Acta Cryst. (1993). D**49**, 429–439

PRISM: Topologically Constrained Phase Refinement for Macromolecular Crystallography

BY DAVID BAKER,* CHRISTOPHER BYSTROFF,†‡ ROBERT J. FLETTERICK† AND DAVID A. AGARD*§

Howard Hughes Medical Institute and the Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143-0448, *USA*

(*Received* 2 *October* 1992; *accepted* 21 *April* 1993)

Abstract

We describe the further development of phase refinement by iterative skeletonization (*PRISM*), a recently introduced phase-refinement strategy [Wilson & Agard (1993). *Acta Cryst.* A**49**, 97–104] which makes use of the information that proteins consist of connected linear chains of atoms. An initial electron-density map is generated with inaccurate phases derived from a partial structure or from isomorphous replacement. A linear connected skeleton is then constructed from the map using a modified version of Greer's algorithm [Greer (1985). *Methods Enzymol.* **115**, 206–226] and a new map is created from the skeleton. This 'skeletonized' map is Fourier transformed to obtained new phases, which are combined with any starting-phase information and the experimental structure-factor amplitudes to produce a new map. The procedure is iterated until convergence is reached. In this paper significant improvements to the method are described as is a challenging molecular-replacement test case in which initial phases are calculated from a model containing only one third of the atoms of the intact protein. Application of the skeletonization procedure yields an easily interpretable map. In contrast, application of solvent flattening does not significantly improve the starting map. The iterative skeletonization procedure performs well in the presence of random noise and missing data, but requires Fourier data to at least 3.0 Å. The constraints of linearity and connectedness prove strong enough to restore not only missing phase information, but also missing amplitudes. This enables the use of a powerful statistical test, analogous to the 'free *R* factor' of conventional refinement [Brünger (1992). *Nature* (*London*), **355**, 472–474], for optimizing the performance of the skeletonization procedure. In the accompanying paper, we describe the application of the method to the solution of the structure of the protease inhibitor ecotin bound to trypsin and to a single isomorphous replacement problem.

Introduction

The crystallographic phase problem remains an impediment to the rapid solution of three-dimensional macromolecular structures. Unfortunately, there is insufficient information in the structure-factor amplitudes alone to determine directly high-resolution phases (for a recent discussion, see Baker, Krukowski & Agard, 1993). Thus, much effort has been directed at practical methods of phase improvement. At different stages in the solution of a molecular structure from diffraction data, varying amounts of physical and chemical information can be utilized to improve the phases. Classical direct methods, which have been used to solve countless small-molecule structures *ab initio*, make use only of positivity and atomicity. Solvent flattening (Wang, 1985), widely used to improve electron-density maps of macromolecules generated using experimental phase information, makes use only of the information that solvent regions in protein crystals are relatively featureless and that the electron density is positive. The full power of chemical knowledge can only be brought into play after an atomic model has been traced through the density. Then, the large store of *a priori* information pertaining to molecular geometry (bond lengths, bond angles and non-bonded interaction energies) can be used to reduce greatly the number of free parameters.

There is a considerable gap between employing solvent flattening and full stereochemical refinement. The information that proteins are composed of linear connected strings of atoms lies somewhere between the knowledge that solvent regions are featureless and the detailed rules of stereochemistry; it is more stringent than the former but does not require the atomic model implicit in the latter.

A previous report (Wilson & Agard, 1993) described a refinement method which exploits the information that proteins are connected linear chains

* Howard Hughes Medical Institute and the Department of Biochemistry and Biophysics.

† Department of Biochemistry and Biophysics.

‡ Current address: Universidad Nacional de Ingeniería, Managua, Nicaragua.

§ Author to whom correspondence should be addressed.

ISSN 0907-4449

through iterative skeletonization of the electron density [an alternate implementation of these constraints is described in Bhat & Blow (1982)]. A linear connected skeleton was generated from an input map in two steps. First, a list of nodes corresponding to local maxima along the x, y, or z axes in the map, and a list of connections between the nodes, were generated using routines in the *GRINCH* (Williams, 1982) package. Second, new nodes were added to link isolated sets of mutually connected nodes to create a completely connected skeleton. Structure factors were then calculated from the skeleton using a carbon-scattering curve for each node. A new electron-density map was calculated using the new phases and Sim-weighted Fourier coefficients, and the skeletonization procedure was repeated. The radius of convergence of this iterative skeletonization procedure was dramatically larger than that of solvent flattening for a variety of simple molecular-replacement test cases.

