Low-Resolution Structures of Proteins in Solution Retrieved from X-Ray Scattering with a Genetic Algorithm

P. Chacón,*# F. Morán,# J. F. Díaz,§ E. Pantos,¶ and J. M. Andreu*

*Centro de Investigaciones Biológicas, C.S.I.C. Velázquez 144, 28006 Madrid, Spain; #Departamento de Bioquímica y Biología Molecular I, Facultad de C.C. Químicas, U.C.M. Ciudad Universitaria s/n, 28040 Madrid, Spain; §Laboratorium voor Chemische en Biologische Dynamica, Celestijnenlaan 200D Katholieke Universiteit Leuven, B-3001 Leuven, Belgium; ¶Daresbury Laboratory, Warrington WA4 4AD, England

ABSTRACT Small-angle x-ray solution scattering (SAXS) is analyzed with a new method to retrieve convergent model structures that fit the scattering profiles. An arbitrary hexagonal packing of several hundred beads containing the problem object is defined. Instead of attempting to compute the Debye formula for all of the possible mass distributions, a genetic algorithm is employed that efficiently searches the configurational space and evolves best-fit bead models. Models from different runs of the algorithm have similar or identical structures. The modeling resolution is increased by reducing the bead radius together with the search space in successive cycles of refinement. The method has been tested with protein SAXS ($0.001 < S < 0.06$ Å^{-1}) calculated from x-ray crystal structures, adding noise to the profiles. The models obtained closely approach the volumes and radii of gyration of the known structures, and faithfully reproduce the dimensions and shape of each of them. This includes finding the active site cavity of lysozyme, the bilobed structure of γ-crystallin, two domains connected by a stalk in βb2-crystallin, and the horseshoe shape of pancreatic ribonuclease inhibitor. The low-resolution solution structure of lysozyme has been directly modeled from its experimental SAXS profile ($0.003 < S < 0.03$ Å^{-1}). The model describes lysozyme size and shape to the resolution of the measurement. The method may be applied to other proteins, to the analysis of domain movements, to the comparison of solution and crystal structures, as well as to large macromolecular assemblies.

INTRODUCTION

Small-angle x-ray scattering (SAXS) data of proteins and other macromolecules in solution can give valuable information on their size and shape parameters. SAXS using synchrotron radiation gives access to time-resolved low-resolution macromolecular structure and permits the study of structural changes from the experimental data, by modeling procedures. 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 (Debye, 1915). The procedure has been widely used in the past to derive the scattering profiles of many simulated structures (Witz et al., 1964; Cantor and Schimmel, 1980; Glatter and Kratky, 1982), and there are now CPU-efficient algorithms for the speedy computation of the scattering profile of structures of even tens of thousands of scattering elements (Pantos et al., 1996).

This work focuses on the inverse scattering problem, which consists of deducing the possible structure(s) or shape(s) or structural changes of a macromolecule from its x-ray scattering profile to a given resolution. Contrary to the direct scattering calculation, the inverse problem cannot be solved analytically, i.e., no "inverse Debye" formula can be constructed to yield 3D position coordinates from scattering data. Moreover, different models can present practically identical profiles to a given resolution, that is, in principle the inverse scattering problem has no unique solution. We will show how this ambiguity can be reduced. Several methods have been developed to date for extracting structural information (shape) other than what classical SAXS parameters offer (e.g., radius of gyration, R_g). These procedures rely on an expert (human) modeler who generates and refines models compatible with experimental data by trial and error through an educated guess (Curmi et al., 1988; Pilz et al., 1990; Garrigos et al., 1992; Wakabayashi et al., 1992; Díaz et al., 1994; Pantos and Bordas, 1994; Fujiwara et al., 1995). This manual procedure is severely limited in its scope by the large number of configurations that need to be manually constructed and tested. It has frequently been used in comparing 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 (Grossmann et al., 1992, 1997; Evans et al., 1994; Perkins et al., 1991, 1993; Mayans et al., 1995; Beavil et al., 1995). The quality of the results relies critically on the use of a priori reasoning and user expertise. In all of these cases the solution of the inverse scattering problem is indirect, that is, they are based on the calculation of profiles of known structures that are compared to the experimental SAXS data, or model structures manually built by the user, one at a time, or also generated automatically by the computer program from a prescribed set of configurations. A Monte Carlo algorithm has recently been developed for rapid computation of scattering profiles of models made from a variety of building blocks such as

Received for publication 3 November 1997 and in final form 5 March 1998.

Address reprint requests to Dr. J. M. Andreu, Centro de Investigaciones Biológicas, C.S.I.C., Velazquez 144, 28006 Madrid, Spain. Tel.: 34-91-564-4562, X4381; Fax: 34-91-562-7518; E-mail: j.m.andreu@fresno.csic.es.

0006-3495/98/06/2760/16 $2.00

prisms, ellipsoids, cylinders, and helices (Henderson, 1996). The presence of an expert user is an intrinsic assumption of the algorithm.

A direct approach is based on expanding the shape function in terms of spherical harmonics (Stuhrmann, 1970; Svergun and Stuhrmann, 1991; Grossmann et al., 1993, 1997; Svergun et al., 1994a,b). It is a powerful method for fitting the SAXS profiles of structures with shapes that can be described by a low number of spherical harmonics, but it is rather limiting when more complex shapes are to be fitted. An elegant related method using icosahedral harmonics has been applied to the study of viral structures (Zheng et al., 1995). Ingenious as they are, and despite the excellent results obtained in specific cases, these methods are limited in the scope of their applicability.

To a first approximation, the inverse scattering problem can be reduced to the search for a model compatible with the scattering profile in a predefined region of space discretized in finite particles (typically spheres). This procedure consists of two steps: 1) the computation of the scattering profile for each possible combination of these elements, and 2) the selection of the one that best fits the experimental profile.

An exhaustive search approach using a CPU-efficient computational method has been employed in a variant of the program DALAI (Pantos, unpublished observations) to characterize the low-resolution structures of tubulin microtubules (Andreu et al., 1992) and tubulin rings (Díaz et al., 1994) in solution. This method selects the best available configuration from a large set of candidates within a configurational space. This direct 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 extent, increases beyond reasonable limits in computer memory and processing time. For example, the number of possible configurations in a space discretized into 79 elements is close to Avogadro's number ($2^{79} = 6.044 \times 10^{23}$), a number too huge to be contemplated. Even if unsuitable configurations (e.g., symmetry-related ones or those with too few or too many spheres) are excluded from the computation of the Debye formula, those remaining to be examined are still too many, even to count, never mind processing them. In fact, the exhaustive search with built-in criteria for avoiding computation of unsuitable configurations (radius of gyration rejection test), using parallel implementation of the DALAI code (Dean et al., 1994), has only been attempted with configuration spaces not exceeding 31 spheres in total (the 32-bit positive integer limit gives $2^{31} - 1 = 2{,}147{,}483{,}647$ possible configurations). This limit can only be exceeded by including a "hard core" of spheres that are always present in the model, and consequently not counting toward the number of configurations generated, but allowing the use of models with larger volumes than that of 31 spheres.

The number of iterations that can be performed to provide a sufficient level of confidence in the solution is limited by practical considerations for the reasons explained above. To overcome this problem, it is necessary to apply an efficient search procedure. To this end, we have combined the Debye formula algorithm used in DALAI (Pantos and Bordas, 1994) with an optimization tool: a genetic algorithm (GA). GAs are search and optimization tools based on natural evolution and genetic mechanisms (Goldberg, 1989; Davis, 1991). In other words, the GA is a general optimization procedure with a wide scope of application, including the solution of several biological problems (Forrest, 1993; Willet, 1995). Fruitful results in molecular modeling problems have been obtained recently using these techniques, such as in protein folding (Dandekar and Argos, 1992, 1994, 1995; Sun, 1995; Pedersen and Moult, 1995, 1997), RNA folding and secondary structure prediction (Van Batenburg et al., 1995; Gultyaev et al., 1995; Ogata et al., 1995), docking and molecular recognition (Jones et al., 1995, 1997), and refinement of NMR solution structure (Li et al., 1997).

The most important advantage of the GA approach is that the algorithm and its implementation are 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 size can be used to great advantage. The only requirements are 1) to be able to codify (map) the problem into bit strings (spheres in a predetermined grid, bit on/off signifies the presence/absence of a sphere), and 2) to define an objective goodness-of-fit function (in our case the reciprocal square root of the sum of squared differences between experimental and calculated profiles).

The use of random choice in guiding the search is a built-in advantage of the GA approach guaranteeing objectivity. The stochastic character of the GA allows the resolution of ambiguities, because each run of the algorithm is an independent sampling of the configurational space, and, consequently, it is possible to obtain different, in principle uncorrelated, structures resulting in a good fit. Finally, genetic algorithms have proved to be highly robust under varying parameter values and problem variables. This work was aimed at building and testing a genetic algorithm method of x-ray scattering simulation and at demonstrating the feasibility of numerically solving the inverse scattering problem in general terms. We present the results and performance achieved, using as input simulated scattering profiles calculated from known protein structures of diverse shapes, as well as the experimental SAXS profile of lysozyme, to validate the general applicability of the algorithm. We show that the low-resolution structure of proteins in solution can be retrieved, in principle, from their x-ray scattering profiles.

MATERIALS AND METHODS

The x-ray scattering analysis method developed is based on the combination of the fast calculation of scattering intensities and of searching for the mass distribution that bests fits the scattering profile using a genetic algorithm. The Debye calculation of x-ray scattering and the approximation

using the algorithm of Pantos and Bordas (1994) will first be outlined. Then the GA implementation of the inverse scattering problem will be described, followed by the method of applying the procedure.

X-ray solution scattering

In SAXS measurements of monodisperse systems, the normalized intensity distribution, $I(S)$, is only determined by the structure of the object. S is the scattering vector modulus, $S = 2(\sin\theta)/\lambda$, where 2θ is the scattering angle and λ is the x-ray wavelength. Using numerical and, in some special cases, analytical methods, it is possible to calculate $I(S)$ from the 3D structure of a given macromolecule. This can be addressed through calculation of the pair distance distribution function $p(r)$ or by direct calculation of scattered intensities (for reviews on SAXS theory, see Pessen et al., 1973; Pilz et al., 1979; Glatter and Kratky, 1982).

