Acta Cryst. (1996). D52, 235-251

A Genetic Algorithm for the *Ab Initio* Phasing of Icosahedral Viruses

Stephen T. Miller,[a] James M. Hogle[a,b] and David J. Filman[b]*

[a]*Committee for Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA, and* [b]*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02155, USA. E-mail: dif@vp2.med.harvard.edu*

(Received 7 June 1995; accepted 21 August 1995)

Abstract

Genetic algorithms have been investigated as computational tools for the *de novo* phasing of low-resolution X-ray diffraction data from crystals of icosahedral viruses. Without advance knowledge of the shape of the virus and only approximate knowledge of its size, the virus can be modeled as the symmetry expansion of a short list of nearly tetrahedrally arranged lattice points which coarsely, but uniformly, sample the icosahedrally unique volume. The number of lattice points depends on an estimate of the non-redundant information content at the working resolution limit. This parameterization permits a simple matrix formulation of the model evaluation calculation, resulting in a highly efficient survey of the space of possible models. Initially, one bit per parameter is sufficient, since the assignment of ones and zeros to the lattice points yields a physically reasonable low-resolution image of the virus. The best candidate solutions identified by the survey are refined to relax the constraints imposed by the coarseness of the modeling, and then trials whose intensity-based statistics are comparatively good in all resolution ranges are chosen. This yields an acceptable starting point for symmetry-based direct phase extension about half the time. Improving efficiency by incorporating the selection criterion directly into the genetic algorithm's fitness function is discussed.

1. Introduction

The goal of this work is to develop the capability to determine phases for viruses and viral components which possess a high degree of internal symmetry, without relying on the availability of previously determined homologous structures or requiring isomorphous heavy-atom derivatives. Screening for heavy atoms in virus crystals is particularly laborious, because millions of reflections must be collected to determine if a candidate compound provides useful phase information.

Although collecting virus data is difficult, there are compensating computational advantages which arise from the high degree of non-crystallographic symmetry of the virus, as foreseen by Main & Rossmann (1966). Once a set of sufficiently non-random preliminary phases have been obtained, a high degree of non-crystallographic symmetry (NCS), such as that in icosahedral viruses, provides powerful constraints on the amplitudes and phases of the data, which are remarkably useful for refining initial phase estimates.

Provided that the amount of detail in the image is limited, the most efficient way to model physical measurements from highly symmetric viruses often involves describing the virus in terms of a small set of symmetry-consistent basis functions. Outside of a crystallographic context, noteworthy developments include the successful use of icosahedral harmonics (see Finch & Holmes, 1967) for the description of viruses in solution scattering experiments by Jack & Harrison (1975) and the use of 522 symmetric basis functions in the reconstruction of icosahedral viruses from electron micrographs (Crowther, 1971). In those applications, only the icosahedral symmetry is relevant, and a single set of analytic functions forms an appropriate basis set for the description of a variety of icosahedral particles.

In a crystallographic context, however, both the crystallographic and NCS operators have to be taken into account. Crowther (1967, 1969) suggested a highly efficient reciprocal-space matrix formulation of the NCS phase-refinement problem, explicitly involving the construction of mutually orthogonal eigendensities. In that formulation, each independent variable is the coefficient of an eigenvector of the NCS-averaging operator defined for a particular arrangement of symmetry-related copies in a specific unit cell. This approach reduces the degrees of freedom of the problem to the minimum number required to describe the non-redundant information content of the resolution-limited discrete Fourier transform of the unit-cell contents.

While highly efficient at low resolution, the computational expense of Crowther's formulation increases dramatically with increasing resolution, making it unsuitable for high-resolution applications. The current widespread use of NCS constraints in high-resolution applications derives from the seminal insight of Bricogne (1974), who noted that the expensive matrix multiplication step required in each iteration of Crowther's successive projection refinement algorithm could be replaced by a formally equivalent, but orders of magnitude less expensive, averaging of symmetry-related electron-density values in direct space.

Direct phase extension is the most dramatic application of the NCS phase-constraint procedure thus far. When phase estimates are available only for the lower resolution data, NCS can be used for extending those phases directly to higher resolution, provided that the upper resolution limit of the data is expanded sufficiently slowly. This phase-extension technique was utilized in low-resolution structure determinations by Argos, Ford & Rossmann (1975), Unge *et al.* (1980) and Rayment, Baker, Caspar & Murakami (1982). It was applied in a high-resolution structure determination by Wim Hol and coworkers (Gaykema, Volbeda & Hol, 1985) and was applied soon thereafter to the determination of high-resolution icosahedral virus structures by Rossmann *et al.* (1985) and Hogle, Chow & Filman (1985). Since that time, the determination of many additional virus structures has involved a direct phase-extension step, usually starting with phase estimates in the 5–8 Å range, but occasionally starting with phase estimates no higher than 13 Å (Valegård, Liljas, Fridborg & Unge, 1990). Direct extension over an even wider resolution range is generally expected to be feasible (Tsao *et al.*, 1992) and would play a critical role in the *ab initio* determination of virus structures, assuming that sufficiently accurate low-resolution phases were available.

Recent ideas for the phasing of low-resolution virus data have focused on the possibility of generating phases *de novo* (Tsao, Chapman & Rossmann, 1992), the incorporation of information from solution scattering (Chapman, Tsao & Rossmann, 1992), or on the use of electron micrographs as sources of low-resolution phases (McKenna, Xia, Willingmann, Ilag & Rossmann, 1992). Previously, the inclusion of electron-microscopy (EM) information has been qualitative, guiding the modeling of polyoma as an icosahedral arrangement of simple geometric solids (Rayment, 1983), or suggesting approximate shapes for envelopes used in subsequent high-resolution 'averaging'. In the past, EM-based phases and envelopes have proven useful, both alone (Jack, Harrison & Crowther, 1975) and in combination with heavy-atom information (Valegård *et al.*, 1990; Tsao, Chapman, Wu *et al.*, 1992; McKenna *et al.*, 1992), or as a guide to the rigid-body placement of previously determined molecular fragments (*e.g.* Cheng *et al.*, 1994; Stewart, Fuller & Burnett, 1993), though such low-resolution information has not yet proven to be sufficient by itself.

One attractive concept for the calculation of phases *de novo* involves applying NCS constraints to randomly generated sets of starting phases (McKenna *et al.*, 1992). This is widely believed to be possible if the data and envelopes were at atomic resolution, but it has not yet proved to be practical given the limits in the achievable quality and resolution of virus data. In practice, refinement from a random starting point (such as that provided by a non-isomorphous heavy-atom derivative) divides the reciprocal sphere into a patchwork of local areas, each of which contains phases that are mutually consistent with respect to the sign, hand and origin of the implicitly specified structure, but which are globally inconsistent with one another (see, for example, Tsao, Chapman & Rossmann, 1992). Direct phase extension usually has failed when too broad a resolution step was attempted, primarily because the barriers between these mutually inconsistent regions have proven impervious to continued refinement.

Recently, a promising purely computational approach was undertaken using atomic model-based synthetic data for canine parvovirus. Starting with a model of a hard uniform spherical shell with adjustable radii (Chapman *et al.*, 1992), an NCS-based phase refinement and direct extension were attempted (Tsao, Chapman & Rossmann, 1992). Because parvovirus crystallized with none of the icosahedral twofold axes parallel with crystallographic twofolds, it was conceivable for the asymmetry of the data to break the spherical symmetry imposed by such a simple model. Not surprisingly, broad shells of correctly phased data were generated, interleaved with shells corresponding to the Babinet opposite of the structure. At that point, the outermost data provided a large enough set of self-consistent phases for direct phase extension to work in an idealized calculation, though unfortunately not in an actual previous attempt to solve the virus structure *de novo*. Paradoxically, the same factors which make it possible to break the symmetry of a spherical phasing model in this space group also make it difficult to position the center of the virus accurately enough for successful direct phasing.

Clearly, there is a need to develop an efficient computational paradigm for the direct-space modeling of icosahedral viruses, to be used both as a means to explore the space of possible low-resolution crystal structures automatically, and to facilitate the incorporation of EM information. The procedure described in this manuscript involves modeling the icosahedrally unique volume of the virus with a small set of immovable lattice points distributed as uniformly as possible, subject to the limitation that the number of such lattice points cannot exceed the total information content of the discrete Fourier transform at the current resolution limit. (This treatment makes every possible model consistent with the known symmetry, and makes all of the basis functions mutually orthogonal eigendensities of the appropriate NCS-averaging operator.) Because each parameter of the model represents the average electron-density value in one particular vicinity, simply turning the density 'on' or 'off' in the earliest stages of the procedure yields a physically reasonable low-resolution portrayal of the virus. This treatment facilitates efficient refinement from multiple random starting points, using genetic algorithms (GA's) (Holland, 1975). These are computational survey methods enjoying what is effectively a built-in though very approximate history mechanism. In addition, a computational strategy has been developed for circumventing the most serious effects of series termination, both in low-resolution phase refinement, and in the direct phase-extension steps which are expected to follow.

A successful computational exercise using perfect data at 24 Å resolution is included to suggest the feasibility of this approach. This represents both an achievable resolution for EM work and a solution space small enough to be searched efficiently by the methods described here.

2. Methods

2.1. *Overview of the method*

This method for the phasing of low-resolution virus data is a four-step process. It consists of a coarse survey of the space of possible solutions using a genetic algorithm, a refinement which optimizes the statistically strongest answers from the survey, a selection procedure which identifies solutions containing no Babinet cross-over regions, and a final refinement which eliminates the coarseness of sampling imposed by the initial parameterization. In the first and most expensive stage of the calculation, trial structures are generated and evaluated to determine the degree to which model-based structure factors agree with the observations. Because the efficiency of the first three steps is critically dependent on the way in which the model is described, special care was taken in selecting a parameterization for the problem. The resulting parameterization differs significantly from traditional Cartesian systems based on the crystallographic asymmetric unit, and offers significant advantages in both conceptualization and efficiency. In the following section, a description of this parameterization is followed by a step-by-step explanation of its use in phase assignment and refinement. Finally, the method is applied to a test system (perfect data from empty capsids of the Mahoney strain of type 1 poliovirus) and its performance is assessed.