Here we describe the further development of this phase-refinement strategy. Significantly more accurate methods are used to skeletonize the input electron-density map and to obtain structure factors from the skeleton. The connectivity of the chain is preserved throughout the skeletonization procedure, thus eliminating the problem of reconnecting disjointed graphs, which grows exponentially with the number of nodes. In this paper the new method is described in detail and is applied to a challenging molecular-replacement test case. In the accompanying paper, we describe the application of the method to the solution of a new protein structure and to the improvement of phases in an otherwise intractable single isomorphous replacement (SIR) problem.

Methods

All calculations were done on a VAX/VMS 9000 or a VAX/VMS 8650 computer. Programs used in this study originated either from the *CCP*4 package of crystallographic programs (*FFT*, *GENSFC*, *SFC*, *LCFUTILS*, *COMBINE*, *ENVELOPE*, *HKLWEIGHT*, *TRUNCMAP*) (SERC Daresbury Laboratory, 1986) or were written by the authors (*SKELETON*, *SIMWT*, *MAPCORREL*, *MAKENV*, *LCFSCALE*, *PURGELCF*, *FREE*). The latter programs were written in C or Fortran and are available upon request. Minor modifications were made in the *CCP*4 programs *TRUNCMAP* and *ENVELOPE* to facilitate interactions with the other routines. Maps and structure-factor data sets were expressed in standard *CCP* 'MAP' and 'LCF' formats, respectively. Coordinates of models and of skeletons were stored in standard Protein Data Bank (PDB; Bernstein *et al.*, 1977) format. The programs *FRODO* (Jones, 1985) and *INSIGHT*II (Biosym Technologies, 1991) were used on either an Evans and Sutherland PS390 or a Silicon Graphics Iris 4D/25 to visualize results. Simple conversion programs (available from the authors) were used to put skeletons or maps into formats compatible with these modeling programs.

The flowchart in Fig. 1 describes the cycle of programs employed in the *PRISM* method. An initial map, which may not be interpretable, is skeletonized and a new map is output. Structure factors are calculated from the skeletonized map and are then scaled to observed amplitudes. For isomorphous replacement problems, phases from the skeletonized map are combined with experimental phase-probability distributions. For molecular-replacement problems, Fourier coefficients $2wF_o - F_c$ (Main, 1979) for acentric reflections and wF_o for centrics are calculated where 'w' is the Sim weight (Sim, 1960). A new map is generated using these Fourier coefficients and input into the next cycle. Solvent flattening with or without non-crystallographic symmetry averaging can be included in the cycle as indicated in the figure.

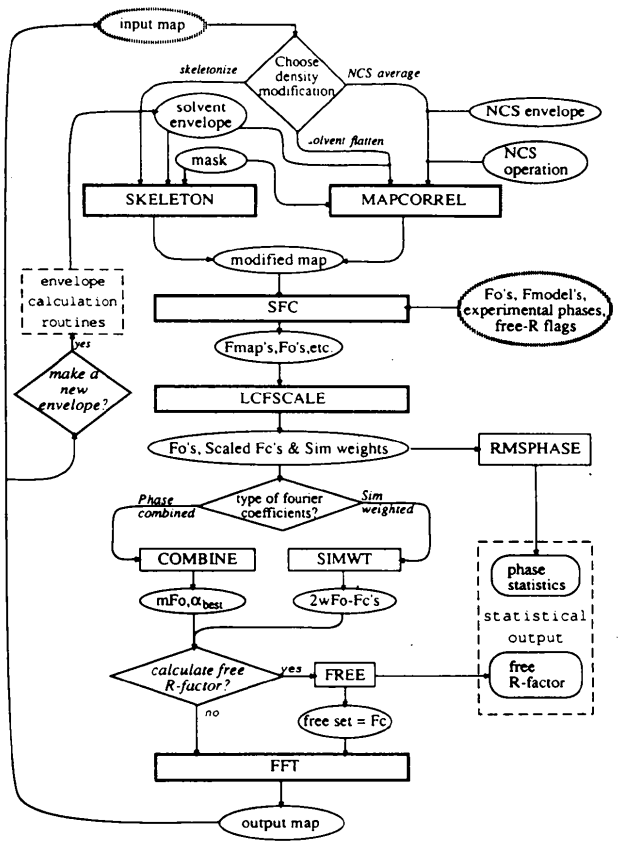

Fig. 1. Flowchart of the cycle of programs employed in the *PRISM* method.