One of the simpler (and therefore more frequently used) methods for calculating the scattered intensities consists of the approximation of a given molecular structure by a finite number of homogeneous spheres. Considering a structure of homogeneous electron density to the resolution of the measurement, the $I(S)$ for a model formed by n beads can be calculated by the Debye formula (Debye, 1915), which reduces to

$$I(S) = I_o(S)\left[n + 2\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\sin(2\pi S r_{ij})}{2\pi S r_{ij}}\right] \quad (1)$$

where the constant $I_0(S)$ is the intensity scattered by each identical sphere, and r_{ij} is the distance between pairs of beads (Pantos et al., 1996). The double summation in Eq. 1 results in a long calculation time that depends on the number of beads. To increase the efficiency of this calculus, we can write the intensity as

$$I(S) = I_o(S)\left(N + 2\sum_{i=1}^{N_{bins}} g(r_i)\frac{\sin(2\pi S r_i)}{2\pi S r_i}\right) \quad (2)$$

where the function $g(r_i)$ is the pair-distance histogram generated by sorting the distance-distance matrix in N_{bins} clusters (for details see Pantos and Bordas, 1994; Pantos et al., 1996). Now the calculation of the summation is done only for N_{bins} terms, which can be on the order of a few thousand, which is rather less than $n(n-1)$ in the Debye formula. This adds significant computational advantages over the direct method presented in Eq. 1. For this purpose we have used as a reference the program DALAI (Pantos and Bordas, 1994), which implements this approximation.

The inverse x-ray scattering problem consists of deducing the structure of an object from its solution scattering profile. One usual modeling method consists of using Eq. 2 to estimate the scattering profile from a model distribution of beads at a given resolution and comparing it with the experimental data. If the two profiles do not match, then the model is redefined (i.e., a new distribution of beads is generated) and tested again until a good approximation is obtained. Several methods have been developed following this approach. The more successful applications are those that take advantage of crystal structural information or those that use a fixed geometry. A common characteristic of all of them is the need to test all possible conformations at a given resolution. This approach is obviously constrained both in resolution and in applicability. Models with high resolution require a large amount of beads, which in turn result in a huge number of different combinations (each one corresponding to one mass distribution) that are eventually impossible to compute. This problem belongs to the class of Np-complete problems for which it is impossible to evaluate all of the possible solutions, and some optimization method must be used instead of the exhaustive exploration. Taking into account the nature of the problem (with a presumably rugged error landscape), we have applied a genetic algorithm to iterative fitting of SAXS data. As will be shown below, this method allows the search for a satisfactory solution in the huge number of different combinations.

Genetic algorithm implementation of the inverse scattering problem

A GA (Goldberg, 1989) essentially consists of a population of elements (called *chromosomes*, by analogy to the function of cellular chromosomes) and some rules for reproduction and selection according to a given fitness criterium. These population members evolve under selection pressure conditions and replicate following genetics rules. Each chromosome represents one point in the search space and hence a possible solution of the problem.

It must be emphasized that there are many ways in which a GA can be implemented. In this work we have used a basic implementation that is schematically shown in Fig. 1. First of all, a suitable representation for the inverse scattering problem must be devised. We define an appropriate initial object formed by spheres with reasonable dimensions (in accordance with the desired resolution) and codify it into a binary array forming the chromosome. The *i*th chromosome bit describes the presence (binary value 1) or absence (binary value 0) of a bead in the corresponding spatial position of this object. With this, each different chromosome represents one possible mass distribution, that is a potential solution. In terms of the GA, the objective will be to find the structure with the scattering profile closest to the target one.

The initial chromosome population is randomly generated, within some restrictions applied to reduce the initial search space. Starting from this initial population, each chromosome is assigned a fitness value according to how well it solves the problem. To achieve this, each chromosome is decoded into the corresponding spatial coordinates set, and the resulting model is processed by the SAXS simulation procedure to obtain a theoretical scattering profile. These calculated profiles are compared with the experimental data to determine the fitness value. 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, the generation of better models. This process, which mimics natural evolution, is repeated until the system converges, or a good

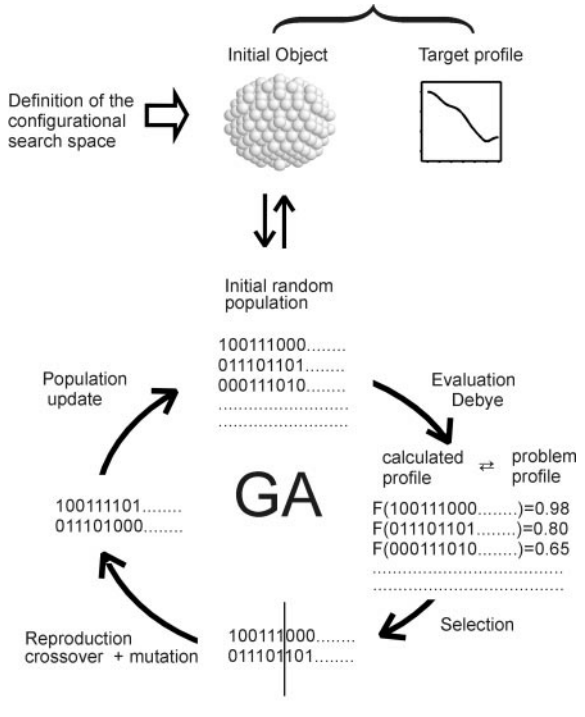

FIGURE 1 Scheme of implementation of the genetic algorithm method of SAXS simulation to numerically solve the inverse scattering problem.

enough solution is found. In the present case, the result of this optimization process yields a population formed by chromosomes (i.e., models) whose corresponding theoretical profiles are very close to experimental ones. In the following sections, we describe in more detail the different elements of the algorithm employed, such as generation of the initial population, evaluation, selection, and reproduction mechanisms.

a) Generating the initial conditions

The genetic algorithm starts with the a priori definition of the configurational search space and the initialization of the population. The definition of the search space is crucial both for overcoming possible ambiguity problems and having a safe and efficient search. We have chosen a search space model with a fixed geometry formed by a uniform hexagonal packing of beads. The number of beads and its radius are selected according to the resolution of the profile and taking into account the dimension of the initial search object. Obviously, this dimension must be large enough to contain the target structure and small enough to constitute a reasonable search space. Two complementary ways of estimating the dimension of the initial model have been developed. On the one hand, it is possible to have a good approximation by a process of trial and error. In fact, when the algorithm is run with a large object with large spheres, its size can be progressively reduced in an iterative process, until a good enough result is obtained, as indicated by higher fitness values. This process is independent of the geometry of the initial model (although most of the first initial models were ellipsoids, other shapes, such as cubic boxes, spheres, etc., may be used). On the other hand, a useful method of estimating the approximate dimensions of the object of study is given by the classical SAXS theory. The maximum particle distance, D_{max}, can be calculated through the pair-distance distribution, because $\rho(r)$ tends to zero for $r > D_{max}$. With this information, the dimension of the configurational search space, consisting of a hexagonal packing of beads within an envelope, can be defined. Experience has shown that the optimum size is the smallest possible giving a volume of the search model ~2–10 times the result model volume. It is also necessary to set the diameter of the beads in this initial pack to the desirable modeling resolution compatible with a reasonable number of beads. Note that for an accurate simulation of the scattering profile, the maximum dimension of the particles employed should be well below $1/(2S_{max})$ (Glatter and Kratky, 1982, pp. 136–137), that is, 8.3 Å for a S_{max} of 0.06 $Å^{-1}$. The resolution of the model is limited by the experimental resolution and the bead size commensurate with that resolution.

After the search scenario has been defined, a population of 400 chromosomes, each of them codifying a given mass distribution of the start model, is randomly generated by filling the starting space with different number of spheres. One restriction imposed on this initial mass distribution is connectivity, that is, the spheres must be relatively near one another. This is easily implemented during the generation of the initial mass distribution only, introducing a proximity weight in the process of selecting random spheres, in such a way that once a bead has been selected, the next one has a 0.95 probability of occupying one of the positions that are at a distance of less than $2R$ in the initial model.

b) Evaluation and fitness criterion

The information contained in each chromosome is decoded to obtain the corresponding mass distribution. The scattering profile of the corresponding beads model is then calculated using the Debye formula. At this point, each population member has a theoretical SAXS profile that can be compared to the problem or reference profile. The evaluation function to be optimized in the GA, the fitness function, is based on this comparison. First of all, to compare different scattering profiles, a normalization procedure is necessary. This is made by dividing the scattering intensity $I(S)$ by the value of the intensity at the origin $I(0)$. In the case of experimental data, the later can be obtained by extrapolation within the Guiner region (Glatter and Kratky, 1982), or substituted by the value at a very small angle. This results in a normalization of every scattering profile to unitary dimensionless intensity at the origin.

We explored different fitness functions to compare the input and the calculated profiles. The algorithm gave better results by comparing intensities on a logarithmic scale than on a linear scale. Among several fitness functions tested, two stand out: the reciprocal of an R-factor (Rf; see footnote to Table 1) and F, the reciprocal of the square root of the sum of square distances between calculated and input profiles:

$$F = \left(\sum_{i} (\log(I_{exp}(S_i)) - \log(I_{model}(S_i)))^2 \right)^{-1/2} \quad (3)$$

Results from both functions were examined in detail, and in general we obtained relatively lower convergence with Rf^{-1}. In this work we will use F as the optimization parameter. Once this fitness value has been calculated for each profile, chromosomes are classified as a function of their proximity to the problem profile. The resulting ranked population is used as a basis for selection, in such a way that those chromosomes situated in the first positions (i.e., with better fitness) of the population have a higher probability of being combined for reproduction.

c) Reproduction

One fraction (P_{elite}) of the total population (P_{tot}), that with higher fitness, is preserved for the next generation. Then the resting members are deleted to make room for the new chromosomes. We have obtained good results by preserving one-half of the population and discarding the other half ($P_{elite} = 0.5P_{tot}$). The regeneration of the population is made by the application of two genetic operators: mutation, which introduces new information in the population; and crossover, which combines the information of the fittest chromosomes. The chromosomes used to create the new generation by application of these operators are selected randomly between the best members (the default value is $0.4P_{tot}$ first elements). The genetic operators are chosen probabilistically according to predefined weights. Typical weight values are 0.4 and 0.6 for the mutation and crossover, respectively. This parent selection avoids premature convergence (trapping into local minima) of the search. The implementation of the two operators is as follows.

Crossover. The crossover operator proceeds to exchange the information of both parent chromosomes to obtain two offspring. For this purpose, we use a uniform crossover (Davis, 1991). This kind of crossover implies that the degree of exchange is fixed to a crossover rate, and the precedence of each bit in the offspring is chosen randomly from the two bits situated in the corresponding position in the parents. In general we have used a crossover rate of 0.2. That is, on average, the offspring has 20% of the bits (genes) from one parent and 80% from the other, and the distribution is random.

Mutation. With this operator a new element is created, copying bit to bit a single parent with a certain error rate, which consists of changing its binary value. As in most GAs, the mutation rate value is small. In this work mutation has been introduced with a probability less than or equal to 0.01.

In addition, to increase the performance of the method, any duplicated chromosome is discarded. With this we maintain the diversity in the population, and by extension reduce premature convergence.

d) Population update

Finally, after mutation, crossover, and rejection of duplicates, the newly generated offspring replaces the fraction ($1 - P_{elite}$) of the population with the worst fitness values.