2.2. *Physical model of the virus*

2.2.1. *The usefulness of symmetry-consistent basis sets.* In any optimization scheme, it is desirable to define the space of possible solutions in terms of the smallest possible number of independent parameters. Efficiency also dictates that the constraints imposed by the mathematical formulation of the model be as consistent as possible with the actual physical attributes of the system. The most effective parameterization examined involved modeling the virus crystal as a linear combination of a very small number of mutually orthogonal icosahedrally consistent basis functions. This formulation represents one of many ways for constructing eigendensities of the non-crystallographic symmetry-averaging operator. The particular basis set described here also has a physical interpretation: each basis function corresponds to the assignment of unit electron density to one lattice point in the icosahedrally unique volume (see Fig. 1). Constructing the basis set in this way has a desirable consequence: the assigning of ones and zeros to the lattice points yields a physically reasonable low-resolution portrayal of the contrast between protein and solvent. Additionally, the GA-based survey becomes more efficient when only one computer bit is needed for each variable.

2.2.2. *Description of the icosahedrally unique volume.* The use of the term 'lattice' here is somewhat atypical, in that the pattern of regular spacing between grid points does not extend perfectly beyond the bounds of each unique volume. Instead, lattice points in non-crystallographic symmetry-related volumes are generated by the application of non-crystallographic symmetry operators to the spatial coordinates of the unique grid points.

The arrangement of lattice points within the icosahedrally unique volume was designed to optimize the uniformity of coverage within these volumes and between their symmetry-related copies. A close-packed regular tetrahedral array of lattice points was deformed slightly to fit within the irregular tetrahedron defined by three adjacent icosahedral fivefold axes and the center of the virus particle (Fig. 2). The distance between adjacent lattice points was scaleable, permitting any degree of coarseness in the final sampling. Because this slightly irregular tetrahedron covers exactly three copies of the icosahedrally unique volume, two thirds of these points, which are redundant due to symmetry need to be removed. Those lattice points lying within or on the borders of the kite-shaped icosahedral asymmetric

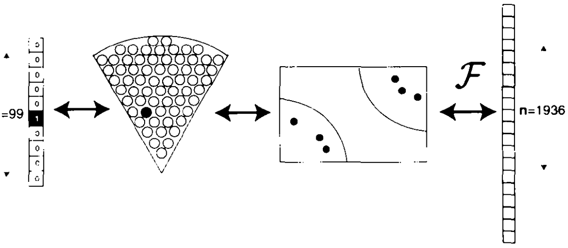


Fig. 1. Diagram of the virus-modeling scheme. From left to right: each element of a real-valued (possibly Boolean-valued) vector of length **m** specifies the electron-density value at one of **m** tetrahedrally arranged volume elements which form the (wedge-shaped) icosahedrally unique volume. By applying icosahedral and crystallographic symmetry operators, $\mathbf{N}_{\mathrm{sym}}$ equivalent points are generated in the crystallographic asymmetric unit (denoted here by a rectangle). Finally, Fourier transformation ($\mathcal{F}$) gives rise to the corresponding complex-valued vector of the **n** unique structure factors, shown at the right. In this parameterization of the virus model, each possible vector of length **m** (on the left) specifies one possible symmetry-consistent model. The symbols **m**, **n** and $\mathbf{N}_{\mathrm{sym}}$ are defined in the text.

volume delimited by one fivefold, one threefold, and two adjacent twofold axes of symmetry (corresponding to the irregular four-sided pyramid seen in Fig. 2) were retained. In addition, an outer spherical cutoff was imposed which was generously greater than the estimated radial extent of the virus (available from experiment), but which required no other prior knowledge of its size and shape.

Finally, all of the lattice points in the unique volume were shifted radially outward a short distance along a vector drawn from the center of the virus to the center of the four-sided base of the pyramid. This translation prevented any of the grid points from lying directly along any of the symmetry axes, and avoided the need to assign different multiplicities to points in different positions, which simplifies subsequent computations. (The extent of the shift, 2.2 grid units, allowed the separation between adjacent points in symmetry-related volumes to approximate the regular grid spacing.) This treatment ensures that the multiplicities of all lattice points are equal, and that each point is an equally weighted contributor to the refinement residual. There are unavoidable slight inequalities caused by the overlap of spheres near packing contacts between viruses, but it is believed that the effect of these small overlaps is minimal.

2.2.3. *Matrix-based Fourier transformation of the model.* One advantage of this approach is that calculating the Fourier transform of the unit-cell contents is greatly simplified. Once preliminary calculations have been completed, it becomes unnecessary to deal explicitly with the high degree of non-crystallographic symmetry. As shown in Fig. 1, each lattice point in the icosahedrally unique volume is expanded by the application of symmetry operators to create the set of all its symmetry-related copies in the unit cell. (Here, in space group $P2_12_12$, each point gave rise to 120 symmetry-equivalent copies.) This set was then Fourier transformed to yield a vector of the unique calculated structure factors, with the vector computed as in an atom-based structure-factor calculation. If the Fourier transform of the virus crystal includes exactly n unique reflections at the current resolution limit, then the reciprocal-space representation of each of the lattice points is a complex-valued column vector of length $\mathbf{n}$. Thus, the contribution of each lattice point to the set of unique $\tilde{F}_{\rm calc}$ is completely specified by multiplying its corresponding vector by the density value at that particular point.

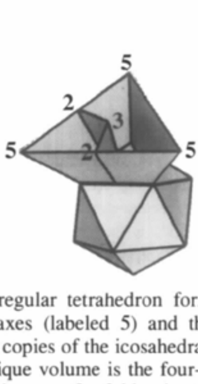


Fig. 2. The slightly irregular tetrahedron formed by three adjacent icosahedral fivefold axes (labeled 5) and the virus particle center contains exactly three copies of the icosahedrally unique volume. One possible choice of unique volume is the four-sided pyramid enclosed by the planes connecting one fivefold axis, two twofolds (labeled 2) and one threefold (labeled 3).

If m represents the number of lattice points in the icosahedrally unique volume, then in this parameterization of the virus crystal, every possible model of the electron density is constrained to be consistent with the symmetry of the virus, and can be specified completely by a real-valued (possibly Boolean-valued) vector of length $\mathbf{m}$. This trial vector of model parameters will be called $\underset{\sim}{\mathbf{w}}$. Once these $\mathbf{m}$ column vectors have been assembled into an $\mathbf{m}$-by-$\mathbf{n}$ matrix (here designated $\mathbf{A}$), calculation of the Fourier transform of each trial model (yielding the unique set of $\tilde{F}_{\rm calc}$, $\{\tilde{F}_{\rm calc}\}$) is accomplished by the computationally inexpensive matrix multiplication,

$$\{\tilde{F}_{\rm calc}\} = A\underset{\sim}{\mathbf{w}}. \qquad (1)$$

This simplification will permit the evaluation of the trial model, namely a comparison of $\{|F_{\rm calc}|\}$ with $\{|F_{\rm obsd}|\}$, at minimal cost.

Once the space group and unit cell are known and the positions and orientations of the virus particles have been specified (conditions that are true at least in favorable cases), there exists a unique matrix, $\mathbf{A}$, for any choice of $\mathbf{n}$ (where $\mathbf{n}$ is specified by the resolution limit) and arrangement of $\mathbf{m}$ lattice points in the asymmetric volume. Possible arrangements include both the complete sampling of the spherically limited icosahedral unique volume described earlier and alternatives in which the same number of lattice points are distributed within some smaller envelope created from an EM image or by a previous calculation at lower resolution. The more restrictive arrangement has the benefit of increasing the density of sampling without increasing the number of variables at any given resolution.

One practical implication of this design is that the cost of calculating the matrix $\mathbf{A}$ is borne only once, when the physical conditions of the experiment are first defined. The matrix $\mathbf{A}$ can then be reused repeatedly until the physical conditions of the experiment are changed, possibly reflecting an increase in resolution, a change in the envelope, or an improved estimate of the unit cell or position or orientation of the virus. (As currently implemented, the procedure runs on a multiprocessor SGI R4000 computer with shared physical memory, optionally allowing the matrix $\mathbf{A}$ to be shared among several coarsely parallelized processes.)

2.2.4. *Choosing the dimensions of* $\mathbf{A}$. The dimensions of $\mathbf{A}$ are determined by the conditions of the experiment. Once the size of $\mathbf{n}$ has been specified by the choice of resolution limit, the approximate size of $\mathbf{m}$ is constrained by the choice of $\mathbf{n}$. At any given resolution limit, the number of crystallographically unique pieces of information in the Fourier transform will be equal to the number of unique centric reflections added to twice the number of unique non-centric reflections (as the non-centrics have both real and imaginary components). Thus, if the crystal exhibits a degree of non-crystallographic symmetry, designated $N_{\rm sym}$ (here 30-fold), and the modeling exercise assumes that an $N_{\rm sym}$-fold expansion of **m** lattice points will completely account for the structure at the selected resolution limit, then an appropriate choice of **m** would be given by,

$$\mathbf{m} = (n_{\rm centric} + 2n_{\rm non\text{-}centric})/N_{\rm sym}. \qquad (2)$$

Fortunately, the present scheme for modeling the virus is not critically dependent upon the optimal choice of **m**. Choosing too large a value simply reduces the efficiency of the selection process by forcing the search of a larger solution space than necessary, while choosing too small a value slightly reduces the range of resolution over which the calculation is valid. Any error of the latter sort is apparent in the resolution dependence of the agreement statistics, and is corrected for in subsequent stages of the calculation.

2.2.5. *The genetic algorithm.* For macromolecules, the phase problem cannot be addressed easily with traditional optimization algorithms because the space of possible structures has an unwieldy number of dimensions, and, for most plausible choices of refinement residual, it is pocked with local minima. Techniques such as simulated annealing (Kirkpatrick, Gelatt & Vecchi, 1983) and Monte Carlo methods (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953) have been applied to some simple systems but may not represent the most efficient way to navigate such complex solution spaces. Genetic programming techniques offer an alternative strategy for searching solution spaces that are not well behaved.

