in the range 1 to 51 spheres including symmetry related configurations.

For example, the number of possible configurations in a space of 10x10x10 nm occupied by hard spheres of 1 nm radius, typical of grid sizes needed for proteins, is 2^{1000}-1, a number too huge to be contemplated even by cosmologists! To illustrate the point, the simple loop counting up to the Avogadro number (2^{79}=6.044x10^{23})

```
DO i=1 , 2^79
  count=count+1
ENDDO
```

would take some 200000 CPU years to execute on a MIPS-R1000 195 MHz processor.

Let us take the SAXS profile in Figure 2a as a simple test case and let us assume that we have some idea about the maximum extent of the target object (from the Rg value, electron microscopy or some other technique) represented by the extent of the configuration space. The number of possible configurations is now 2^{51}-1 = 2.25x10^{15}. Each configuration corresponds to a potential solution for the mass distribution.

2. Generation of an initial population of chromosomes.

We now generate randomly a population of chromosomes each of them representing a different model structure within the configuration space. A set of 200 to 1000 is adequate in sampling the full mass range. Each chromosome is decoded into a set of coordinates by means of a look-up table containing all the possible (51 in this example) coordinates. The resulting structural models are processed by the SAXS simulation procedure to obtain the scattering profiles, which are then compared with the experimental data by computing the fitness value defined as the normalised sum of residuals

$$F = \sum | \log I_{calc} - \log I_{exp} | / n$$

where n is the number of points in the profile and I_{calc} and I_{exp} the computed and experimental intensities, respectively, normalised at the first point. The profile of the best structure and the fitness values of the initial population are shown in Figure 3. It is obvious that the starting best-guess structure is far from ideal.

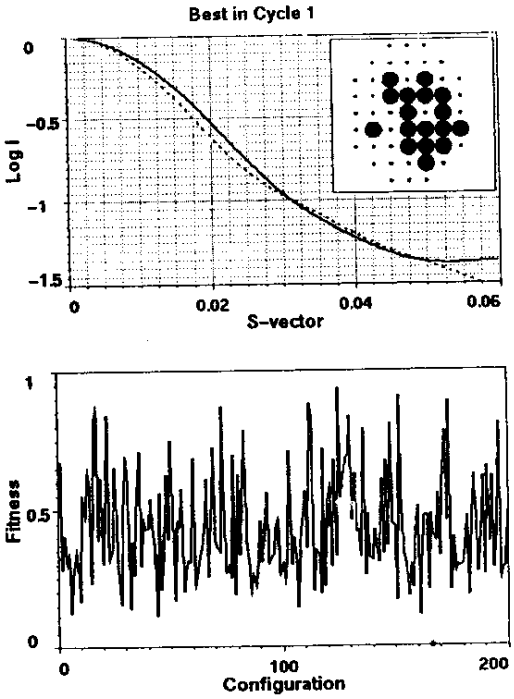

Fig. 3. Top: The profile of the first best-guess configuration (dashed line) compared with the target profile (solid line). Both profiles are normalised at the first point. The inset shows the first best guess structure of 17 spheres. Bottom: The fitness values of the starting guess structures. Their Rg and mass values span the range 7-27 and 4-36, respectively.

3. Evolution of the breeding stock.

The algorithm now proceeds by selecting the elite portion of the current population (typically 10-30% of the total) which is used as the "breeding stock". Chromosomes are selected at random from this set and recombined by operators simulating genetic mechanisms: a) Uniform crossover takes two individuals, so-called parents, and mixes their genes randomly. b) Random mutation changes genes of offspring from 0 to 1 or from 1 to 0. The new set of offspring is processed by the SAXS simulation procedure after discarding accidental identical twins. A new ranking order is established and offspring that fit the target profile best are promoted into the breeding stock while unfit ancestors are discarded.

Table I. The 10 best configurations from generation 12. Columns 2 to 5 give the fitness value, the radius of gyration, the number of spheres and the configuration bit-string.