The above process (steps b–d) is cyclically repeated until the maximum fitness does not improve for a certain number of iterations (convergence). The complete GA is schematically represented in Fig. 1. Most of the parameters in the algorithm are user-defined. For example, the population size (P_{tot}), the corresponding parts of the population that are recombined, the genetic operator weights, and the crossover and mutation rates are the result of previous parameterization. In terms of robustness, the default values give the best results for models starting with a size no larger than 1000 beads. Obviously, for a given model, the values of the parameters could be improved according to its characteristics. For example, if the

TABLE 1 Results of modeling calculated SAXS profiles*

	Search space			Model				
Protein	R (Å)	N	Maximum dimensions (Å)	N	R_g (Å)	r.m.s.	Rf	V_{model}/V_{pdb}
βb2-Crystallin	6	62	$48 \times 34 \times 29$	23	21.29	5.28×10^{-2}	4.87	1.61
2bb2.pdb	6	736	$118 \times 108 \times 108$	23	24.23	1.28×10^{-2}	1.30	1.61
R_g(pdb) = 24.7 Å	6	254	$96 \times 83 \times 88$	23	24.38	6.95×10^{-3}	0.69	1.61
max. dimensions	6	154	$90 \times 55 \times 48$	23	24.38	6.95×10^{-3}	0.69	1.61
$= 76 \times 45 \times 38$	6	102	$84 \times 49 \times 48$	23	24.38	6.95×10^{-3}	0.69	1.61
V_{pdb} = 12,879 Å^3	4	235	$84 \times 48 \times 45^{\#}$	63	24.63	1.27×10^{-3}	0.07	1.31
	2	676	$76 \times 60 \times 55^{\#}$	381	24.72	3.65×10^{-4}	0.03	1.01
γ-Crystallin	6	236	$72 \times 76 \times 68$	22	15.93	1.26×10^{-3}	1.19	1.52
g2c.pdb	6	92	$60 \times 48 \times 48$	22	15.93	1.26×10^{-3}	1.19	1.52
R_g(pdb) = 16.65 Å	4	194	$60 \times 46 \times 41^{\#}$	68	16.53	9.31×10^{-4}	0.07	1.40
max. dimensions	2	637	$56 \times 41 \times 43^{\#}$	423	16.69	1.83×10^{-4}	0.02	1.09
$= 54 \times 40 \times 34$								
V_{pdb} = 13,051 Å^3								
Ribonuclease In.	6	340	$96 \times 83 \times 88$	54	24.81	8.97×10^{-3}	1.20	1.60
lbnh.pdb	6	168	$72 \times 69 \times 68$	55	24.73	7.51×10^{-3}	0.84	1.63
R_g(pdb) = 24.94 Å	4	441	$80 \times 72 \times 60^{\#}$	150	24.85	2.13×10^{-3}	0.18	1.20
max. dimensions								
$= 70 \times 61 \times 60$								
V_{pdb} = 30,475 Å^3								
Lysozyme	6	236	$76 \times 72 \times 68$	16	13.73	1.58×10^{-3}	1.29	1.36
6lyz.pdb	6	92	$60 \times 48 \times 48$	16	13.73	1.58×10^{-3}	1.29	1.36
R_g(pdb) = 14.03 Å	4	155	$48 \times 46 \times 45^{\#}$	46	13.76	8.67×10^{-4}	0.07	1.16
max. dimensions	2	823	$50 \times 48 \times 35^{\#}$	316	14.05	5.13×10^{-4}	0.03	1.00
$= 44 \times 39 \times 34$								
V_{pdb} = 10,655 Å^3								
2xLysozyme	6	340	$96 \times 88 \times 83$	32	20.38	8.61×10^{-3}	0.83	1.36
lrcm.pdb	6	236	$76 \times 72 \times 68$	32	20.08	6.77×10^{-3}	0.54	1.36
R_g(pdb) = 20.6 Å	4	275	$72 \times 58 \times 46^{\#}$	94	20.27	9.90×10^{-4}	0.10	1.18
max. dimensions	2	921	$68 \times 58 \times 45^{\#}$	651	20.58	3.42×10^{-4}	0.03	1.02
$= 62 \times 58 \times 44$								
V_{pdb} = 21,353 Å^3								

*R is the bead radius, N is the number of beads. The radius of gyration (R_g) was calculated as $R_g = (\Sigma_i z_i R_i / \Sigma_i z_i)^{1/2}$, where z_i is the atomic number of atom i and R_i is the atomic distance from the center of the electron charge distribution in the molecule (the hydrogens are not taken into account). In the case of the models of identical beads, the computation is simpler, because z_i is constant. The goodness of fit was measured using the root mean squared deviation, r.m.s. = $(\Sigma_i(\log(I_{exp}(S_i)) - \log(I_{model}(s_i)))^2/N)^{-2}$, and an R factor,

$$\mathrm{Rf} = \sum_i \left\| \frac{\log(I_{exp}(S_i)) - \log(I_{model}(S_i))}{\log(I_{exp}(s_i))} \right\|.$$

#the search space is within an envelope generated from the preceding model with a large bead radius (see Materials and Methods).

problem needs more exploration, it might be necessary to increment the weight of the mutation rate in the search process.

The program was written in FORTRAN and compiled for Silicon Graphics workstations. As pointed out above, the computer time depends critically on the number of spheres involved (resolution) and on the dimension of the initial model. In a standard run with a starting model of 300 beads, the average time employed was ~3 h on a dedicated SG Indigo 2 workstation with R4400 at 150 Mhz and 32 MB of RAM.

e) Increasing resolution

The higher the resolution, the longer the computing time. A procedure was devised to increase the resolution and to prevent and detect possible local minima. This procedure, referred to as mask strategy, consists of progressive reduction of the search space and the size of the bead by creating an envelope or mask from the resulting model of a previous search (see Fig. 2 *B*). In practice, the model formed by beads with radius R_1 is placed within an hexagonal packing of beads of radius $R_2 < R_1$. The smaller beads at a center-to-center distance smaller than $2R_1$ from the larger initial beads constitute the search space for the next run of the genetic algorithm. This yields two advantages: it overcomes problems of convergence and permits working with smaller models, and with the corresponding reduction in the search space, it saves calculation time. This strategy is also useful when the dimension of the initial object is unknown. In such cases, the search can be started with large models that are progressively reduced, with a minimum cost in computing time.

Performance of the genetic algorithm with simple objects

To tune up the parameters of GA and evaluate its performance, several tests were carried out with the scattering profiles calculated for simple beaded

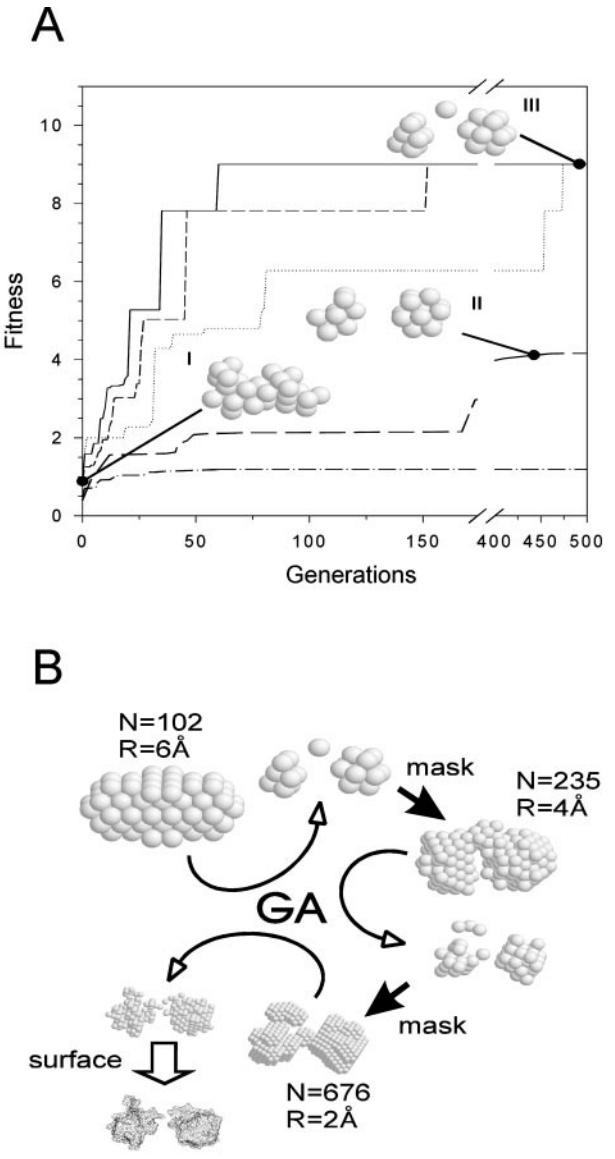

FIGURE 2 (*A*) Representation of the search performance with different dimensions of the initial object, for the βb2-crystallin fragment. The fitness parameter *F* (see Materials and Methods) of the best member of the population is plotted versus the number of generations of the genetic algorithm. Solid line, performance of the method with an initial search model formed by 102 beads. Dashed line, 154 beads; dotted line, 254 beads; long dashed line, 736 beads; dash-dotted line, 64 beads. Note that as the dimension increases, the performance of the search decreases. Note as well how with the 64 bead search space there is no optimization, because the search space does not contain the problem structure. The bead model I corresponds to the best start model in the randomly generated initial population, and model II is the local minimum in which the algorithm falls with an initial object formed by 736 beads. Model III is the same final best-fit model at which the algorithm arrived with initial models made of 102, 154, or 254 beads. (*B*) An example of mask strategy with the same test object. The result of successively increasing resolution by creating a new search space with smaller beads (an envelope to the previous best-fit model) can be observed. *N* is the number of beads of each initial object for each of the three runs of the algorithm, and *R* is the bead radius.

objects (such as hexagons, cubes, and letters) manually built with the program LEGO (Pantos, unpublished), and letting the algorithm find the shape of the object. In these examples the main difficulty of the inverse scattering problem arises: with different runs of the algorithm, it is possible to obtain different model structures. In practice the models converge; they are consistently similar to one another and to the shape object of the test (Chacón et al., 1998). Occasionally the algorithm fell into a local minimum with the simple examples, resulting in a model with a good profile fit, but away from the global optimum that characterizes the structural shape. This convergence/ambiguity problem is directly related to the resolution and the configurational space dimension. In fact, these first results showed how the ambiguity decreased with increasing resolution and decreasing model size. Ideally, to reduce this problem it is necessary to work with the maximum resolution and the smallest possible configurational space. In practice, defining the initial configurational space following the rules given above and making different runs to accumulate statistics is enough to get a successful result. Because of the stochastic character of the GA, 10 different runs are found to be sufficient in each test to safeguard the convergence and confidence in the result. In these preliminary tests the usefulness of the stepwise reduction of resolution and search space (the mask strategy) was corroborated.