Genetic programming methods are based on a rough analogy with Darwinian selection. In the same way that a population of natural life forms evolves, on average, toward a greater fitness for its environment, it is expected that a population of trial solutions to an optimization problem can be made to evolve into an improved set of solutions, assuming that both the appropriate selective pressures and a mechanism for change have been provided by the *in silico* environment. Often, the mechanism for applying a selective pressure is simply to bias the probability of reproduction in favor of those trials having greater fitness. This kind of machine learning procedure first was developed in the mid-1970's by Holland (Holland, 1975). Currently it is the topic of active research in computer science (see, for example, Forrest, 1993; Srinivas & Patnaik, 1994) and has provided useful solutions for a variety of problems. Examples of the successful use of genetic programming can be found in Davis (1991) and Chang & Lewis (1994).

Genetic algorithms constitute one subset of genetic programming techniques. In a genetic algorithm (GA), parameters expressing the set of possible solutions are encoded as bit strings (that is, strings of ones and zeros), where the fitness score of each bit string corresponds to some selected cost function (refinement residual) that results from an evaluation of the encoded parameter values. To create new trial solutions, the current large population of bit strings is regarded as a parental generation. Then, using the metaphor of sexual reproduction, members of the parental generation are selected and paired randomly to produce offspring. A similar number of progeny are created, each of whose bit strings represent some combination of its two parental strings. A selective pressure is applied to the process simply by giving the strings with better fitness scores a higher probability of reproducing. Over many generations, the population as a whole is expected to evolve towards a greater average fitness. This population shift increases the likelihood of discovering the correct solution to the optimization problem among the most fit members of the evolving population. This contrasts with an unbiased, random approach, such as Monte Carlo, which surveys the solution space in a more uniform way, and with simulated annealing which is an inherently more local approach.

In common implementations of a GA, the single most important operation in the production of progeny strings is called crossing over (though as a secondary operation, random single-bit mutations also are introduced with low frequency to prevent premature convergence of the population and to make all regions of the solution space theoretically accessible). In crossing over, part of the progeny string is contributed by one of the two parents, and the remainder of the string is contributed by the other parent, with the locations of the cross-over points chosen at random. This tends to preserve successful bit patterns which are located close together in the linear sequence of the genome, which, by design in this approach to virus modeling, frequently corresponds to lattice points which are close together in space.*

Genetic algorithms are believed to search the space of possible solutions more efficiently than purely random trials would, due to the phenomenon of implicit parallelization (Holland, 1975). In effect, GA's contain a built-in history mechanism wherein information about the fitness values of previously encountered bit strings is retained in the current population. This information is at least somewhat predictive of the fitness values of other strings not yet sampled which share some of the same bit patterns (though clearly, the accuracy of the predictions improves as the topography of the space being sampled becomes less complicated). For

* Indexing the lattice points approximately in the order of increasing radius, as done here, may facilitate shell formation, though this surmise has not yet been tested.

that reason, GA's have been recommended as potentially useful optimization procedures to try when attempting to solve computational problems, such as *ab initio* virus modeling, which have no known efficient solutions.

The specific implementation of the GA used in these experiments is the *GAucsd* program package (Schraudolph & Grefenstette, 1992). This well designed package requires the user to supply the subroutine to be used for evaluating the bit strings, but handles in a fairly transparent fashion the generation, maintenance and reproduction of populations of bit strings, in accordance with a set of user-selected attributes. A so-called 'roulette wheel' is employed by *GAucsd* to make the probability of reproduction of a bit string proportionate to its fitness. Here, each individual is assigned a segment of the wheel proportional in size to its user-defined fitness, and an individual is selected for reproduction when a 'spin of the wheel' lands within its angular range. In the case of virus modeling, whenever the fitness-evaluating subroutine receives a bit string from the main program (here interpreted as a Boolean vector of length **m**), the vector is multiplied by the matrix **A**, yielding the n unique Fourier coefficients corresponding to the model implicitly specified by the bit string. The fitness score then reported back to the main program could be any one of several plausible choices. To date, the two quadratic residuals used most often have compared the model based transform with the set of unique observations, expressed either as reflection intensities (I_{obsd} and I_{calc}) or as structure-factor magnitudes ($|F_{\text{obsd}}|$ and $|F_{\text{calc}}|$).

$$Q_{2I} = 1 - \langle I_{\text{obsd}} I_{\text{calc}} \rangle^2 / \langle I_{\text{obsd}}^2 \rangle \langle I_{\text{calc}}^2 \rangle. \quad (3)$$

$$Q_{2F} = 1 - \langle |F_{\text{obsd}}||F_{\text{calc}}| \rangle^2 / \langle |F_{\text{obsd}}|^2 \rangle \langle |F_{\text{calc}}|^2 \rangle. \quad (4)$$

In addition, the residuals Q_{1I} and Q_{1F} have been defined as the square roots of Q_{2I} and Q_{2F}, respectively. Each of these expressions can be recognized as a simple algebraic rearrangement of the unweighted normalized mean-square discrepancy, assuming that the least-squares value of the linear scale factor between I_{calc} and I_{obsd} or $|F_{\text{calc}}|$ and $|F_{\text{obsd}}|$ already has been incorporated into the expression. The fraction contained in each of these expressions resembles the square of the linear correlation coefficient, except that the best fitting line which is implicit in a correlation coefficient is constrained in this case to pass through the origin.

2.3. *First refinement step*

Once several runs of the GA (involving perhaps tens or hundreds of millions of trials) have identified a few promising regions of the parameter space, the next step of the procedure involves refining each of the candidate solutions. The purpose of this refinement is to relax the artificial constraint which limits each lattice point to only one of two possible values. This relaxation allows a more accurate modeling of the virus.

The refinement is a steepest descent minimization of Q_{2F} with respect to the electron densities at the **m** unique lattice points, here defined as $\mathbf{w}(1) \ldots \mathbf{w}(m)$. In each iteration of the refinement, a search direction for line-search minimization, here denoted $\Delta\mathbf{w}(1) \ldots \Delta\mathbf{w}(m)$, is obtained from the multiplication,

$$\Delta \underset{\sim}{\mathbf{w}} = B \underset{\sim}{\Delta} \tilde{F}_{\text{error}}. \quad (5)$$

Here, $\underset{\sim}{\Delta} \tilde{F}_{\text{error}}$ represents the vector of n unique difference coefficients,

$$\Delta F_{\text{error}}(h) = [|F_{\text{obsd}}(h)| - |F_{\text{calc}}(h)|] \exp[i\varphi_{\text{calc}}(h)] \quad (6)$$

(for $h = 1, \ldots, n$) with the set of F_{calc} appropriately scaled. The matrix **B** is defined as,

$$\mathbf{B} = [\text{Re}(\mathbf{A}^{*T}\mathbf{A})^{-1}]\mathbf{A}^{*T}. \quad (7)$$

The vector of density shifts thus obtained, $\Delta \underset{\sim}{\mathbf{w}}$, can be recognized as a slope-over-curvature expression for the least-squares minimization of Q_{2F}. As with the use of the matrix **A** described above, the matrix **B** needs to be calculated only once.

2.4. *Selection*

In the results which follow, the overall agreement statistics which consider all of the reflections at once are shown to have only a limited value in predicting the quality of the associated phase sets. In particular, high values of the residuals are diagnostic of poorly phased solutions, but low values of the residuals do not distinguish between good and bad answers. As will be shown below, these limitations of the residual were overcome by considering small resolution ranges individually rather than collectively. The selection procedure acts to bin the data into some small number of resolution ranges (which may overlap) and to evaluate the residual [(3) or (4)] in each resolution bin for each of the refined trial structures. Then, rather than accepting a trial solution because its bin-specific statistics are good relative to the remainder of the population ('survival of the fittest'), the selection process instead rejects those solutions which have the poorest bin-specific statistics in any of the bins (effectively 'elimination of the least fit').

Two methods for selection have been investigated which follow this paradigm. The results presented in this manuscript are based on the simpler of the two, which is easier to implement as it merely involves choosing the single best trial produced by each of the 200 GA runs, evaluated after they have been refined. Eliminating all but one of these 200 candidates was accomplished by visiting each of eight non-overlapping resolution bins in succession, and discarding some fixed percentage of the worst scoring trials in that bin, repeating the series of bins if needed, until only one (or a few) candidate solutions was left. Using this approach, the selection procedure would accept trial models which gave only moderately good statistics in any or all resolution ranges, so long as unusually bad statistics occurred in none of the resolution ranges. The second method, developed subsequently, involves a modification of the GA, and is discussed below.

2.5. *Final refinement step: increasing the fineness of sampling*

At the end of the procedure, a final refinement step was carried out to eliminate the coarseness of sampling imposed by the coarse-grained icosahedral lattice, but without changing the resolution limit used in the preceding steps. Thus, after a small number of promising candidates from the GA had been refined to remove the two valued density constraint, a still smaller number of the remaining candidates which survived the selection procedure were subjected to this final refinement. Using a locally developed memory-resident implementation of the method of Bricogne (1976) and a conventional $d/8$ grid, 20 cycles of iterated direct-space averaging were applied to initial phase sets specified by the trial models. This refinement, carried out at the same resolution limit as the previous steps, acted to eliminate the coarseness of sampling imposed by the coarse-grained icosahedral lattice.

2.6. *The need for improvements of the averaging calculation*

In the course of developing the memory-resident scripted map averaging procedure for use at extremely low resolution, an interesting phenomenon came to light. First, it was noted that even when a perfect set of low-resolution phases and amplitudes was used as a synthetic standard, one or more cycles of averaging at 24 Å resolution yielded calculated structure factors which differed from the input reference standard by a surprisingly large 18% in $R_1^{\rm cryst}$*. That calculation had used a 170 Å radius envelope which was spherical except in the region where spheres would overlap. When vector difference maps of the asymmetric unit (not shown) were calculated by Fourier inversion of the appropriately scaled vector difference coefficients,

$$\Delta\tilde{F}(h)_{\rm vec} = \tilde{F}_{\rm obsd}(h) - \tilde{F}_{\rm calc}(h), \qquad (8)$$

the source of this large discrepancy in Q_F became evident. A large negative difference density feature was seen at about 150 Å radius, just below the point on the spherical surface which contacts the neighboring virus particle. This feature nearly disappeared when the calculation was repeated with identical envelopes at a higher resolution limit.