Rank	F_v	Rg	m	Configuration bit string
1	8.839	15.918	14	00000100 00101001 01111000 10100011 01010000 00000000 000
2	5.764	15.622	15	00000100 00101101 00111000 11101001 01000000 10000000 000
3	5.743	15.333	14	00000000 00111100 01010100 01101001 01010000 01000000 000
4	5.478	15.597	14	00000100 00101001 01111000 10101011 01000000 00000000 000
5	5.277	15.821	15	00000100 00101101 01111000 10101011 01000000 00000000 000
6	5.161	15.753	14	00000100 00101101 00110000 11101001 01000000 10000000 000
7	5.059	16.559	15	00000000 00101111 11110000 10101011 01000000 00000000 000
8	4.849	16.055	16	00001010 00101000 01111000 11101011 01000000 00000100 000
9	4.749	15.969	14	00000100 00101001 01111000 10100011 01000000 00000100 000
10	4.499	15.097	14	00000000 00101000 01011100 01101001 01010000 11000000 000

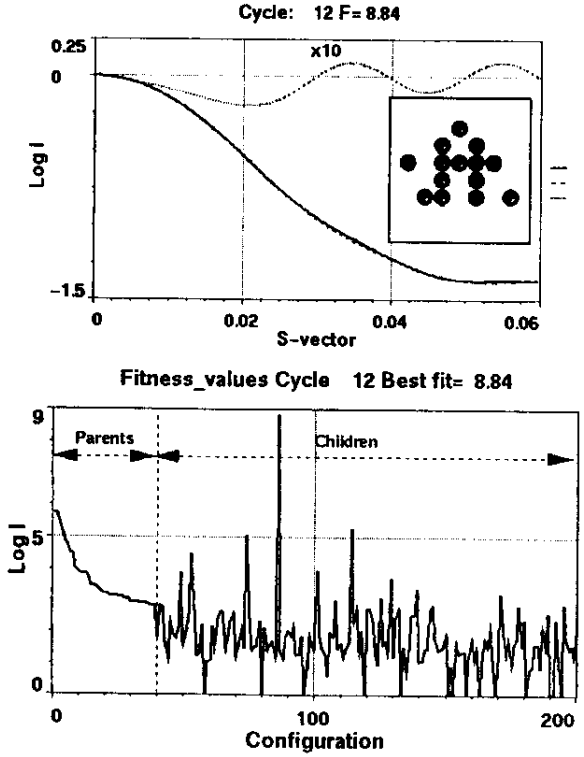

Fig. 4. Top: Target profile (line) and current best fit (dashed line) produced at iteration 12. The dotted line shows the residual amplified x10. The current best structure is shown in the insert. Bottom: The fitness values of the current population. The top 30% make up the breeding stock ranked according to fitness. The rest are the new offspring, several of which will be promoted into the breeding stock in the next cycle. A new leader has been born with fitness value = 8.84. Configurations with zero fitness value are those which have been rejected on the basis of their mass or Rg value being outside a preset range.

This process is repeated in subsequent cycles. It is easy to see that the population will evolve over successive generations towards a global maximum.

Table I and Figure 4 reflect the situation at an early stage in the evolution of the reproductive population. The algorithm has already generated top configurations with mass and Rg values not far from those of the target. The similarity with the target configuration is now beginning to emerge. This early success can be used to advantage by narrowing down the mass range and Rg value of configurations needing to be processed by the SAXS simulation procedure in subsequent generations. Figure 5 summarises the steps followed so far.