X-ray solution scattering profiles

We used the simulation program DALAI (Pantos and Bordas, 1994) to generate the theoretical SAXS profiles from the atomic coordinates of test proteins taken from the Protein Data Bank (Abola et al., 1988). The scattering vector range taken for the calculation of the profile was 0.001–0.06 $Å^{-1}$, by analogy with the experimental range that can be practically achieved with SAXS measurements employing synchrotron radiation. It is worth mentioning that the program CRYSOL (Svergun et al., 1995) could also be used to obtain the theoretical profiles of test models from the atomic coordinates of proteins with known tertiary structures. To better reproduce an experimental scattering spectrum, noise has been added to these profiles. The noise introduced follows a Gaussian distribution (Press et al., 1989) with an amplitude approximately equal to the magnitude of the intensity at the highest angle values.

The experimental x-ray solution scattering data of lysozyme (from hen egg white; Sigma) in 50 mM sodium phosphate buffer (pH 7) were collected at station 2.1 of the Daresbury Laboratory Synchrotron Radiation Source. Data acquisition and processing were performed as previously described (Díaz et al., 1994; Andreu et al., 1992). Absolute values of the scattering vector were calibrated using all of the observable diffraction orders of the 67-nm repeat in wet tail collagen. A 3-m camera was employed, effectively covering an *S* range from 0.003 to 0.035.

RESULTS

To evaluate whether it is possible to objectively retrieve the low-resolution structure of a protein from its x-ray solution scattering profile, the performance of the genetic algorithm method has been tested with synthetic scattering profiles, with an *S* range from 0.001 to 0.06 $Å^{-1}$, calculated for proteins of known atomic structures. In the set of test proteins some representative shapes are included to check the ability of the search procedure to find different models compatible with both SAXS and shape. To better reproduce the experimental scattering spectra, noise has been added to the profiles. In the following sections we will analyze how the genetic algorithm operates (Fig. 2), analyze the significance of results obtained for each model protein structure, and finally present initial results of modeling an experimental scattering profile with this method. For each example, the problem SAXS profile, the best model profiles, and the corresponding bead models are collected in Fig. 3, *A–E*. For comparison, Fig. 4 shows three different orthogonal views

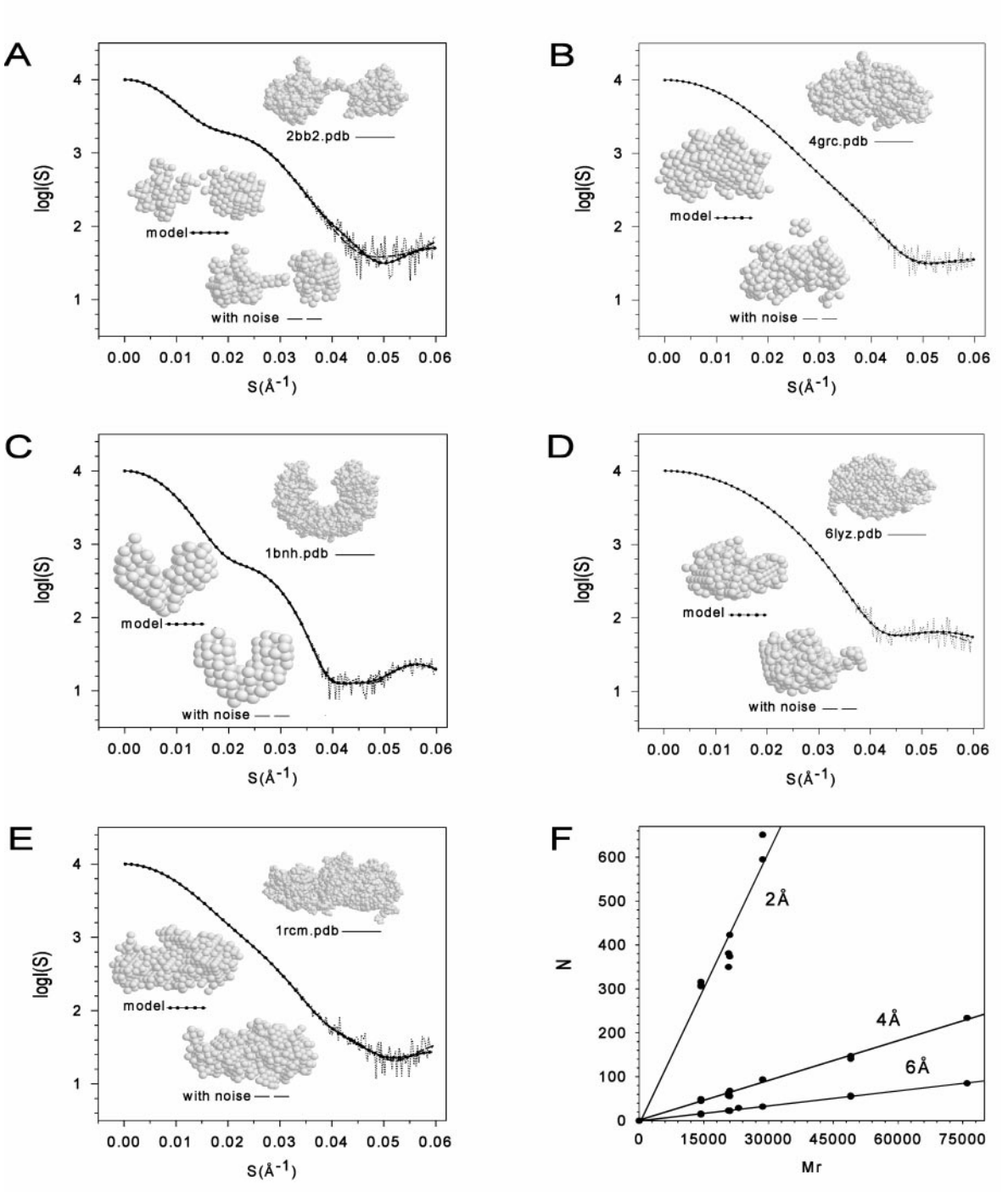

FIGURE 3 Calculated and fitted scattering profiles for the known protein structures of (*A*) βb2-crystallin (2bb2.pdb), (*B*) γ-crystallin (4grc.pdb), (*C*) ribonuclease inhibitor (1bnh.pdb), (*D*) lysozyme (6lyz.pdb), and (*E*) two molecules of lysozyme (1rcm.pdb). Solid lines: synthetic profiles calculated from each pdb file with the program DALAI (Pantos et al., 1996; hydrogen atoms and water molecules have not been taken into account). Points (*circles*), DALAI_GA simulated profiles. Dotted lines, synthetic profiles with added noise. Dashed lines, corresponding simulated profiles. A CPK view of the problem structure is shown in the upper right part of each panel, and the best fitted bead models are shown below it. (*F*) The numbers of beads N in models with different bead radii are plotted versus the anhydrous molecular masses of the problem structures. The corresponding linear regression parameters ($N = a + bM_r$) are $a = 0.497$, $b = 1.13 \times 10^{-3}$, $r = 0.998$; $a = 1.151$, $b = 3.0 \times 10^{-3}$, $r = 0.996$; $a = -8.70$, $b = 20.66 \times 10^{-3}$, $r = 0.974$ for bead radii of 6, 4, and 2 Å, respectively. Data include those from the models derived from scattering profiles with and without noise (*A–F*) and similar experiments with lactoferrin (1lfd.pdb, M_r 76,000).

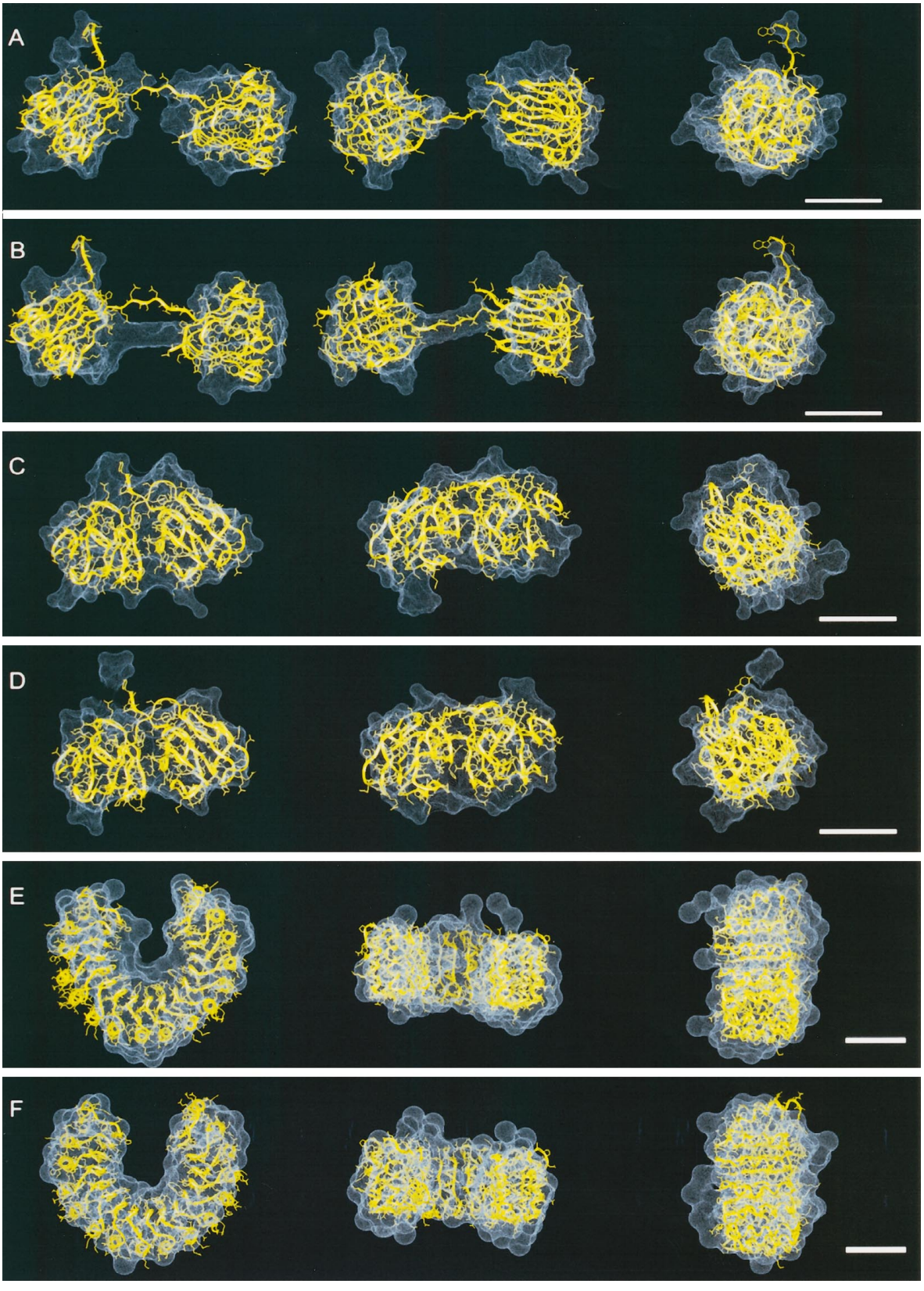

FIGURE 4 Bead models obtained from synthetic SAXS solution profiles in comparison with the known protein x-ray crystal structures. To get a reasonably good comparative display, the PDB structures are in a ribbon peptide chain and wireframe side-chain representation (*yellow*), and the models are displayed as a dotted Connolly surface (*blue*; probe radius 4.00 Å) generated from the sphere model with Insight II (version 95.0). Note that the comparison of each pair of high-low-resolution structures is not a simple task, and has to be made using graphic tools for molecular representation. In this work we have manually overlaid the structures; the automation of this procedure is an open problem. The proteins modeled are (*A*) the βb2-crystallin fragment and (*B*) the same with noise added to the SAXS profile; (*C*) γ-crystallin and (*D*) the same with noise; (*E*) pancreatic ribonuclease inhibitor and (*F*) the same with noise. In each panel the left view has been rotated 90° in the *x* and *y* axes to generate the center and right views, respectively. The bars indicate 20 Å.