In effect, the unaveraged map contains an unavoidable series termination ripple from the neighboring particle and from the solvent mask, which affects the icosahedral subunits adjacent to it much more than the distant ones. Because the electron density in affected subunits is no longer identical to the other symmetry-related subunits, any direct-space averaging operation which includes these affected grid points must yield a corrupted estimate of the averaged density. This makes it impossible to simultaneously satisfy both the constraints imposed by symmetry and the constraints of consistency with the low-resolution data until after the resolution of the calculation has been extended. Statistically, however, most of the effect appears to have been due to solvent-flattening alone, as observed previously in polyoma (Rayment, 1983), as $R_1^{\rm cryst}$ was reduced only to 16% when 30-fold averaging was omitted.

2.7. *Modifications to averaging and phase extension*

Although well known series termination effects in the application of non-crystallographic symmetry previously have been described (*e.g.* Rayment, 1983), they are pointed out here for three reasons: (1) because an appreciation of their severity is required to account for the high values of the refinement statistics; (2) because eventually they bear directly upon the goal of extending phases from low resolution to higher resolution; and (3) because they suggest an unconventional strategy for direct-space averaging at extremely low resolutions, which has, in fact, been utilized here.

Noting that the most severe of the series-termination effects with perfect data were localized to specific spots on the outermost portion of the spherical envelope, simple variants of the averaging process were devised to prevent the most severely affected points from undergoing density modification. Here, rather than applying 'solvent flattening' to the points in the input map which lie outside of the envelope, this routine simply leaves them alone, retaining their original values. Points within the envelope continue to be averaged with values interpolated at their symmetry-equivalent positions, as usual. With this procedural variant in place, the 170 Å radius envelope yielded an $R_1^{\rm cryst}$ of 10% in the test with perfect low-resolution phases. Limiting the radius to 155 Å had the effect of reducing the statistical discrepancy to 6%, and further shrinking the envelope to 140 Å lowered the discrepancy to only 2%.

Obviously, using this approach involves a deliberate decision to sacrifice some of the power of the non-crystallographic symmetry-constraint procedure by discarding many of the constraint relationships theoretically available, and its use probably should be discontinued once data of sufficiently high resolution are included. The benefit, however, is that the averaging system behaves more nearly as it should, preserving an input set of perfect phases and amplitudes nearly intact. With adequate data, the large set of constraint relationships which remain should be sufficient to refine the input phases towards symmetry-consistent values.

* $R_1^{\rm cryst} = \langle \big| |F_{\rm obsd}| - |F_{\rm calc}| \big| \rangle_h / \langle |F_{\rm obsd}| \rangle_h$.

A better compromise, to be investigated in the future, might involve creating a more sophisticated envelope having a larger radius and 'cut outs' near the particle contact points. Unlike the traditional averaging approaches which include solvent flattening, this unconventional averaging approach should be fairly tolerant of errors in the shape of the envelope. Using the envelope solely to distinguish points which are averaged from points which are not, automatically eliminates the potential danger of 'chopping off' and flattening portions of a macromolecule lying outside of an accidentally misconfigured envelope in the early stages of a structure determination.

2.8. *Improvements in the GA*

Given the success of the rather simple selection procedure outlined above, an additional effort was made to incorporate the resolution-dependent 'elimination of the least fit' principle directly into the GA. This was expected to be a much more efficient process than refining against one criterion and then selecting against a different one. With experimentation, it was discovered that a refinement residual could be designed which had the desired characteristic that trial solutions which were particularly poor in any resolution range (relative to the other contemporaneous trials) were penalized severely, but that trial solutions which were particularly good were given only a slight advantage. One such residual, normalized for the number of bins (N), took the form,

$$\mathbf{Q}_{\text{exp}} = (1/N)\sum_{N} \exp[k(\mathbf{Q}_{\text{bin}} - \mu_{\text{bin}})/\sigma_{\text{bin}}] \qquad (9)$$

where $\mathbf{Q}_{\text{bin}}$ is the Q_{1I} or Q_{1F} residual of a particular bit-string trial in a specific resolution range (overlapping or non-overlapping), and k is an empirically determined constant. μ_{bin} and σ_{bin} are recently updated estimates of the mean and standard deviations of all previously encountered values of $\mathbf{Q}_{\text{bin}}$, weighted in a way that progressively decreases the influence of older individuals. Computationally, it is convenient to calculate μ and σ by updating the bin-specific running sums S_0, S_1, and S_2, every time a new bit string (or generation of bit strings) is evaluated. Thus, for $j = 0$, 1 and 2, the current value of S_j is,

$$S_j \leftarrow tS_j + \mathbf{Q}^j_{\text{bin}}, \qquad (10)$$

where t is a multiplier slightly less than one. The mean and standard deviation for each bin are then given by,

$$\mu = S_1/S_0, \qquad (11a)$$

$$\sigma = [(S_2/S_0) - \mu^2]^{1/2}. \qquad (11b)$$

It must be noted, however, that this approach deviates significantly from the simple paradigm for the GA wherein any given bit string always is characterized by the same fitness value, regardless of when it is evaluated. Instead, the Q_{exp} score of any particular string becomes worse as the overall fitness of the population improves with time. One way to avoid mistakenly choosing a poor string which happens to score relatively well early in the GA simply requires that a few of the best individuals in each generation be retained into the following generation (a strategy sometimes referred to as 'incomplete replacement', or 'elitist survival', which is a standard option of the *GAucsd* package). With this approach, the best bit string generated by this procedure is then easy to identify as the best surviving individual in the final generation. Thus far, preliminary tests with this latter method have yielded answers comparable with the simpler selection method, though in considerably less time (data not shown).

3. Results

3.1. *Test structure for the calculations*

The experiments reported here are idealized attempts to generate phases for poliovirus at 24 Å resolution. These experiments begin with error-free data calculated from an atomic model, a generous overestimate of the radius of the virus (which could easily have been obtained from the dimensions of the unit cell and packing considerations, or from electron micrographs of frozen hydrated samples), and knowledge of the position and orientation of the virus in the crystal. In many instances, the orientation of the particle can be determined experimentally (very accurately at low resolution) by the use of icosahedrally locked rotation searches *versus* the higher resolution data (Tong & Rossmann, 1990). Meanwhile, the position may be known from the space group and packing considerations.* These 24 Å experiments are a necessary precursor to experiments at higher resolution, which will be forthcoming.

These test calculations are based on an actual crystal structure. The high-resolution structure of native antigenic empty capsids of the Mahoney strain of type 1 poliovirus has been reported by Basavappa *et al.* (1994) These virus assembly intermediates crystallize in space group $P2_12_12$ with $a = 322.9$, $b = 358.0$ and $c = 380.1$ Å with one half virus particle per asymmetric unit. The virus particle center is located on the crystallographic twofold axis (c), almost exactly at $z = 1/4$, with one of the icosahedral twofolds coincident with the crystallographic twofold. The orientation of the virus particle in the $P2_12_12$ cell can be described by noting that two of the other icosahedral twofold axes which are perpendicular to the c axis are rotated 2.3° away from the directions of the crystallographic principals a and b. Crystals of the native empty capsids are nearly isomorphous with crystals of mature poliovirions, and the orientation

* Though an insufficiently accurate particle position is sometimes fatal, the phase-determination procedure suggested here would increase the resolution limit of the calculation very slowly, permitting the particle position to be refined as the phase determination progressed.

parameters of the mature virus (Hogle *et al.*, 1985) were determined originally from the observed data using an icosahedrally locked rotation function at high resolution.

The values for **m** and **n** follow directly from the resolution of the experiment. At 24 Å there are 1936 unique reflections, and this specifies the value for **n**. **m** was chosen to be 99, based on (2), given that $\mathbf{N}_{\rm sym}$ is 30 and that approximately half of the unique reflections are centric. Hence, in these experiments, **A** is a 99-by-1936 matrix with complex-valued entries, and **B** is 1936-by-99.

3.2. *The treatment of bulk solvent*

Several different GA based searches were conducted, of which only the two most promising are described here. These searches differed only in their description of the density level of the bulk solvent relative to that of the protein. Because the volume of the unit cell outside of the spherical envelope is not populated by any of the **m** unique lattice points or their symmetry-related copies, this entire volume implicitly has a density of zero. However, this model is not entirely realistic, as the actual bulk-solvent region has a non-zero average density. In the highly simplified virus model utilized here, there are several plausible ways to account for the influence of the bulk solvent. In the most straightforward, the bit-string values of 1 and 0 can be mapped to any two specified electron-density levels which, in the absence of a known F_{000} term, represent offsets from the bulk solvent level, which is then mapped to zero.

In the first of the GA experiments, the bit values 1 and 0 were mapped to 1.0 and 0.0, respectively, so that the bulk solvent was equivalent to the background level in the interior of the virus particle. In the second set of experiments, the two bit-values were mapped to 1.0 and -1.0, causing the bulk-solvent level to be exactly half of the highest protein density. This latter mapping should result in the creation of positive and negative images of the virus equally often, and either answer is acceptable. Efforts to optimize the relative bulk-solvent level might be productive in attempts to solve a real crystal structure; however, the current use of ideal data calculated from an atomic model makes that line of investigation irrelevant.*

3.3. *GA test calculations*

Two separate genetic algorithm experiments were undertaken using the *GAucsd* package, differing only in their description of the bulk-solvent background. In each case, on the order of 200 trials were run in each of which a random starting population of 500 bit strings of length 99 was subjected to evolutionary pressure until either 42 000 individuals had been evaluated or the population had reached convergence (which occurred in about 15% of the trials). The entire population was replaced in each generation and cross-over and mutation rates were 1.4 and 0.000170, respectively, which correspond to the default values suggested by *GAucsd*. Each experiment consumed 6 d of computation time on a Silicon Graphics R4000 processor and represents the evaluation of more than seven million test models. For each of the roughly 200 trials in each experiment, the single Boolean vector with the best fitness score was recorded for future use. Thus, the seven million evaluations yielded about 200 candidate solutions in each experiment.