The graphs in Figure 6 are instructive on how convergence to the best fit is reached. In the early stages better candidates are produced rapidly. Improvements in the lead candidate occur at successively longer intervals as the breeding stock improves steadily (Darwinian selection pressure). It is this feature of the genetic algorithm that makes it so powerful. The procedure inevitably leads to the rebirth of chromosomes that have existed in a previous cycle, only to be discarded again. This apparent wastefulness of the algorithm is offset by the overall improvement achieved in the end. The total number of configurations required to be examined to reach the perfect fit for this example (Figure 7) was nearly $2x10^6$ (generated in 700 iterations), six orders of magnitude smaller than the number of configurations of mass 15, N(15)= $3.2x10^{12}$ and 9 orders of magnitude smaller than the total number of configurations an exhaustive search would have had to process to be sure of finding the best fit. The lower graph in Figure 6 tells us that as time passes fewer and fewer offspring are likely to be better than their parents. The reason for this is that the number of "better" configurations that can be produced is decreasing rapidly, they are more

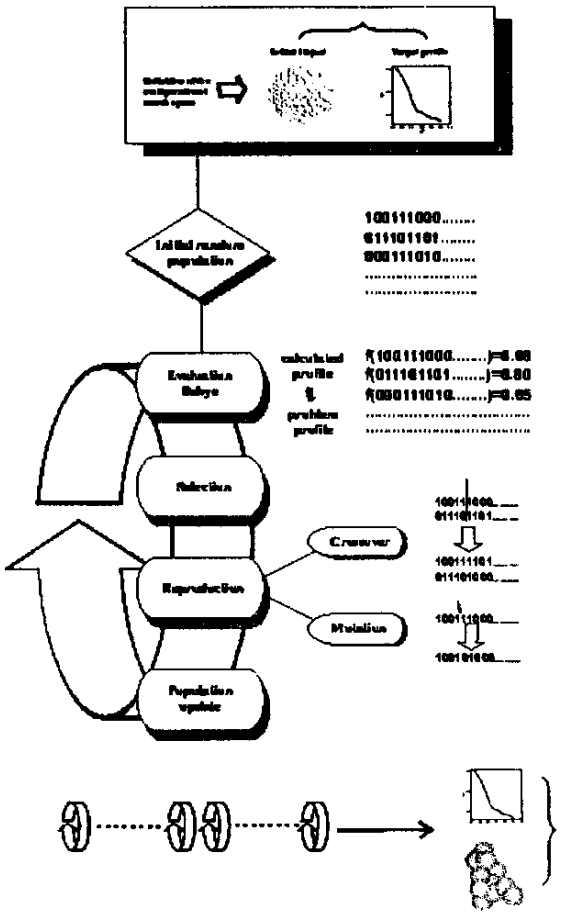

Fig. 5: Schematic of the genetic algorithm implementation

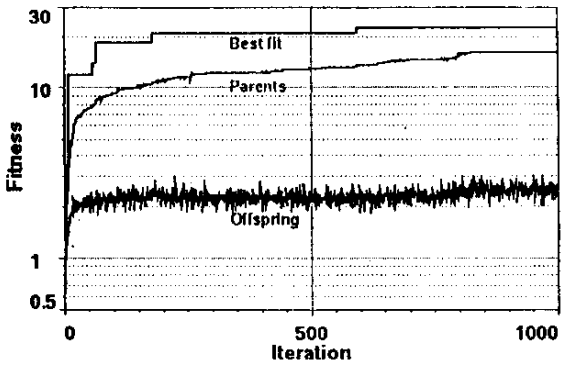

Fig. 6. Semilog plots of the evolution of the best fit (upper curve), the average fitness value of the parents (middle curve) and the average fitness value of the ofrspring (lower curve). The quality of the breeding stock improves with iteration number.

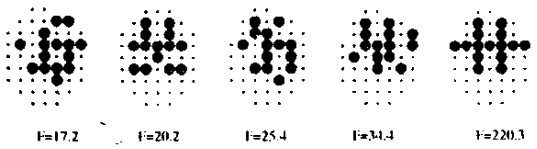

Fig. 7. Evolution of successive best-fit structures. The final structure is one of the 13 symmetry related ones to the target object in Figure 3a. Independent runs of the program (different starting seeds for the random number generator) may result to another solution (target shape in different orientation) at different times.

difficult to find. The search for better fits is driven mainly by the mutation process. At this point the system has converged and the search for better fits is driven mainly by the mutation process.