of a 3D overlay of the corresponding problem structure (*in yellow ribbon and side chain representations*) and the fitted structure (*as a dotted blue surface of the bead model*). The results are summarized in Table 1 (without noise) and Table 2 (with noise). Note that the r.m.s. deviation and R factor are not directly comparable between the two tables, because the noise in the problem profile obviously increases their values. As a reference, the corresponding r.m.s. and Rf values between the original problem profile and the same with noise, for each test protein, are indicated in the first column of Table 2, providing an approximate indication of the lowest limit range that each fit could achieve. To be confident about the convergence of the results, and because of the stochastic character of the method, 10 different initial simulations with spheres of radius 6 Å were repeated for each test protein. All of these bead models had converging sizes and shapes; actually the majority of the runs at this resolution resulted in practically the same best-fit solution. The resolution was increased by employing this initial solution to constrain the starting configurational space for additional search cycles with decreasing bead radius.

Performance of the genetic algorithm method of x-ray scattering simulation

βb2-Crystallin

The SAXS pattern was calculated from the atomic coordinates of the 2bb2 PDB entry (Bax et al., 1990). The structure of this βb2-crystallin fragment shows two linked but separated globular domains, one of them with a protuberance (approximately residues 168–175) outside the globular shape. Although this structure is not properly representative of a real SAXS experiment, because the link between domains may be flexible in solution, it is a good illustrative example for showing the performance of the method proposed here. It will be used to illustrate how to define the initial conditions, and how the search space influences the behavior of the genetic algorithm.

A set of ellipsoids of different sizes, all of them formed by a hexagonal packing of 6-Å-radius beads, was employed as initial searching spaces. The results of a set of representative runs are shown in Fig. 2 *A*. The best fit evolves over successive generations in such a way that the fitness value

TABLE 2 Results of modeling calculated SAXS profiles with added noise*

	Search space			Model				
Protein	R (Å)	N	Maximum dimensions (Å)	N	R_g (Å)	r.m.s.	Rf	V_{model}/V_{pdb}
βb2-Crystallin	6	154	$90 \times 55 \times 48$	23	24.68	9.07×10^{-2}	7.45	1.62
R_g(pdb) = 24.7 Å	6	102	$84 \times 49 \times 48$	23	24.89	9.06×10^{-2}	7.43	1.62
max. dimensions	4	241	$84 \times 46 \times 45^{\#}$	57	25.07	8.97×10^{-2}	7.24	1.19
$= 76 \times 45 \times 38$	2	623	$76 \times 44 \times 39^{\#}$	350	24.99	8.96×10^{-2}	7.16	0.91
r.m.s. = $9.57\ 10^{-2}$								
Rf = 7.50								
γ-Crystallin	6	236	$72 \times 76 \times 68$	23	16.73	5.35×10^{-2}	5.35	1.59
R_g(pdb) = 16.65 Å	6	92	$60 \times 48 \times 48$	23	16.53	5.05×10^{-2}	5.05	1.59
max. dimensions	4	198	$60 \times 46 \times 42^{\#}$	56	16.64	4.71×10^{-2}	4.71	1.15
$= 54 \times 40 \times 34$	2	566	$52 \times 49 \times 46^{\#}$	374	16.79	4.69×10^{-2}	4.69	0.96
r.m.s. = $5.82\ 10^{-2}$								
Rf= 4.63								
Ribonuclease Inc.	6	340	$96 \times 83 \times 88$	55	26.42	7.71×10^{-2}	7.71	1.63
R_g(pdb) = 24.94 Å	6	168	$72 \times 69 \times 68$	57	25.30	7.24×10^{-2}	7.24	1.69
max. dimensions	4	525	$72 \times 71 \times 58^{\#}$	147	24.79	6.30×10^{-2}	6.30	1.29
$= 70 \times 61 \times 60$								
r.m.s. = 5.6410^{-2}								
Rf = 6.90								
Lysozyme	6	236	$76 \times 72 \times 68$	15	13.00	4.06×10^{-2}	3.19	1.27
R_g(pdb) = 14.04 Å	6	92	$60 \times 48 \times 48$	15	13.00	4.06×10^{-2}	3.19	1.27
max. dimensions	4	141	$48 \times 42 \times 40^{\#}$	49	13.98	3.67×10^{-2}	2.56	1.23
$=44 \times 39 \times 34$	2	480	$52 \times 43 \times 41^{\#}$	309	14.08	3.66×10^{-2}	2.56	0.97
r.m.s. = $3.79\ 10^{-2}$								
Rf= 2.56								
2xLysozyme	6	340	$96 \times 83 \times 88$	32	20.53	5.35×10^{-2}	4.03	1.36
R_g(pdb) = 20.6 Å	6	236	$76 \times 72 \times 68$	32	20.38	5.05×10^{-2}	4.12	1.36
max. dimensions	4	281	$78 \times 52 \times 44^{\#}$	93	20.59	4.71×10^{-2}	3.89	1.17
$=62 \times 58 \times 44$	2	915	$69 \times 56 \times 48^{\#}$	595	20.69	3.86×10^{-2}	3.86	0.93
r.m.s. = $4.71\ 10^{-2}$								
Rf = 3.75								

*Symbols and definitions are as in Table 1.

#The search space is within an envelope generated from the preceding model with a larger bead radius (see Materials and Methods).

improves until the system converges (the same occurs with the average fitness of the population; not shown). Note that to have a fully unbiased search, the program assigns no penalty for unconnected structures. The results show how the dimension of the starting model influences the search process, and how it is possible to estimate an appropriate dimension by a simple process of trial and error. For instance, starting with model dimensions progressively larger than the real structure (the models made of 102, 154, and 254 beads) decreases the performance of the search. As the conformational space size increases the search becomes more difficult, and thus the time needed to get the maximum fitness values increases. This behavior is clearly illustrated in the evolution trajectories, in which the convergence to maximum fitness is slower for the larger start models. In an extreme case, if the conformational space becomes too large, the system gets to a local minimum, as exemplified with the model formed by 736 spheres. On the other hand, if the dimension of the search space is too small, the fitness value remains continuously low (model of 62 spheres in Fig. 2). This is because the solution structure does not fit into the configurational space searched (the optimization has no sense). It is possible therefore to guess the approximate size of the input model by using successive start models and by monitoring the evolution of the fit value. A similar conclusion can be derived by examining the similarity of shapes between model and real structures. In all of the cases (except for the insufficient search space), the model reproduces the two domain shape of the βb2-crystallin fragment. The models of 102, 154, and 254 spheres yield the same best fit shape (model III in Fig. 2 *B* and Table 1). Even in the case of 736 beads, the fitting results in a model with characteristics that resemble the two-domain input shape (Fig. 2, model II). Similar results have been obtained in the case of profiles with added noise (not shown).

Following the initial search, it is possible to improve the resolution by progressively decreasing the radius of the model beads from 6 Å to 4 Å and 2 Å and employing a mask strategy (Fig. 2 *B*). The best structure resulting from the previous fit is used as a mask to define the next start model with a smaller bead radius. The benefits of this strategy can be noticed in the progressive resemblance of the models to the scattering profile of the problem molecule, as can be seen in Table 1. The R_g of the model gradually converges to that calculated for the known structure, 24.70 Å (from 24.38 at 6 Å to 24.72 at 2 Å). The value of the r.m.s. deviation between the input and the calculated scattering patterns falls from 6.9×10^{-3} to 3.6×10^{-4}. The volume of the model also converges to the one of the problem structure, as can be seen by the ratio V_{model}/V_{pdb}, which changes from 1.61 to 1.01. It can be appreciated in Fig. 3 *A* how the SAXS scattering pattern of the calculated model is indistinguishable from that of the input structure. The structure of the final beads model is also shown; the similarity between them is clear from a visual comparison of the corresponding bead models. A more precise comparison is presented in Fig. 4 *A*, where the GA obtained solution (*blue surface*) is superimposed on the crystal structure (*yellow ribbon*). As can be observed, not only are the two domains correctly modeled, but the protrusion in one of the domains is also predicted and the link between the domains emerges. The overall resemblance is considered to be very good. Even though the crystal structure has a resolution of 2.1 Å and the simulated solution scattering data extend only to 16.6 Å, the later contains enough information to closely define the volume and overall structure of this protein.

The results of fitting the synthetic SAXS profile with added noise are summarized in Table 2. The noise perturbation introduced in the synthetic SAXS spectrum results in an r.m.s. deviation near 10^{-1}, and an Rf of 7.5, obviously restricting any possible improvement of the numerical fit below this limit (order of magnitude). The decrease in the size of the beads does not exert much influence on the fit values, reflecting this limiting effect. Nevertheless, the volume of the model, compared with the real one, presents a noticeable improvement (the ratio V_{model}/V_{pdb} varies from 1.62 to 0.91 with decreasing bead size). The final fit (Fig. 4 *B*) is a less accurate solution than that obtained without noise (Fig. 4 *A*). However, the resulting model is still fully compatible with the crystal structure of the βb2-crystallin fragment shape, showing an acceptable overall resemblance. In particular, it reproduces the two-domain structure, the projection in one of the domains still can be found, and, interestingly, it locates the connecting arm, even if it is slightly misplaced.

γ-Crystallin

The next test structure used was γ-crystallin, also a protein from the eye lens. The atomic coordinates were taken from the 4gcr entry of PDB (Wistow et al., 1983). The shape of the protein can be described as a "kidney shape" formed by two adjacent lobes (Figs. 3 *B*, 4 *C*, and 4 *D*). As in the case of βb2-crystallin, a small protrusion can be distinguished between the two lobes. The calculated scattering profiles are displayed together with the corresponding modeling results in Fig. 3 *B*. As in all of the cases examined, the profiles modeled without noise are practically indistinguishable from the reference profiles. The results of different fits of increasing resolution are presented in Tables 1 and 2. The overlay of the fitted models on the crystal structure reveals a good similarity (Fig. 4 *C*), even in the case of the SAXS profile with noise (Fig. 4 *D*). The "kidney shape" is reasonably reproduced in both cases. Notice that in the case with noise, the projection is represented by a small set of beads separated from the rest of the model structure.