3.4. *Behavior of the fitness function during the procedure*

In the earliest generation of the GA, with coarse sampling and binary-valued parameters, the average Q_{1I} score tended to average 0.58 or 0.96, depending on whether a zero or a non-zero bulk-solvent model had been included. These values were consistently better than 1.0, which is the theoretical value for an entirely random structure, presumably because the appropriate icosahedral symmetry has automatically been imposed. By the final generations of the GA, Q_{1I} scores around 0.21 and 0.18 were typical (and on average, the best scores in each experiment were only 0.01–0.02 better). Given that the GA is explicitly a minimizer of Q_{1I}, these relatively high values suggest that the coarseness of the model has limited the extent to which the observed and calculated transforms can possibly agree.

After the GA, the lowest scoring trial from each random starting population was refined to relax the binary valuation constraint, assuming each of three possible solvent models during the least-squares minimization of Q_{2F}. Over the course of the refinement, the Q_{1F} values which were monitored dropped significantly. Thus, when the bulk-solvent level was matched to the background density in the interior of the protein, the average overall Q_{1F} score for the best solution in each of the 203 trials dropped from 0.61 to 0.31 over the course of the refinement. In the other reported search, the bulk-solvent level was equivalent to one half of the protein density, and the average Q_{1F} residual among 178 candidate solutions dropped from 0.52 to 0.24 as the refinement progressed.

Finally, in the last refinement stage, the lattice coarseness constraint was relaxed using conventional iterated map averaging (Bricogne, 1976). Here the number of $\mathbf{w}(x)$ variables used to describe the solution had increased

* The assignment of 0.5 to the bulk solvent causes a physically unrealistic situation in which some regions of the virus are assigned densities below the bulk-solvent level. In practice, however (see Table 1, below), these solvent models yield correct phases more often than those where the bulk-solvent level is 0.0. A likely explanation is that at the outset of the GA, a random arrangement of 'black' and 'white' points creates a 'gray' sphere which contrasts strongly with a white background (0.0), but not with a gray one (0.5). This artifact makes the starting low-resolution phases much better than random, though they fail to improve as much over the course of the GA experiment (data not shown), presumably because of a built-in bias toward an incorrect structure. In contrast, the population which is entirely unbiased at first refines to a better set of solutions, even though the models are forced to remain non-committal about the contents of the solvent region until the refinement stage.

dramatically, while the information content of the low resolution Fourier transform remained the same as before. Not unexpectedly, dramatic improvements often were seen in Q_{1F}, which was reduced to the 0.06 to 0.11 range.*

3.5. *Phasing results of the GA*

In order to track the performance of the algorithms under development it was necessary to calculate comparisons with correct phases, even though the algorithms themselves are based on amplitude information alone. In view of the reasonably good correlation between the centric and non-centric statistics (see Fig. 3*a*, for example), it is sufficient to focus only upon the centric statistics. The panels in Fig. 3 are scatter plots, with each point representing one of the 200 candidate solutions from a single GA experiment. Here, the percentage of correct centric phases is plotted *versus* the Q_{1F} residual, either with all of the unique reflections included (Fig. 3*b*), or with different resolution ranges considered separately (Figs. 3*c*–3*h*). It is clear that the overall residual score is positively correlated with the internal self-consistency of the phase set, though the degree of correlation varies among resolution ranges. Significantly, this correlation is due primarily to the fact that test solutions with largely incorrect phases yield high residuals, while lower values of the residual are not particularly reliable indicators of phase consistency.

3.6. *Phasing results of refinement*

The behavior of the centric phases over the course of the procedure is illustrated in Figs. 4 and 5. As examples, two of the final models identified by the selection procedure are included, one typical of an acceptable solution (Fig. 4), and one typical of an inconsistently phased solution (Fig. 5). Although the solution plotted in Fig. 5 was obtained from the void bulk-solvent experiment and the solution in Fig. 4 was from the 0.5 solvent level search, these solutions were selected as illustrative and typical of the results obtained from either of the searches.

3.7. *The good solution*

In the acceptable solution, the bit-string trial identified by the GA (Fig. 4*a*) yielded an overall Q_{1I} residual slightly better than average (0.1564), and the centric phases initially were self-consistent to 50 Å. Relaxation of the two-valued constraint on the density in the first refinement step acted to improve the phase set so that no extensive cross-over regions remained, though randomization of the phase set was seen at several points in the higher resolution end of the transform (Fig. 4*b*). The corresponding map (Fig. 4*e*) showed substantial improvement. The averaging step (Fig. 4*c*) improved the phase consistency to the point that 75% of the centric phases were correct and the corresponding map (Fig. 4*f*) was qualitatively similar to a map phased perfectly at 24 Å (Fig. 4*g*). Although the phases beyond 30 Å still need improvement, this map would be sufficiently accurate to be the basis for tighter non-geometric envelopes in subsequent stages of the structure determination (see below). Though the phases from the model in Fig. 4 happen to be generally in agreement with the reference phases, a set with the opposite phases would have been equally acceptable.

3.8. *The poor solution*

In contrast, the phase plots for the poorly phased solution contain several oscillations (Fig. 5*a*–5*c*), and the corresponding maps (Figs. 5*d*–5*f*) look very little like the reference structure. Each peak and trough in the plots represents a local group of reflections whose phases are consistent with respect to sign and hand, but phases in the peak and trough regions are not consistent with one another. Most often, the incorrectly phased data represent the negative (or Babinet opposite) of the structure, though the hand of the corresponding structure may vary in certain space groups, and the conditions of the experiment enforce a consistent choice of origin. Similar patterns commonly have been observed previously in NCS refinement starting with random or inconsistent phases (see above). These phase cross-overs, present in the vast majority of the millions of random trials, are precisely what the binned selection procedure seeks to identify and eliminate. The task of determining a self-consistent starting point for phase extension thus can be reduced to finding a transform with no cross-over points or, alternatively, to finding one which is self-consistent out to as high a resolution as possible. Observe in Figs. 5(*b*) and 5(*c*) that despite continued refinement, most of the phase cross-overs have persisted. This suggests that selection from a large population, rather than minimization, is likely to be the most fruitful approach.

The significant phase changes caused by refinement can be seen in Figs. 4 and 5 by comparing parts (*a*) and (*b*). Changes seen in Fig. 5, which are less dramatic than in Fig. 4, are fairly typical of the large majority of the trials. In the experiments described in Table 1, refinement caused 27–48% of the centric reflections to change. Subsequent relaxation of the lattice coarseness constraint clearly affects the phases to a lesser extent, as seen by comparing parts (*b*) and (*c*). In both Figs. 4 and 5, small improvements at higher resolution are

* The Q_{1F} and Q_{1I} residuals share a common minimum, though Q_{1I} is more strongly influenced by the largest reflections, including several in the lowest resolution ranges. The Q_{1I} was used in the GA, and reported here, because acceptable answers were produced more often than when Q_{1F} was used (perhaps because the coarse model is much less accurate at higher resolution). It is expected that the method is not critically sensitive to the choice of residual, and no systematic effort has yet been undertaken to optimize it.

achieved at the expense of worsening agreement in the lowest resolution portion of the transform.

3.9. *Phasing results of selection*

After the initial refinement step, the simpler selection procedure was run in two slightly different ways, to select a single best answer from each of the six reported experiments. Details of the experiments are given in the legend of Table 1. The Q_{1F} residual and the fraction of correct centric phases (denoted Fract) are listed individually for each resolution bin and collectively for the 24 Å data. Both high and low percentages of correct centric phases are indicative of acceptable solutions, as they both represent self-consistent solutions. It is clear from this table that the overall residual is not sufficient to discriminate between acceptable and poor solutions, as the solution with the best overall residual (0.1431)