Application to a known protein structure

A more realistic example of testing the efficiency and effectiveness of the algorithm would be one where the size of the configuration space is representative of actual target structure dimension and the resolution of experimental data. Figure 8 shows a configuration space of 254 spheres of radius 6Å. The spheres are hexagonal packed to provide the best mass sampling and are all within an ellipsoid of revolution of major/minor axes of 100/70Å. This space is large enough to contain the crystal structure of two adjacent fibronectin type III repeats from the Drosophila neural cell adhesion molecule neuroglian [25], 1cfb entry in Brookhaven database).

In place of experimental data we use the simulated SAXS profile produced by program DALAI using all 1914 atoms in the 1cfb.pdb file. We limit the range to be fitted to $S=0.06Å^{-1}$, corresponding to real space resolution of $1/2S= 8.3Å$.

Figure 9 shows the SAXS fit (fit and target profile can hardly be distinguished) and the average fitness of the parent population with the values for the best fits at a given iteration superimposed as spikes. Successively better fits improve the "genetic stock" until little further improvement can be

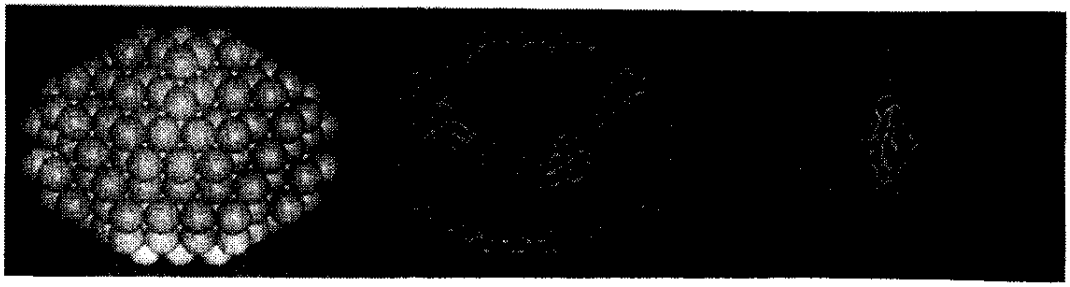

Fig. 8. a) The configurational space of 254 spheres of radius 6Å. b) and c) Connoly surface representation of the configuration space with ribbon representation of target structure embedded in it at two orthogonal projections.

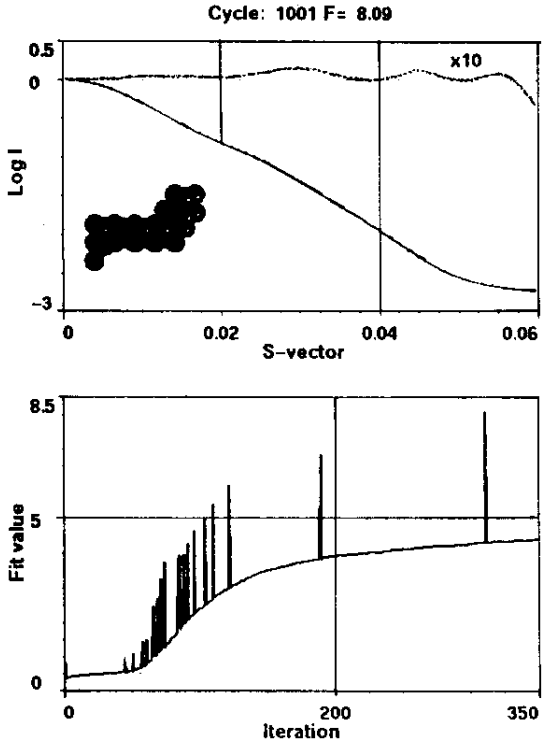

Fig. 9. Top: Best fitted profile and residual (dashed line x10) after 330 iterations. The residual (dotted line) is scaled x10. Bottom: Average population fitness as a function of iteration number. The spikes superimposed on the smooth graph give the fit value of the best fit attained at a given iteration number.

achieved. Despite the enormous size of the configuration space (2^{254-1}) the pertinent shape features of the structure at that resolution have been quickly identified. Only 29 spheres out of the 254 are retained. It is also clear that the size of the spheres is too large to accommodate fine details in the periphery of the molecule. This is reflected in ripples in the residual curve in Figure 9.