Ribonuclease inhibitor

Porcine pancreatic ribonuclease inhibitor is a cytoplasmic protein with a "horseshoe-like" shape. The crystal structure at 2.5-Å resolution was taken from the 1bnh PDB entry (Kobe and Deisenhofer, 1993). This is the larger of the

protein structures used to test the method (Table 1). The problem SAXS profile can also be well fitted, and the correct size and shape are deduced (Fig. 3 *C*). In the overlay of the bead model and the crystal structure, it can be observed that there is a small lack of mass in one of the arms (see Fig. 4 *E*). Nevertheless, in all 10 runs, the models converged to the characteristic horseshoe shape with little difference from those displayed in Fig. 4 *E*. The same behavior occurs when the profiles with added noise are used. As can be seen in Fig. 4 *F*, the model horseshoe has a small torsion, yet curiously the contours of the arms are better defined than in Fig. 4 *E*. Because of the large size of the problem structure, this model was only refined up to a bead radius of 4 Å, because the huge number of beads required beyond this resolution exceeds the capabilities of the current implementation of the algorithm.

Monomer and dimer structures. Size of models

Among the structures of lysozyme available in the database, we have used two different crystal structures: 1) native chicken egg white lysozyme (Diamond, 1974; 6lyz PDB entry) and 2) two molecules of Cys-6-,Cys-127-carboxymethylated lysozyme in one unit cell (Hill et al., 1993; 1rcm PDB entry). The shape of the two-domain structure can be roughly described as globular, containing a small hole that corresponds to the active site (Figs. 3 *D* and 6). The improvement produced by the successive radius reduction can be regarded in Table 1. In the first step of refinement, in which the bead radius decreases from 6 Å to 4 Å, the fitting presents a significant improvement. In the second refinement step, using the 2-Å radius bead model, the fitted SAXS is indistinguishable from the reference profile (Fig. 3 *D*), with a R_g difference of only 0.02 Å. The volumes of the bead models deduced from the scattering profiles with and without noise are, respectively, 3% and $<1\%$ smaller than the van der Waals volumes calculated from atomic coordinates (Tables 1 and 2).

Whereas the globular shape of the structure is correctly deduced independently of the bead size, the characteristic hole cannot be noticed with models made of 6-Å-radius beads, and it shows up in the successive refinements of the initial model using the mask strategy. In Fig. 6, *A* and *B*, the 2-Å bead radius models of lysozyme are superimposed on the crystallographic structure. The differences with the reference structure are somewhat more pronounced when the input has noise (Fig. 6 *B*, note one single bead off the 309 bead model structure). In any event, the cleft is correctly located in both cases.

To test the correspondence of monomeric and dimeric models, two molecules of lysozyme from a unit cell were analyzed as a dimer. After refinement, the final models resemble the real structure quite well (Fig. 3 *E*). As can been seen in Tables 1 and 2, the resulting fit values for the lysozyme crystal dimer are comparable to those obtained for a single molecule, and the models have double mass (double number of beads) within a few percent error, irrespective of the addition of noise to the problem scattering profiles. In fact, the number of beads in the models is proportional to the molecular mass of the problem structure, for all of the structures tested (Fig. 3 *F*).

Modeling an experimental scattering profile: lysozyme

To make a first test of the applicability and effectiveness of the procedure with real experimental SAXS data, the scattering profile of lysozyme in the *S* range from 0.003 to 0.03 $Å^{-1}$ has been employed. This has been compared with synthetic data in the same *S* range. The experimental SAXS pattern is plotted together with the fitted model profile in Fig. 5. The differences between them are quite small, and the two profiles maintain identical shapes. The results of both experiments, with real and synthetic data, are summarized in Table 3. As in previous cases, 10 runs of the genetic algorithm proved sufficient to ensure convergence. It can be observed that the volume ratio in the theoretical case is close to unity, whereas it is slightly larger in the experimental case (the volume of the model derived from experimental data is somewhat larger than the volume of the anhydrous crystal structure). The predicted radii of gyration are close to the respective reference values.

The low-resolution bead models of lysozyme obtained from its synthetic and experimental SAXS profiles are compared with the crystal structure in Fig. 6, *C* and *D*, respectively (in the later case six different views have been rep-

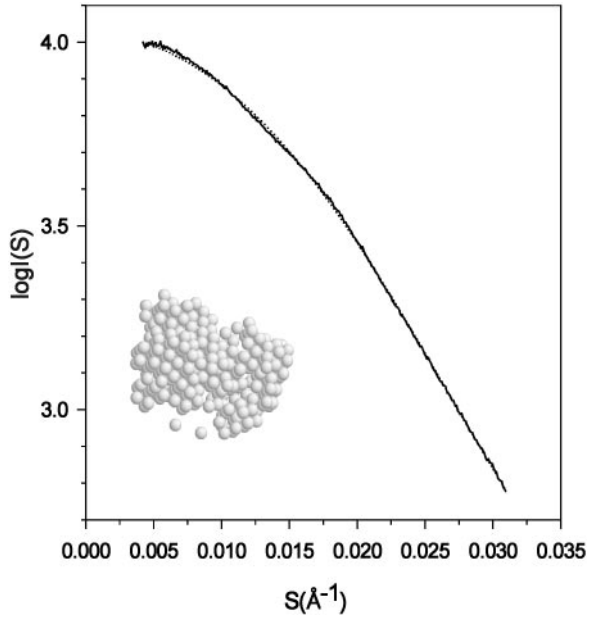

FIGURE 5 Experimental (*solid line*) and simulated SAXS profile (*dots*) of lysozyme. The best fit bead model (2-Å bead radius) is shown. The molecular mass values estimated for lysozyme from models with 6, 4, and 2 Å bead radius, employing the linear regression shown in Fig. 3 *F*, are 15,400, 18,200, and 17,500, respectively.

TABLE 3 Results of modeling experimental and calculated SAXS profiles of lysozyme*

	Search Space			Model				
Lysozyme	R (Å)	N	Maximum dimensions (Å)	N	R_g	r.m.s.	Rf	V_{model}/V_{pdb}
6lyz.pdb	6	92	$60 \times 48 \times 48$	14	13.51	6.37×10^{-5}	0.01	1.19
R_g(pdb) = 14.03 Å	4	147	$58 \times 44 \times 42^{\#}$	45	13.81	5.53×10^{-5}	0.003	1.13
V_{pdb} = 10655	2	810	$44 \times 42 \times 41^{\#}$	306	14.04	5.20×10^{-5}	0.003	0.96
S = 0.0001–0.03								
SAXS exp.	6	236	$76 \times 72 \times 68$	17	15.34	6.69×10^{-3}	0.48	1.44
R_g/exp) = 15.85 Å[§]	6	92	$60 \times 48 \times 48$	17	15.44	6.69×10^{-3}	0.48	1.44
S = 0.0024–0.031	4	174	$48 \times 46 \times 45^{\#}$	56	15.54	6.69×10^{-3}	0.46	1.40
	2	91	$50 \times 48 \times 35^{\#}$	352	15.74	6.67×10^{-3}	0.46	1.11

*Symbols and definitions are as in Table 1.
[#]The search space is within an envelope generated from the preceding model with a larger bead radius (see Materials and Methods).
[§]R_g value determined from the scattering pattern employing the Guinier approximation.

resented). In both cases the resolution employed is not enough to identify the cleft previously characterized with an S range twice as large (Fig. 6, *A* and *B*), but is clearly good enough to identify the globular structure and correct size of lysozyme.

DISCUSSION

A general method for retrieving the low-resolution structure of macromolecules in solution from their SAXS profiles has been presented. The method consists of fitting a theoretical scattering profile to the target experimental profile. The theoretical profile is obtained by Debye calculation of beaded models. The models are optimized by means of a genetic algorithm that searches the huge space of possible mass distributions. With this method, it is possible to determine the best distribution of mass units that generate a profile that fits the target. The spatial distribution of the spherical beads simulates the structure of the protein at the resolution given by the size of the spheres, as well as its radius of gyration, volume, and mass. This has been tested with scattering profiles calculated from known crystallographic protein structures with and without added noise, and with the experimental profile of lysozyme. The reliability of the method has been verified using protein structures with different sizes and shapes: globular, dimeric, bilobed, horseshoe shaped, and one consisting of two domains connected by a stalk.

Performance of the genetic algorithm method used to numerically explore the inverse x-ray scattering problem and retrieve the low-resolution structure

The algorithm simulates very accurately theoretical x-ray solution scattering profiles in the absence of noise. When noise is included in the problem profiles, the fitting goes to the limit allowed by the deviation induced by the noise (reflected in the r.m.s values presented in Table 2). Because of its stochastic nature, it might be expected in principle that, for a given structure, different model structures would appear in different runs of the algorithm. In practice the method is robust enough to yield, after 10 runs, 10 different structures that can only be distinguished by the disposition of a few beads. This set of solutions characterizes a concrete structural shape and dimensions of the particle. In other words, the method yields good reproducibility. Another potential problem is the intrinsic ambiguity of the inverse scattering problem, that is, theoretically there is not a unique model structure compatible with a given SAXS pattern at a given resolution. The ambiguity is related to the resolution and the model dimensions: the larger the resolution, the smaller the ambiguity. The convergent results presented indicate that this problem is overcome in practice by the capabilities of the GA search combined with the reduction of the search space implemented in the method.

In general terms, the application of the method to the protein structures tested leads to fast convergence to a best fitted model with a shape similar to that of the crystallographic structure. This is based on the dramatic reduction of processing time required to reproducibly explore the whole configurational space with the genetic algorithm, instead of calculating every possible mass distribution at a given resolution. This optimization method opens a set of new possibilities for extracting information on the structure of proteins in solution.

To increase the resolution of the models and to speed up the convergence and reduce the calculation time, a parallel computing version of the algorithm could be implemented, because both Debye calculation and GA are very suitable for parallel implementation. In addition, it may be possible to improve the structure of the GA by introducing new genetic operators.