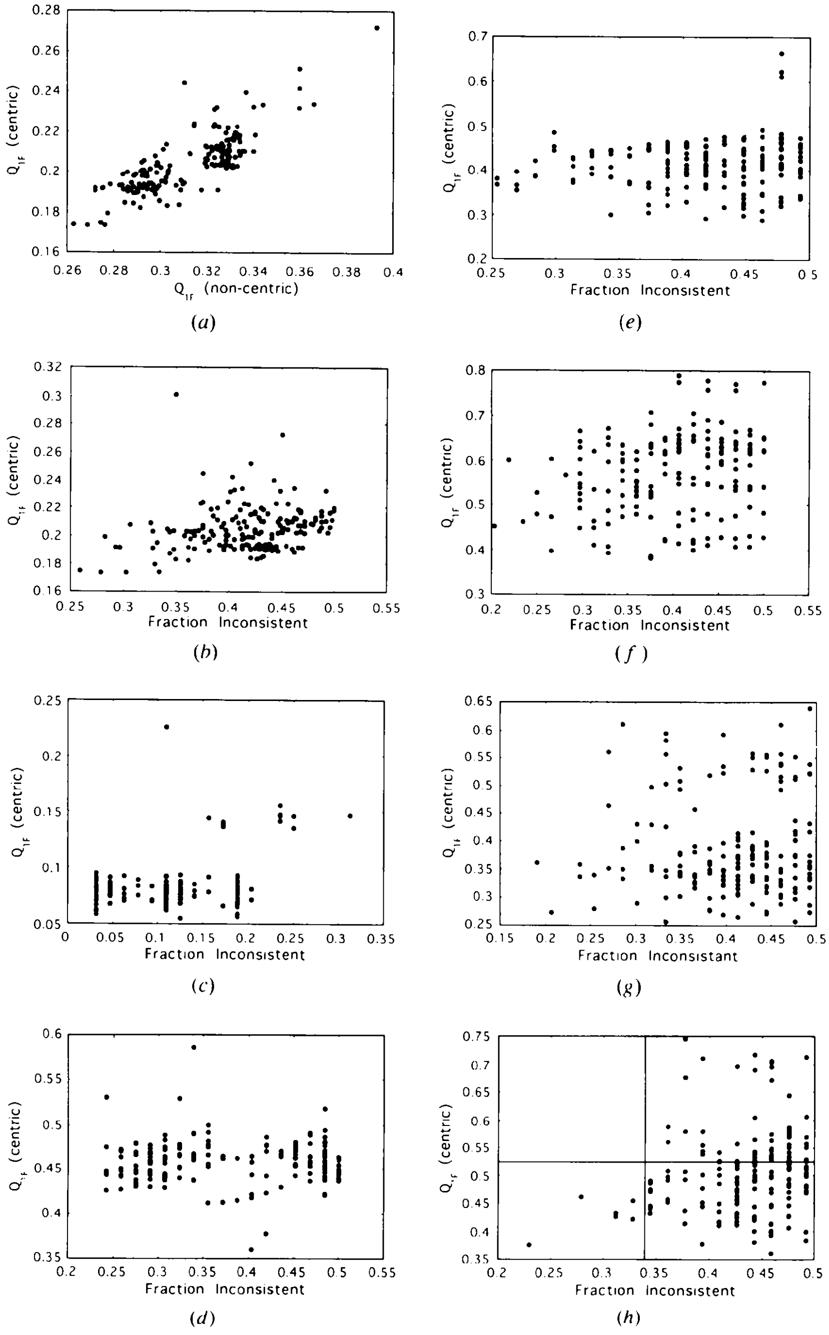
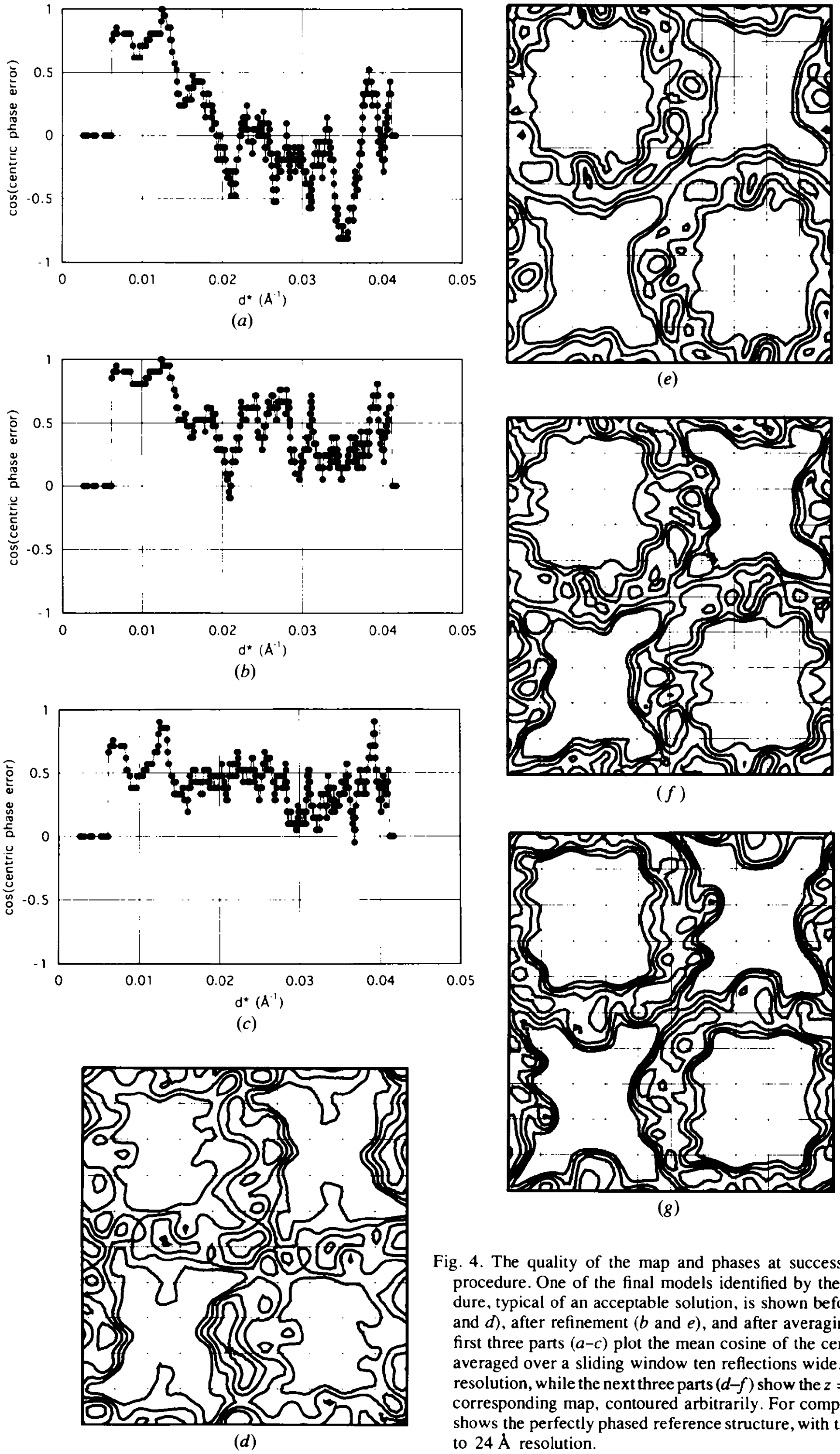


Fig. 3. Scatter plots indicating the extent to which the amplitude-based statistic, Q_{1F}, serves as a predictor of phase consistency. Each of the ~200 points in each panel represents the best-scoring outcome of a single GA experiment. (*a*) The correlation between centric and non-centric statistics. (*b–h*) The correlation between centric statistics and phase consistency. Panel (*b*) includes all centric reflections to 24 Å, while each of the panels (*c–h*) was calculated using reflections exclusively from one of the resolution bins (1–6) specified in Table 1. Here, the abscissa represents a measure of internal self-consistency: the fraction of the centric phases which agree or disagree with the reference standard, whichever is the lesser. Thus, zero would be perfectly consistent, and 0.5 represents a random outcome. The crossed lines in panel (*h*) are drawn to emphasize the salient point: that values of the residual which are high relative to the other scores in the same bin (*i.e.*, points lying above the horizontal line) often identify reflection sets which are among the most inconsistently phased (points lying to the right of the vertical line).



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 4. The quality of the map and phases at successive steps of the procedure. One of the final models identified by the selection procedure, typical of an acceptable solution, is shown before refinement (*a* and *d*), after refinement (*b* and *e*), and after averaging (*c* and *f*). The first three parts (*a*–*c*) plot the mean cosine of the centric phase error, averaged over a sliding window ten reflections wide, as a function of resolution, while the next three parts (*d*–*f*) show the $z = 0$ section of the corresponding map, contoured arbitrarily. For comparison, panel (*g*) shows the perfectly phased reference structure, with the data truncated to 24 Å resolution.

includes at least one severe cross-over between bins 4 and 5, and the overall fraction of self-consistent centric phases (46%) is close to a random outcome.

Even though some of the chosen phase sets are incorrect, results of the sort presented in Table 1 would be a satisfactory outcome in the solution of an actual

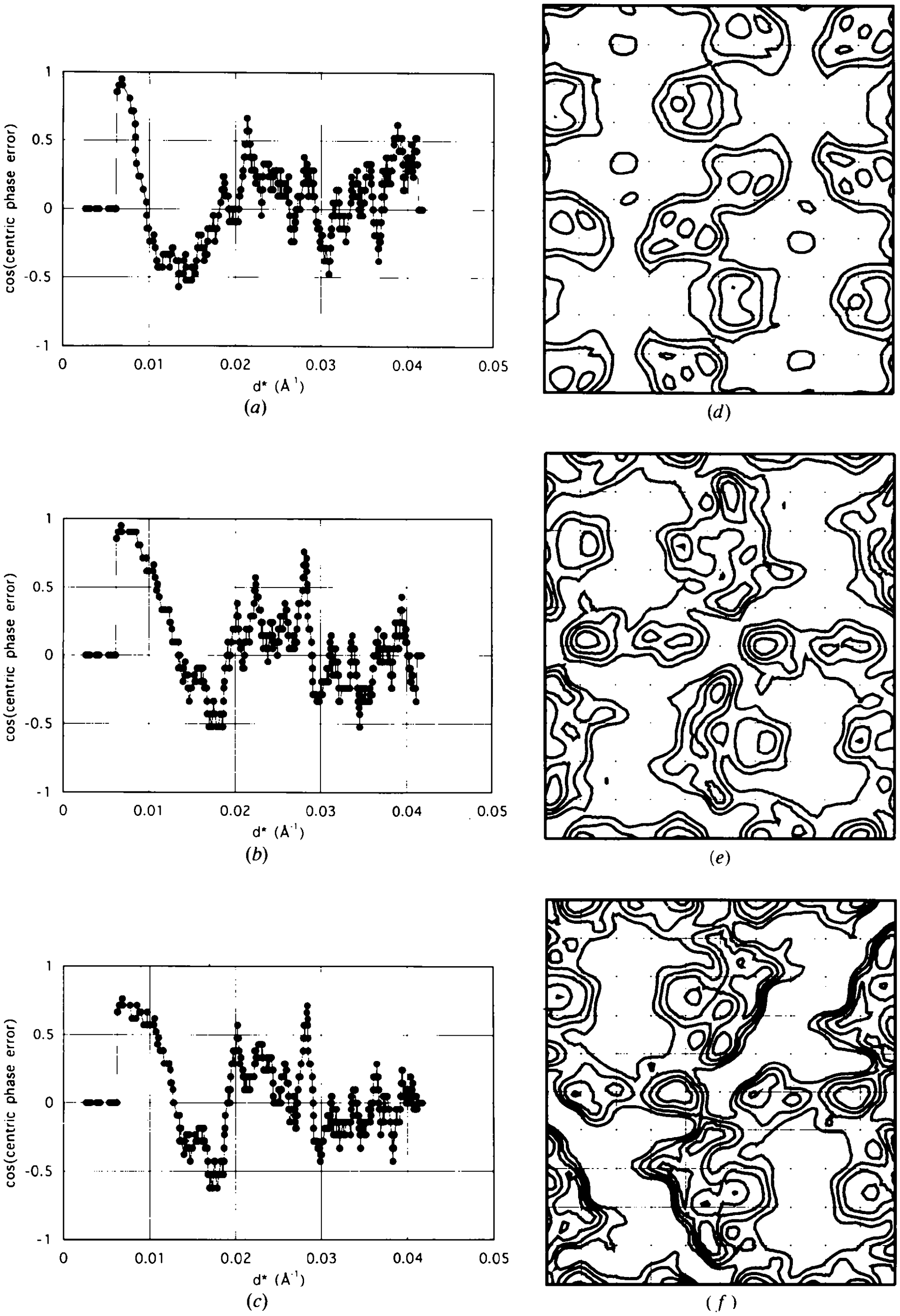


Fig. 5. The effect of successive steps of the procedure on an inconsistently phased trial. One of the final models identified by the selection procedure, typical of an unacceptable solution, is shown before refinement (*a* and *d*), after refinement (*b* and *e*), and after averaging (*c* and *f*). The plots (*a*–*c*) and the maps (*d*–*f*) are explained in Fig. 4. When first selected by the GA, this solution had an overall Q_{11} residual slightly better than average (0.1791). In this poorly phased solution, all of the refinement steps have tended to create a commonly seen pattern in the transform in which resolution shells of correctly phased data alternate with shells of data whose phases are consistent with the non-crystallographic symmetry, but inconsistent with the phases in other resolution ranges.