Increase in the quality of fit can now only be achieved by increasing the resolution of the configuration space (smaller spheres in a finer grid). Alternatively, as the low resolution shape of the structure is now known, the best fitted structure can now be used as a "mask", augmented by a margin around it and imposed on a finer resolution configuration space to define a subset of the spheres that would be contained in the original ellipsoid. This reduces the memory and processing time requirements significantly without loss in information to be extracted (Figure 11).

Conclusions

A general method for estimating the low resolution structure of macromolecules from the solution SAXS profiles has been presented. The method consists of fitting the scattering profile computed from packed-sphere models of the molecule using the Debye formula. The models are optimized by means of a genetic algorithm that searches the huge space of possible mass distributions.

Genetic algorithms have proved to be highly robust under varying parameter values and problem variables. The use of random choice in guiding the search is a built-in advantage of the GA approach guaranteeing impartiality. This stochastic character allows the resolution of ambiguities since each run of

Fig. 10: Three projections of the fitted structure with the crystal structure shown in ribbon representation and the best fit of 6Å spheres as a Conolly surface produced by program Insight II.

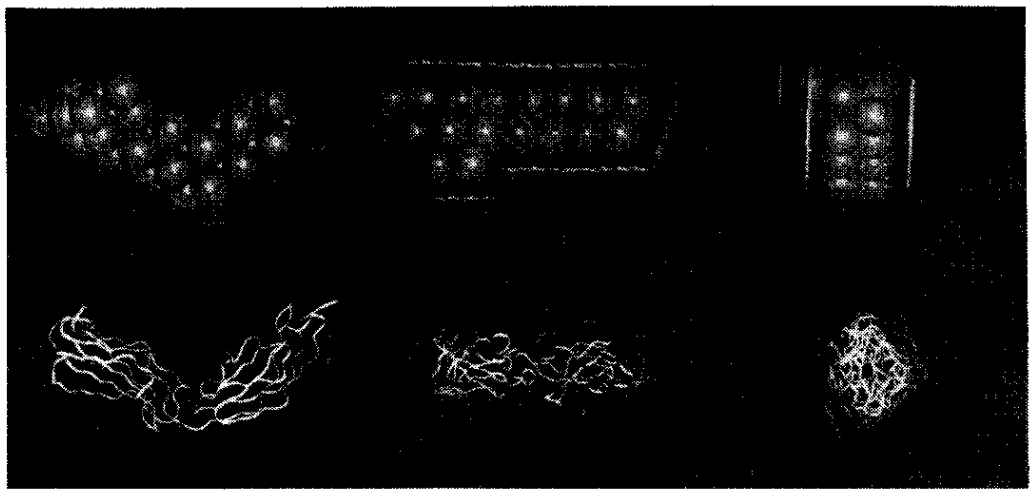

Fig. 11 Top: Three orthogonal projections of finer resolution configuration space of 536 spheres of radius R=3Å produced by adding a surface layer of 12Å on the best fit from the previous run. Bottom: The best fitted structure (F=143) obtained by rerunning the algorithm.

the algorithm is an independent search and, consequently, it is possible to obtain different, in principle uncorrelated, structures resulting in a good fit.

The various steps employed in the procedure have been outlined in some detail using a synthetic two-dimensional example for the shake of clarity. Application to a three-dimensional case of a known protein structure demonstrates that reliable results can be obtained describing the size and shape of the target structure with ambiguity limited only by the resolution and quality of the data rather than user expertise and a-priory knowledge.

The reliability of the method with protein structures of different sizes and shapes (globular, bi-lobed, dimeric, horseshoe shaped and two-domain proteins) is demonstrated elsewhere [24] where the effect of experimental noise in the data is also investigated.