Comparison of models with problem structures

The simulations were considered to describe size and shape successfully, because only small differences could be identified by overlaying the test and simulated structures (Figs. 4 and 6). Note that the goodness of the fit has been quan-

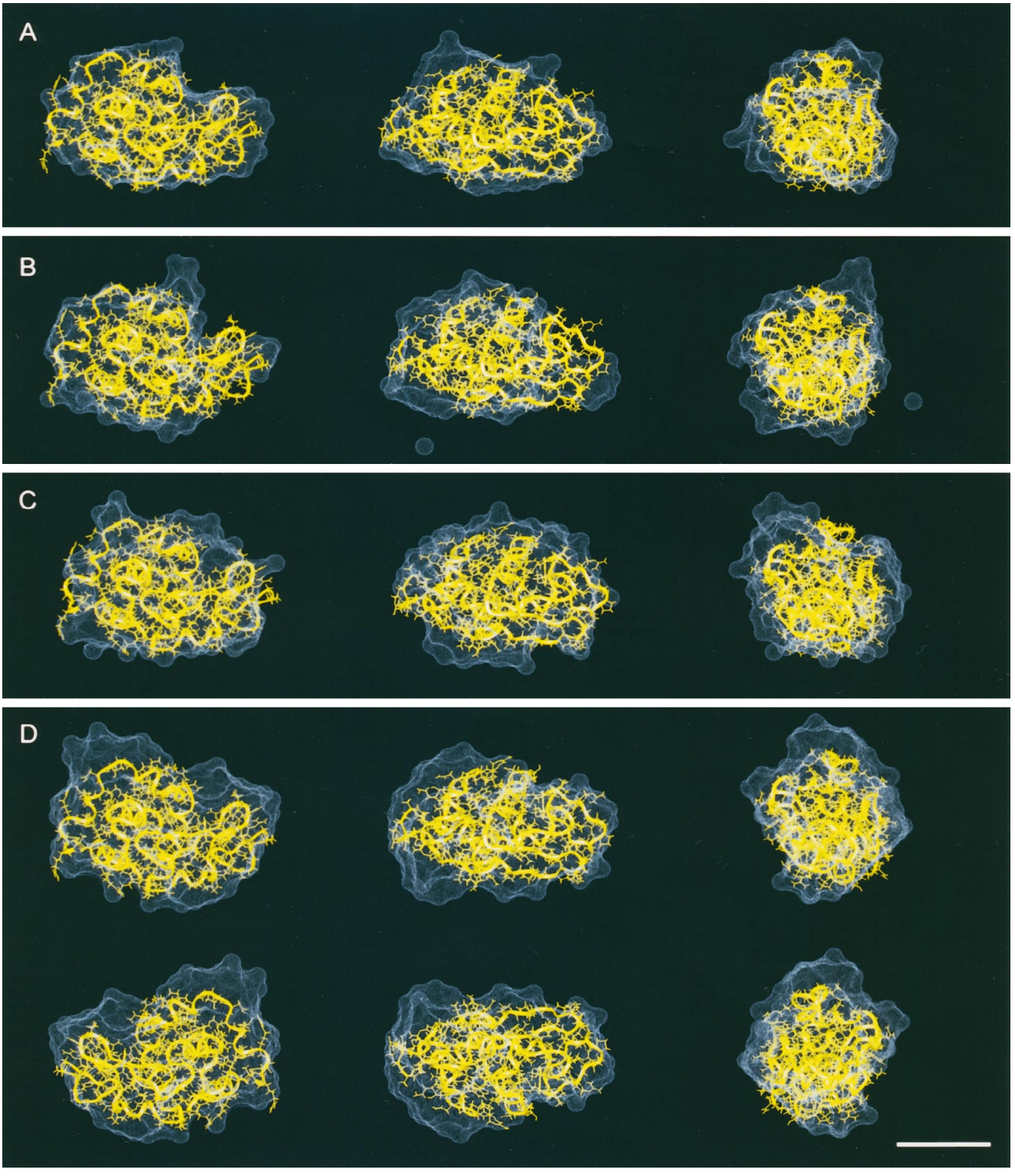

FIGURE 6 Bead models (*blue*) obtained from SAXS solution data of lysozyme in comparison with the x-ray crystal structure (*yellow*). The structures are represented as in Fig. 4. (*A*) Model derived from the calculated SAXS profile to 0.06 A^{-1}; (*B*) same with noise added to profile; (*C*) model obtained from the calculated profile limited to 0.03 A^{-1}. *D* shows the model derived from the experimental SAXS profile of lysozyme to 0.03 A^{-1} (Fig. 5), in six different projections. The bar indicates 20 Å.

titatively evaluated only in the reciprocal space (Tables 1 and 2). The average relative values of volume employing beads of 2-Å radius give maximum deviations of 9% and 6%, with and without noise, respectively (Table 4). The estimation R_g values deviate by less than 1% from the reference values. A good linear correlation has been found

TABLE 4 Summary of results of modeling calculated SAXS profiles

Bead radius (Å)	$R_{g_{model}}/R_{g_{pdb}}$		V_{model}/V_{pdb}	
	Without noise	With noise	Without noise	With noise
6	0.98 ± 0.01	0.99 ± 0.05	1.49 ± 0.13	1.49 ± 0.16
4	0.99 ± 0.00	0.99 ± 0.02	1.26 ± 0.10	1.21 ± 0.06
2	1.00 ± 0.00	0.99 ± 0.01	0.98 ± 0.04	0.94 ± 0.03

between the number of beads of the models and the molecular weight of the problem crystal structures (Fig. 4 *F*).

A closer view of the model versus problem structures reveals that in some regions there is a lack of mass, whereas in others there is an excess of mass. We have observed an apparent mass balance, that is, the resulting structure practically maintains the mass (or volume), even if slightly misplaced. This effect of mass conservation may be interpreted as an inherent property of the algorithm, which in this way finds the correct total mass of the problem protein. This effect can be appreciated in the case of βb2-crystallin (Fig. 4, *A* and *B*), in which the model distribution of mass in the link between domains does not exactly fit the crystallographic structure. This is more noticeable in the model obtained from the scattering profile with noise (Fig. 4 *B*), in which the link is displaced from its original position. Here the effect of the fluctuations induced in the scattering profile is translated into the structure, and the algorithm is able to find a representative structure of the link, which resembles an average of different possible positions of the chain between the two domains. In this same structure the projection of the left domain is split into two adjacent projections in the model. In another example (Fig. 4 *E*), the left arm of the modeled protein has a definite mass deficiency that is compensated for by some extra beads in other parts of the molecule, as can be observed in this and the other two views of the molecule.

The effects of noise are clearly exemplified by the case of lysozyme. It can be observed that in the absence of noise, the method is able to locate the cavity of the active center of the enzyme (Fig. 6 *A*, *left view*). The input file with noise results in a model in which the overall globular shape is correctly predicted, but because of a small dispersion of beads, the cavity is less accurately detected. This effect is also present in the models obtained at lower resolution and from experimental SAXS profile (Fig. 6, *C* and *D*). Nevertheless, the differences, like dispersion of mass in the hole region and the presence of an isolated sphere, are small enough for the models to have an acceptable overall similarity.

Low-resolution protein structure deduced from the x-ray solution scattering profile

The x-ray scattering simulation algorithm provides, in principle, a new procedure for deducing the size and shape of any protein or macromolecular assembly from its solution scattering profile. To our knowledge, this is possible for the first time for any kind of shape, and is applicable to any size, spanning the range from that of smaller proteins to an upper limit dictated only by the lower angle cutoff of the scattering data. The method can operate solely from a scattering profile without any other knowledge about the protein size. The output is a scattering model of the protein, typically consisting of several hundred beads. Maximum information content is extracted from the scattering curve, in contrast to classical parametrical approaches such as radii of gyration and other dimensions of the particle. The procedure is applicable to homogeneous proteins, although currently it is not convenient for the calculation of conformationally flexible ensembles.

The results on the shape of several proteins from their calculated scattering profiles (Figs. 4) offer a very good prospect for doing the same from experimental x-ray solution scattering measurements of proteins to ~ 0.06 Å^{-1}. Nevertheless, the approach will have to be tested with a representative set of experimental scattering profiles of proteins with known crystal structures to prove its practical effectiveness. The effects of possible experimental errors, particularly at the higher angles, will have to be analyzed.

The first results obtained from the experimental scattering profile of lysozyme show that even at limited resolution, the globular shape is correctly deduced, and the size of the protein is very well defined in terms of the volume of the scattering bead model, which includes that of the crystallographic structure (Fig. 6). Note that the size is modeled directly from the normalized scattering profile; that is, it does not require any instrumental calibration of the scattering intensity at the zero angle. The x-ray solution scattering volume of lysozyme, measured from its 2-Å-radius bead model, is 1.10 times that of the anhydrous crystal structure. The effective molecular weight of lysozyme estimated from the scattering model, employing the linear correlation for anhydrous models (Fig. 4 *F*; 2-, 4-, and 6-Å bead radius), is 17,000 ± 1500, which is larger than the chemical molecular weight of lysozyme (14,300). These differences, to be systematically investigated, may correspond to a hydration of $0.19 \pm 0.10\ g_{H_2O}/g_{protein}$, which is contained within the estimated hydration of lysozyme, $0.33\ g_{H_2O}/g_{protein}$ (Kuntz, 1971), or may also contain some measurement bias of the modeling procedure.

Application to large protein assemblies is constrained only by the CPU and memory available to handle the number of spheres of a size commensurate with the resolution of the data. In the favorable cases of repetitive structures, related algorithms that search for the structure of the subunits within a determined lattice may be implemented. It is to be anticipated that the present method may be applied to the accurate analysis of interdomain movements induced by ligands and by protein-protein interactions in solution, as well as by the different environments in crystal and in solution.

We are indebted to Prof. Joan Bordas for suggesting to us the feasibility of deducing low-resolution structures with extensive Debye simulation of the SAXS profiles, and to Prof. Francisco Montero for indicating the use of a genetic algorithm to optimize the problem and for useful discussions. We thank Dr. Antonio Romero for reading the manuscript.

This work was supported by grants DGES PB950116 (JMA) and CICyT BIO960895 (FM). PC had a predoctoral fellowship from the Ministerio de Educación (Spain). JFD was a research associate supported by the Fund for Scientific Research (Flanders) (32.0163.94) and by the Research Council of the Katholieke Universiteit Leuven (OT/93/20). We thank Prof. Yves Engelborghs for making computer facilities available at KUL.

REFERENCES

Abola, E. E., F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng. 1987. Protein data bank. *In* Crystallographic Databases-Information Content, Software Systems, Scientific Applications (F. H. Allen, G. Bergerhoff, and R. Sievers, eds.). Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester. 107–132.

Andreu, J. M., J. Bordas, J. F. Díaz, J. García de Ancos, R. Gil, F. J. Medrano, E. Nogales, E. Pantos, and E. Towns-Andrews. 1992. Low resolution structure of microtubules in solution. *J. Mol. Biol.* 226: 169–184.

Bax, B., R. Lapatto, V. Nalini, H. Driessen, P. F. Lindley, D. Mahadevan, T. L. Blundell, and C. Slingsby. 1990. X-ray analysis of beta B2-crystallin and evolution of oligomeric lens proteins. *Nature.* 347: 776–780.

Beavil, A. J., R. J. Young, B. J. Sutton, and S. J. Perkins. 1995. Bent domain structure of recombinant human IgE-Fc in solution by x-ray and neutron scattering in conjunction with an automated curve fitting procedure. *Biochemistry.* 34:14449–14461.