Table 1. *Resolution-dependence of the phasing statistics for 12 automatically selected models*

The unique centric reflections, listed in increasing resolution order, have been grouped into eight equal non-overlapping bins. For each resolution bin, and for the centric data as a whole, the fraction of correct centric phases ('Fract.') and the value of the amplitude-based selection criterion (Q_{1F}) have been listed. The models presented here were chosen as follows: First, the GA experiments labeled I18h and I28h were run repeatedly, using Q_{1I} as a selection criterion, and assuming solvent levels of zero and 0.5, respectively. Then, the single best trial from each experiment was further refined *versus* Q_{2F}, to eliminate the binary-valuation constraint. These refinements, labeled F0h, F1h and F2h, represent mapping of the bulk solvent levels to 0, 1, and 0,5, respectively. Finally, a single 'best' answer was selected automatically from among the 200 refined possibilities in each of the six experiments by considering the resolution bins in succession, and discarding one or more of the worst scoring trials in that bin. If necessary, this process was repeated until only a single representative of each experiment was left. As it turned out, the particular set of six solutions that was chosen changed when the bins were considered in a different order, though the quality of the results were similar in both of the sets of six, shown above.

	Bin 1		Bin 2		Bin 3		Bin 4	
Resolution (Å)	∞–67.85		67.85–47.98		47.98–39.17		39.17–33.62	
Experiment	Q_{1F}	Fract.	Q_{1F}	Fract.	Q_{1F}	Fract.	Q_{1F}	Fract.
I18h.01.F0h	0.0390	0.0312	0.2457	0.1129	0.3914	0.2836	0.4583	0.4219
I18h.01.F1h	0.1575	0.2031	0.3381	0.3871	0.3418	0.3881	0.3743	0.4375
I18h.01.F2h	0.0517	0.0938	0.2919	0.2903	0.2895	0.3731	0.3770	0.4219
I28h.01.F0h	0.0567	0.0312	0.4335	0.2742	0.3922	0.2687	0.4015	0.2344
I28h.01.F1h	0.0579	0.0312	0.4313	0.2581	0.3564	0.2388	0.3722	0.2656
I28h.01.F2h	0.0620	0.8906	0.4300	0.7097	0.3665	0.7463	0.4519	0.7969
I18h.01.F0h	0.0523	0.0312	0.4434	0.2903	0.3107	0.3881	0.4019	0.5312
I18h.01.F1h	0.1171	0.2188	0.4968	0.5645	0.2851	0.3881	0.4071	0.3906
I18h.01.F2h	0.1259	0.2188	0.3805	0.4032	0.3622	0.4179	0.3996	0.3594
I28h.01.F0h	0.0567	0.0312	0.4335	0.2742	0.3922	0.2687	0.4015	0.2344
I28h.01.F1h	0.0579	0.0312	0.4313	0.2581	0.3564	0.2388	0.3722	0.2656
I28h.01.F2h	0.0615	0.0312	0.4325	0.2742	0.3658	0.2687	0.4619	0.2344

	Bin 5		Bin 6		Bin 7		Bin 8		Overall	
Resolution (Å)	33.62–30.34		30.34–27.70		27.70–25.64		25.64–24.00		∞–24.00	
Experiment	Q_{1F}	Fract.	Q_{1F}	Fract.	Q_{1F}	Fract.	Q_{1F}	Fract.	Q_{1F}	Fract.
I18h.01.F0h	0.3586	0.4127	0.5095	0.2295	0.5722	0.2769	0.6618	0.3500	0.1472	0.2648
I18h.01.F1h	0.4025	0.6349	0.5525	0.6066	0.6722	0.5538	0.6579	0.4667	0.2121	0.4585
I18h.01.F2h	0.2781	0.6508	0.4898	0.6721	0.5300	0.5385	0.7386	0.5500	0.1431	0.4466
I28h.01.F0h	0.2729	0.4127	0.3467	0.4426	0.5577	0.3538	0.5842	0.3333	0.1735	0.2925
I28h.01.F1h	0.2642	0.4286	0.3955	0.3934	0.5863	0.3231	0.5976	0.2333	0.1718	0.2708
I28h.01.F2h	0.2728	0.7937	0.3933	0.5410	0.5166	0.6000	0.5878	0.6833	0.1735	0.7213
I18h.01.F0h	0.2547	0.5873	0.4217	0.5902	0.6369	0.5231	0.6729	0.4667	0.1708	0.4249
I18h.01.F1h	0.3115	0.5714	0.3936	0.6230	0.6075	0.4769	0.5891	0.4833	0.2038	0.4625
I18h.01.F2h	0.3478	0.4444	0.4216	0.5574	0.6162	0.4154	0.6236	0.4167	0.1946	0.4032
I28h.01.F0h	0.2729	0.4127	0.3467	0.4426	0.5577	0.3538	0.5842	0.3333	0.1735	0.2925
I28h.01.F1h	0.2642	0.4286	0.3955	0.3934	0.5863	0.3231	0.5976	0.2333	0.1718	0.2708
I28h.01.F2h	0.2567	0.3333	0.3781	0.3934	0.5413	0.3077	0.5823	0.2333	0.1746	0.2589

crystal structure because a sufficiently large proportion of the chosen solutions (here, about half) are acceptably self-consistent. Direct phase extension to high resolution could affordably be applied to several of the candidate solutions. Eventually, the agreement of any of the high-resolution images with amino-acid sequence and stereochemical information would provide the definitive objective test of the accuracy of the image.

4. Discussion

4.1. *Inadequacies of an overall residual*

The results above demonstrate that the simple resolution-independent overall residuals tested here were not sufficient either to evolve or to select acceptably self-consistent solutions to the virus phasing problem. However, alternative formulations of the residual or of the selection criterion which grouped the reflections according to resolution have given acceptable results with ideal test data, and have suggested promising avenues for future investigation.

It should be emphasized that the inability of the unmodified Q_F and Q_I residuals alone to provide a sufficiently powerful filter could not have been predicted in advance, partly because they represent an entirely conventional approach to the application of NCS constraints. If the reflections are not weighted, the simple Q_{2F} refinement residual is the very same residual implicitly minimized by successive projection in the methods of Crowther (1969) and Bricogne (1974); and similarly, Jacobson, Elkin, Hogle & Filman (manuscript in preparation) have successfully applied NCS constraints to a high-resolution virus structure by the minimization of Q_I. It was conceivable that the ability of the GA to escape from local minima would prove sufficient, providing that the residuals themselves were serving as sufficiently accurate indicators of global phase consistency.

Neither is the incorporation of a resolution-dependent binning scheme into NCS refinement a particularly radical idea. Beginning with the structure of *P*1/Mahoney in 1985 (Hogle *et al.*, 1985), all of the high-resolution icosahedral virus structures solved in this laboratory have utilized this common technique. An important benefit of binning the reflections according to resolution is that binning provides a reliable way to avoid potential mis-scaling of the transform of the model as a function of resolution. Such mis-scaling may have contributed to the inadequacy of the overall statistic.

Two principal reasons can be offered to explain why the simple overall residuals tested here were not sufficient.

4.1.1. *Low values do not guarantee good solutions.* Ideally, if the residual utilized as a fitness function were applying the proper evolutionary pressure, there would be a strong positive correlation between the statistical fitness of the trial structures and the extent to which their model-based centric phases were correct. Fig. 3 serves to illustrate one of the most salient observations: that while such a positive correlation does exist, it is primarily due to the fact that poor fitness values are diagnostic of bad phase sets. In particular, the models yielding the best of the phase sets do not necessarily produce particularly good overall statistical values, and, indeed, the lowest residuals often are associated with unacceptable solutions that include phase cross-overs.

The ability of inconsistently phased transforms to produce low overall values of Q_F makes these unweighted statistics poor detection criteria. Presumably, the unfavorable influence of the relatively small number of reflections in the narrow cross-over bands between adjacent Babinet-related shells (which were discussed above) was vastly outweighed by the collective influence of the many reflections whose phases were consistent with those of their immediate neighbors. Nevertheless, the results presented in Table 1 suggest that the narrow shell of reflections that is directly involved in a phase cross-over region may be able to produce diagnostically bad shell-specific statistics, provided that a meaningfully large enough number of reflections is present in the shell, and provided that the shell is narrow enough to exclude large numbers of self-consistently phased ones. Hence, the success of the resolution-binning procedure may partly be due to the increased relative influence of the narrow shells of poorly phased reflections.

4.1.2. *Lattice coarseness.* The second major problem associated with the use of a simple, resolution-independent residual arises from the coarseness of the model and from series termination error, both of which have a more pronounced effect on the higher resolution reflections. These effects have made it impossible for the statistics the highest resolution shells to become very favorable on an absolute scale, while the modeling of the virus in the lowest resolution shells has the potential to become quite accurate.