References

1. P. Debye, *Ann. Phys. (Leipzig)*, **46**, 809 (1915).
2. O. Glatter and Kratky. Small Angle X-ray Scattering. Academic Press London (1982).
3. C.R. Cantor and P.R. Schimmel, Biophysical Chemistry. Part II. Techniques for the study of biological structure and function. W.H. Freeman, New York. 811-819(1980).
4. J.G. Grossmann, M. Neu, E. Pantos F.J. Schwab, R.W. Evans, E. Townes-Andrews, P.F. Lindley, H. Appel W. Thies, and S.S., Hasnain, *J. Mol. Biol.*, **225**, 811 (1992).
5. R.W. Evans, J.B. Crawley, R.C. Garrant, J.G. Grossmann, M. Neu, A. Aitken, K. J. Patel, A. Meilak, C. Wong, J. Singh, A. Bomford, and S.S. Hasnain, *Biochem.* **33**, 12512 (1994).

6. S.J. Perkins, A.S. Nealis, B.J. Sutton, and A. Feinstein, *J. Mol. Biol.*, **221**, 1345 (1991).
7. S .J. Perkins, K.F. Smith, J.M. Kilpatrick, J. E. Volanakis, and R.B. Sim *Biochem. J.*, **295**, 87 (1993).
8. (a) M.O. Mayans, W.J. Coadwell, D. Beale, D.B.A. Symons, and S.J. Perkins *Biochem. J.*, **311**, 283 (1995).
   (b) E. Pantos, and J., Bordas, *J. Pure & Appl. Chem.*, **66**, 77 (1994).
9. A.J. Beavil, R.J. Young, B.J. Sutton, and S.J. Perkins, *Biochem.*, **34**, 14449 (1995).
10. J.M. Andreu, J. Bordas, J.F. Díaz, J. Garcia de Ancos, R. Gil, F.J. Medrano, E. Nogales, E. Pantos, and E. J., Towns-Andrews *Mol. Biol.*, **226**, 169 (1992).
11. J.F. Díaz, E. Pantos, J. Bordas and J.M. Andreu *J. Mol. Biol.*, **238**, 214 (1994).
12. C.E. Dean R.C. Denny P.C. Stephenson, G.J. Milne and E. Pantos, *J. de Phys. III, Colloque*, **C9**, 445 (1994).
13. E. Pantos, H.F. van Garderen, P.A.J. Hilbers, T.P.M Beelen. and R.A van Santen. *J. Mol. Struct.*, **383**, 303 (1996).
14. D.E. Golberg, Genetics Algorithms in Search, Optimisation and Machine Learning Addison-Wesley, Reading , MA(1989).
15. L. Davis, Editor of Handbook of Genetics Algorithms Van Nostrad Reinhold, New York. (1991).
16. T Dandekar, and P. Argos, *Protein Eng.*, **5**, 637 (1992).
17. T Dandekar, and P. Argos, *J. Mol. Biol.*, **236**, 844 (1994).
18. T Dandekar, and P. Argos, *J. Mol. Biol.*, **256**, 645 (1996).
19. S. Sun, *Biophys. J.*, **69**(2), 340 (1995).
20. J.T. Pedersen, and J. Moult *Proteins.*, **23**, 454 (1995).
21. F. H. D. Van Batenburg, A. P. Gultyaev, and C.W. A. Pleij, *Theor. Biol.*, **174** (3), 269 (1995).
22. C. M. Osihiro, I.D. Kuntz, and Scott J. Dixon, *J. of Computer-Aided Molecular Design*, **9** 113 (1995).
23. G. Jones, P. Willett, and R. C. Glen, *J. Mol. Biol.*, **245** (1) 43(1995).
24. P. Chacón F. Morán J. F. Díaz E Pantos and Andreu , Submitted to *Biophys.J.* (1997).
25. Huber-AH; Wang-YM; Bieber-AJ; Bjorkman-PJ AD, *Neuron.*; **12**(4): 717 (1994)