Boehm, M. K., M. O. Mayans, J. D. Thornton, R. H. J. Begent, P. A. Keep, and S. J. Perkins. 1996. Extended glycoprotein structure of the seven domains in human carcinoembryonic antigen by x-ray and neutron solution scattering and an automated curve fitting procedure: implications for cellular adhesion. *J. Mol. Biol.* 259:718–736.

Cantor, C. R., and P. R. Schimmel. 1980. Biophysical Chemistry. Part II. Techniques for the Study of Biological Structure and Function. W. H. Freeman, New York. 811–819.

Chacón, P., F. Morán, J. F. Díaz, E. Pantos, and J. M. Andreu. 1998. A genetic algorithm for low resolution protein structure determination. *In* Proceedings of the 5th International Symposium on Protein Structure-Function Relationship. Z. Zaidi, editor. University of Karachi, Pakistan (in press). http://www.dl.ac.uk/SRS/FCSI/DALAI_GA/index.html.

Curmi, P. M. G., D. B. Stone, D. K. Schneider, J. A. Spudich, and R. A. Mendelson. 1988. Comparison of the structure of myosin subfragment 1 bound to actin and free in solution: a neutron scattering study using actin made "invisible" by deuteration. *J. Mol. Biol.* 203:781–798.

Dandekar, T., and P. Argos. 1992. Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Eng.* 5:637–645.

Dandekar, T., and P. Argos. 1994. Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* 236:844–861.

Dandekar, T., and P. Argos. 1996. Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* 256:645–680.

Davis, L. 1991. Handbook of Genetics Algorithms. Van Nostrand Reinhold, New York.

Dean, C. E., R. C. Denny, P. C. Stephenson, G. J. Milne, and E. Pantos. 1994. Computing with parallel vitual machines. *J. Phys. III. Colloque.* C9:445–448.

Debye, P. 1915. Zerstreuung von röntgenstrahlen. *Ann. Phys.* 46:809–823.

Díaz, J. F., E. Pantos, J. Bordas, and J. M. Andreu. 1994. Solution structure of GDP-tubulin double rings to 3 nm resolution and comparison with microtubules. *J. Mol. Biol.* 238:214–225.

Diamond, R. 1974. Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.* 82:371–391.

Evans, R. W., J. B. Crawley, R. C. Garratt, J. G. Grossmann, M. Neu, A. Aitken, K. J. Patel, A. Meilak, C. Wong, J. Singh, A. Bomford, and S. S. Hasnain. 1994. Characterisation and structural analysis of a functional human serum transferrin variant and implications for receptor recognition. *Biochemistry.* 33:12512–12520.

Forrest, S. 1993. Genetic algorithms: principles of natural selection applied to computation. *Science.* 261:872–878.

Fujiwara, S., F. J. Kull, E. P. Sablin, D. B. Stone, and R. A. Mendelson. 1995. The shapes of the motor domains of two oppositely directed microtubule motors, ncd and kinesin: a neutron scattering study. *Biophys. J.* 69:1563–1568.

Garrigos, M., S. Mallam, P. Vachette, and J. Bordas. 1992. Structure of the myosin head in solution and effect of light chain 2 removal. *Biophys. J.* 63:1462–1470.

Glatter, O., and O. Kratky. 1982. Small Angle X-Ray Scattering. Academic Press, London.

Goldberg, D. E. 1989. Genetics Algorithms in Search, Optimisation and Machine Learning. Addison-Wesley, San Mateo, CA.

Grossmann, J. G., Z. H. L. Abraham, E. T. Adman, M. Neu, R. R. Eady, B. E. Smith, and S. S. Hasnain. 1993. X-ray scattering using synchrotron radiation shows nitrite reductase from *Achromobacter xylosoxidans* to be a trimer in solution. *Biochemistry.* 32:7360–7366.

Grossmann, J. G., S. S. Hasnain, F. K. Yousafzai, B. E. Smith, and R. R. Eady. 1997. The first glimpse of a complex of nitrogenase component proteins by solution x-ray scattering: conformation of the electron transfer transition state complex of the *Klebsiella pneumoniae* nitrogenase. *J. Mol. Biol.* 266:642–648.

Grossmann, J. G., M. Neu, E. Pantos, F. J. Schwab, R. W. Evans, E. Townes-Andrews, P. F. Lindley, H. Appel, W. Thies, and S. S. Hasnain. 1992. X-ray solution scattering reveals conformational changes upon iron uptake in lactoferin, serum and ovo-transferrrins. *J. Mol. Biol.* 225:811–819.

Gultyaev, A. P., F. H. D. van Batenburg, and C. W. A. Pleij. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* 250:37–51.

Henderson, S. J. 1996. Monte Carlo modeling of small-angle scattering data from non-interacting homogeneous and heterogeneous particles in solution. *Biophys. J.* 70:1618–1627.

Hill, C. P., N. L. Johnston, and R. E. Cohen. 1993. Crystal structure of a ubiquitin-dependent degradation substrate: a three-disulfide form of lysozyme. *Proc. Natl. Acad. Sci. USA.* 90:4136–4140.

Jones, G., P. Willett, and R. C. Glen. 1995. Molecular recognition of receptors sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* 245:43–53.

Jones, G., P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267:727–748.

Kobe, B., and J. Deisenhofer. 1993. Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature.* 366: 751–756.

Kuntz, I. D. 1971. Hydration of macromolecules. IV. Polypeptide conformation in frozen solutions. *J. Am. Chem. Soc.* 93:516–518.

Li, L., T. A. Darden, S. J. Freedman, B. C. Furie, B. Furie, J. D. Baleja, H. Smith, R. G. Hiskey, and L. G. Pedersen. 1997. Refinement of the NMR solution structure of the gamma-carboxyglutamic acid domain of coagulation factor IX using molecular dynamics simulation with initial Ca^{2+} positions determined by a genetic algorithm. *Biochemistry.* 36: 2132–2138.

Mayans, M. O., W. J. Coadwell, D. Beale, D. B. A. Symons, and S. J. Perkins. 1995. Demonstration by pulsed neutron scattering that the arrangement of Fab and Fc fragments in the overall structures of bovine IgG_1 and IgG_2 in solution is similar. *Biochem. J.* 311:283–291.

Ogata, H., Y. Akiyama, and M. Kanehisa. 1995. A genetic algorithm based molecular modelling technique for RNA stem-loop structures. *Nucleic Acids Res.* 23:419–426.

Pantos, E., and J. Bordas. 1994. Supercomputer simulation of small angle x-ray scattering, electron micrographs and x-ray diffraction patterns of macromolecular structures. *J. Pure Appl. Chem.* 66:77–82.

Pantos, E., H. F. van Garderen, P. A. J. Hilbers, T. P. M. Beelen, and van R. A. Santen. 1996. Simulation of small angle scattering from large assemblies of multi-type scatterer particles. *J. Mol. Struct.* 383:303–308.

Pedersen, J. T., and J. Moult. 1995. Ab initio structure prediction for small polypeptides and protein fragments using genetics algorithms. *Proteins.* 23:454–460.

Pedersen, J. T., and J. Moult. 1997. Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.* 269:240–259.

Perkins, S. J., A. S. Nealis, B. J. Sutton, and A. Feinstein. 1991. Solution structure of human and mouse immunoglobulin M by synchrotron x-ray scattering and molecular graphics modelling: a possible mechanism for complement activation. *J. Mol. Biol.* 221:1345–1366.

Perkins, S. J., K. F. Smith, J. M. Kilpatrick, J. E. Volanakis, and R. B. Sim. 1993. Modeling of serin-proteinase fold by x-ray and neutron scattering and sedimentation analysis: occurrence of the fold in factor D of the complement system. *Biochem. J.* 295:87–99.

Pessen, H., T. F. Kumosinski, and S. N. Timasheff. 1973. Small-angle X-ray scattering. *Methods Enzymol.* 27:151–209.

Pilz, I., O. Glatter, and O. Kratky. 1979. Small-angle x-ray scattering. *Methods Enzymol.* 61:148–243.

Pilz, I., E. Schwarz, D. G. Kilburn, R. C. Miller, Jr., R. A. Warren, and N. R. Gilkes. 1990. The tertiary structure of a bacterial cellulase determined by small-angle x-ray scattering analysis. *Biochem. J.* 271: 277–280.

Press, W. H., B. F. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1989. Numerical Recipes in Pascal. The Art of Scientific Computing. Cambridge University Press, New York.

Stuhrmann, H. B. 1970. Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particule-scattering function. *Acta Crystallogr.* A26:297–306.

Sun, S. 1995. A genetic algorithm that seeks native states of peptides and proteins. *Biophys. J.* 69:340–355.

Svergun, D. I., C. Barberato, and M. H. J. Koch. 1995. Crysol—a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28:768–773.

Svergun, D. I., M. H. J. Koch, and I. N. Serdyuk. 1994a. Structural model of the 50S subunit of *Escherichia coli* ribosomes from solution scattering. I. X-ray synchrotron radiation study. *J. Mol. Biol.* 240:66–77.

Svergun, D. I., J. S. Pedersen, I. N. Serdyuk, and M. H. J. Koch. 1994b. Solution scattering from 50S ribosomal subunit resolves inconsistency between electron microscopic models. *Proc. Natl. Acad. Sci. USA.* 91:11826–11830.

Svergun, D. I., and H. B. Stuhrmann. 1991. New developments in direct determination from small-angle scattering. 1. Theory and model calculations. *Acta Crystallogr. A.* 47:736–744.

Svergun, D. I., V. V. Volkov, M. B. Kozin, and H. B. Stuhrmann. 1996. New developments in direct shape determination from small angle scattering. 2. Uniqueness. *Acta Crystallogr.* A52:419–426.

Van Batenburg, F. H. D., A. P. Gultyaev, and C. W. A. Pleij. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* 174:269–280.

Wakabayashi, K., M. Tokunaga, I. Kohno, Y. Sugimoto, T. Hamanaka, Y. Takezawa, T. Wakabayashi, and Y. Amemiya. 1992. Small-angle synchrotron x-ray scattering reveals distinct shape change of the myosin head during hydrolysis of ATP. *Science.* 258:443–447.

Willet, P. 1995. Genetic algorithms in molecular recognition and design. *Trends Biotechnol.* 13:516–521.

Wistow, G., B. Turnell, L. Summers, C. Slingsby, D. Moss, L Miller, P. Lindley, and T. Blundell. 1983. Gamma-II crystallin at 1.9 angstrom resolution. *J. Mol. Biol.* 170:175–202.

Witz, J., S. N. Timasheff, and V. Luzatti. 1964. Small-angle x-ray scattering investigation of the geometry of β-lactoglobulin A tetramerization *J. Am. Chem. Soc.* 86:168–173.

Zheng, Y., P. C. Doerschuck, and J. E. Johnson. 1995. Determination of three-dimensional low-resolution viral structure from solution x-ray scattering data. *Biophys. J.* 69:619–639.