For example, when the statistics for the automatically selected models are examined in detail by tabulating them in bins by resolution, as in Table 1, it is clear that minimization of an overall residual has produced particularly good agreement with the standard in the lowest resolution ranges (*e.g.*, 0.03 to 0.13 in the innermost bin), while statistical agreement in the highest resolution ranges is much poorer (*e.g.*, 0.58 to 0.74 in the outermost bin), though in every case, the level of agreement is considerably better than random (which would be 1.00 for both the Q_F and Q_I cost functions). A very similar distribution of scores was seen even when the averaging process was initiated with perfect phases (data not shown), indicating that the high values of the fitness scores were not merely consequences of poor phase choices. Consistent with this pattern, a very wide range of Q_{IF} values was seen in the lowest resolution shells prior to refinement, but only a much narrower range of scores was possible at higher resolution.

Unfortunately, these artifacts of the modeling process made it more difficult to identify and select against poorly phased trials. The availability of a greater dynamic range at very low resolution artificially increased the influence of the lower resolution data on the GA and on the refinement process, beyond the influence they would have exerted if the model had been less coarse, or if the Fourier series had not been truncated. With an overall residual, this built-in bias is unavoidable because any model coarse enough for efficient survey can provide only an approximate description of the higher resolution information.

Fortunately, however, the simple decision to focus on narrower ranges of resolution addresses both of the major shortcomings of the overall unweighted refinement residual, both by making the residual more sensitive to the presence of phase cross-over regions, and by providing a way to correct for the unintentional resolution-dependent weighting of the procedure. By regarding every bin as an equally important contributor to the selection criterion and scoring, it was possible to correct automatically for the biased treatment of reflections without explicitly needing to understand or to model exactly how that bias behaves as a function of resolution. Using that approach, any built-in statistical error in the model can be tolerated, provided that the bin-specific score of each trial depends only on how the trial ranks relative to the population of other trials subjected to similar constraints and pressures. The use of relative ranking *per se* as a fitness criterion in GA's already has several precedents in computer science (see Srinivas & Patnaik, 1994).

4.2. *Implications of the choice of residual for NCS refinement*

In summary, the picture which emerges from these and previous experiments is of broad bands of consistently phased reflections with low Q_F values, separated by narrower bands of inconsistent ones with higher associated Q_F scores. The simple minimization of an overall unweighted resolution-independent residual usually is unable to eliminate most of the phase cross-over points, as such an improvement would require worsening the agreement in some of the extensive self-consistent regions as they moved toward global agreement.

Curiously, the approaches to NCS refinement adopted by Rayment (1983) and by Arnold & Rossmann (1988) were diametrically different from the approaches found to be effective here. Thus, Rayment down-weighted the influence on the refinement of individual reflections for which $|F_{\rm obsd}|$ and $|F_{\rm calc}|$ disagreed, using the empirical weight $\omega_{\rm exp} = \exp(-|\Delta|F||/|F_{\rm obsd}|)$, while Arnold & Rossmann achieved a similar down-weighting using the geometric mean between $\omega_{\rm exp}$ and Sim's weight (Sim, 1960). This approach was believed to facilitate refinement by making it easier for reflections with good statistical agreement (presumed an indicator of phase correctness) to overcome the influence of reflections with poorer statistical agreement. No doubt such an approach would be helpful in direct phase extension if the vast majority of reflections with low $|\Delta|F||$ were phased in a self-consistent way. In contrast, the approach developed here involves focusing on the statistical behavior of resolution shells, rather than on individual reflections; and it involves increasing, rather than decreasing, the influence of the data with poorest agreement. This latter approach may be necessary when the phase set already has been corrupted by the inclusion of mutually inconsistent regions; when the aim is to identify problematic phase sets rather than repairing them; and when the refinement method is (like the GA) capable of taking steps which are uphill relative to the local gradient to obtain global improvements in the parameter set.

4.3. *Future directions*

Ultimately, the goal of this work is to find ways to determine high-resolution crystal structures with no previously determined known homologue and without depending on the availability of isomorphous heavy-atom derivatives.

Experimental techniques for the collection of extremely low resolution data from virus crystals are now being considered. Careful collection of the low-resolution data would require specific strategies to compensate for the large and rapidly varying Lorentz correction, including the use of an atypically large oscillation range and the positioning of each reflection as far from the projected spindle axis as possible. How well the explicit direct-space model of the virus could tolerate measurement errors and incomplete sampling of the transform has yet to be determined, though it certainly would be desirable to have collected all of the data. If the current method should prove insufficient when using authentic virus data, additional constraints are available, both in direct and reciprocal space, which could straightforwardly be incorporated into the penalty function of the GA. Once the experimental techniques have been satisfactorily worked out, calculations with experimentally observed structure-factor amplitudes should be forthcoming.

Once the initial, low-resolution phase estimates have been obtained, and initial envelopes created, either by the *ab initio* survey methods presented above, or by means of cryoelectron microscopy, the task still remains of propagating phase and envelope information from a very low resolution (such as 24 Å) to fairly high resolution (say 3 Å). Direct NCS-based phase extension over such a wide resolution range has not yet been accomplished in practice, though it is expected to be possible. Some of the results presented above (particularly concerning the suppression of series termination errors) are pertinent to the phase-extension process, and suggest ways in which that might be accomplished.

Used together, the coarse tetrahedral lattice parameterization of trial structures and the matrix formulation for the structure-factor calculation are considerably more efficient than iterated direct-space averaging, as long as the resolution remains low. The opposite will become true as the low-resolution phases determined by the present approach are extended to higher resolution. In either approach, it is anticipated that information about the shape of the molecule from previous lower resolution calculations could be used to create increasingly more detailed envelopes. In the coarse lattice model, the envelope would specify the locations of whatever number of lattice points (**m**) were permitted by the resolution of the data used. Reducing the volume of the envelope permits a finer sampling of the protein region, and a more accurate modeling of the virus, without requiring a corresponding increase in the number of Fourier terms. Recognizing that the directly generated phase information is more reliable at the lower resolution end of the current resolution sphere, as shown in Fig. 4(*c*), propagation of information *via* increasingly detailed envelopes could furnish an appropriately cautious approach to direct phase extension. Eventually, at higher resolution, after a switch is made to the Bricogne (1976) direct-space averaging method, it would seem a sensible precaution to initiate averaging with only the lower resolution portion of the transform specified by the coarse model, and to utilize the envelope principally to distinguish grid points whose values could be modified by averaging from those whose values were to be preserved intact. With these precautions, it should be possible to successfully extend the solutions to high resolution.

The authors thank M. Wien for critical reading of the manuscript, and members of the Hogle and T. Ellenberger laboratories for useful discussions. This work was supported by NIH grant AI20566 (to JMH), and by a grant from the Keck Foundation to establish a structural biology facility in the Department of Biological Chemistry and Molecular Pharmacology. STM was supported in part by Biophysics training grant TM#T32-GM08313.

References

Argos, P., Ford, G. C. & Rossman, M. G. (1975). *Acta Cryst.* A**31**, 499-506.

Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270-282.

Basavappa, R., Syed, R., Flore, O., Icenogle, J. P., Filman, D. J. & Hogle, J. M. (1994). *Protein Sci.* **3**, 1651-1669.

Bricogne, G. (1974). *Acta Cryst.* A**30**, 395-405.

Bricogne, G. (1976). *Acta Cryst.* A**32**, 832-847.

Chang, G. & Lewis, M. (1994). *Acta Cryst.* D**50**, 667-674.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Acta Cryst.* A**48**, 301-312.

Cheng, R. H., Reddy, V. S., Olson, N. H., Fisher, A. J., Baker, T. S. & Johnson, J. E. (1994). *Structure*, **2**, 271-282.

Crowther, R. A. (1967). *Acta Cryst.* **22**, 758-764.

Crowther, R. A. (1969). *Acta Cryst.* B**25**, 2571-2580.

Crowther, R. A. (1971). *Philos. Trans. R. Soc. London Ser. B*, **261**, 221-230.

Davis, L. (1991). *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold.

Finch, J. T. & Holmes, K. C. (1967). *Methods in Virology*, Vol. 3, pp. 351-474. New York: Academic Press.

Forrest, S. (1993). *Science*, **261**, 872-878.

Gaykema, W. P. J., Volbeda, A. & Hol, W. G. J. (1985). *J. Mol. Biol.* **187**, 255-275.

Hogle, J. M., Chow, M. & Filman, D. J. (1985). *Science*, **229**, 1358-1365.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.

Jack, A. & Harrison, S. C. (1975). *J. Mol. Biol.* **99**, 15-25.

Jack, A., Harrison, S. C. & Crowther, R. A. (1975). *J. Mol. Biol.* **97**, 163-172.

Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. (1983). *Science*, **220**, 671-680.

McKenna, R., Xia, D., Willingmann, P., Ilag, L. & Rossmann, M. G. (1992). *Acta Cryst.* B**48**, 499-511.

Main, P. & Rossmann, M. G. (1966). *Acta Cryst.* **21**, 67-72.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087-1092.

Rayment, I. (1983). *Acta Cryst.* A**39**, 102-116.

Rayment, I., Baker, T. S., Caspar, D. L. D. & Murakami, W. T. (1982). *Nature (London)*, **295**, 110-115.

Rossmann, M. G., Arnold, E. Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Nature (London)*, **317**, 145-153.

Schraudolph, N. N. & Grefenstette, J. J. (1992). *A User's Guide to GAucsd* 1.4. University of California San Diego, La Jolla, CA, USA.

Sim, G. A. (1960). *Acta Cryst.* **13**, 511-512.

Srinivas, M. & Patnaik, L. M. (1994). *Computer*, **27**, 17-26.

Stewart, P. L., Fuller, S. D. & Burnett, R. M. (1993). *EMBO J.* **12**, 2589-2599.

Tong, L. & Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 783-792.

Tsao, J., Chapman, M. S. & Rossmann, M. G. (1992). *Acta Cryst.* A**48**, 293-301.

Tsao, J., Chapman, M. S., Wu, H., Agbandje, M., Keller, W. & Rossmann, M. G. (1992). *Acta Cryst.* B**48**, 75-88.

Unge, T., Liljas, L., Strandberg, B., Vaara, I., Kannan, K. K., Fridborg, K., Nordman, C. E. & Lentz, T. J. (1980). *Nature (London)*, **285**, 373-377.

Valegård, K., Liljas, L., Fridborg, K. & Unge, T. (1990). *Nature (London)*, **345**, 36-41.