Descriptions of the individual programs

SKELETON employs a modified version of Greer's algorithm (Greer, 1985) to reduce an electron-density map to a set of connected grid points, or 'nodes'. Based on these nodes, a 'skeletonized' map is then generated with the density falling off smoothly with distance from the skeleton.

An electron-density map is read in, and all grid points above a user-defined minimum density (minden) are stored as 'nodes' together with a list of nearest neighbors. Nodes at the six grid points ($x \pm 1$, y, z), (x, $y \pm 1$, z) and (x, y, $z \pm 1$) are considered nearest neighbors of a node at (x, y, z). Two nodes are considered to be connected if there is a continuous path of nearest neighbors between them. A 'graph' is a set of mutually connected nodes. A 'cut-point' is a node the removal of which disconnects a graph. An 'end-point' is a node with a single neighbor.

Nodes are then removed, starting from the lowest density, unless they are either cut-points, or are end-points and have a density greater than a user-defined value (epden). Once a user-defined maximum density (maxden) is reached, no further nodes are removed. Graphs containing less than a certain user-defined number of nodes (mingraph) are then removed. Because the nodes are more closely spaced than atoms along a peptide chain, a single overall thermal B factor is insufficient to shape properly the electron-density distribution. To accomplish this, a new map is generated from the selected set of nodes using the formula $\rho = \exp[-(\pi^2/\beta)r^2]$ where r is the distance from the nearest node and β is user-defined and acts much like an anisotropic B factor. Instead of using a uniform peak height for the skeleton, using the original electron densities at the node points as weighting factors was examined, but did not improve the performance of the method. Throughout the paper, the removal of nodes not required for the connectivity of the chain is referred to as thinning, and the removal of small disconnected graphs, as pruning.

As the skeleton, and hence the highest peaks in the skeletonized map, are restricted to grid points, optimal results are obtained when the input map is calculated on a finely sampled grid. The experiments described here utilized a grid spacing of $\frac{1}{4}$ the resolution limit unless otherwise indicated.

LCFSCALE scales calculated amplitudes (F_{calc}) to observed amplitudes (F_{obs}) and calculates Sim weights. Zonal scaling was used for all of the work described here. In molecular-replacement cases where the scattering density in a portion of the asymmetric unit was known, input maps were calculated with F_{000} equal to zero, and a mask was used to zero the known portion prior to density modification. The transform (F_{known}) of the masked portion of the map (usually structure factors from a molecular-replacement model) was then combined through vector addition with the transform (F_{skel}) of the skeletonized portion of the map to give the output F_{calc}'s

$$F_{calc} = a_1 F_{known} + a_2 F_{skel},$$

where the scale factors, a_1 and a_2, were obtained for each resolution zone by minimizing $\Sigma(F_{obs}^2 - |a_1 F_{known} + a_2 F_{skel}|^2)^2$. Performing the scaling in reciprocal space rather than real space has the advantage that the intrinsic differences between a skeleton and an atomic model can be partially compensated for by resolution-dependent scale factors.

MAPCORREL performs real-space non-crystallographic symmetry averaging for grid points within an envelope. A molecular envelope can be created either using the filtering approach of Wang (Wang, 1985) or from the skeleton coordinates using *MAKENV*. When there is non-crystallographic symmetry (NCS), related grid points within the envelope are averaged and grid points outside the envelope are set to their mean value. In the absence of an input NCS operator, the program reads a solvent envelope, flattens the solvent region and optionally truncates negative density in the protein region.

A single 'master' VMS command language file is available from the authors which runs the entire *PRISM* package, performing any combination of NCS averaging, solvent flattening and skeletonization. Fig. 1 shows the programs included in the package and the order in which they run. The input to the command file is a text file containing keywords and values defining the run.

Results and discussion

The phase-refinement strategy is described in Fig. 1. As in traditional density-modification methods, an input map is modified by the application of constraints, in this case connectivity and linearity, and new structure factors are calculated from the modified map. The new phases are then combined with the experimental amplitudes and a new map is generated. The process is repeated until convergence is reached. The radius of convergence of such a method is determined by the power of the applied constraints. The accuracy is limited by the errors which necessarily accompany the enforcement of the constraints.

Accuracy of skeletonization

To test the fidelity of skeletonization, a perfect map was subjected to multiple rounds of the iterative skeletonization procedure. Structure factors and a

perfect starting map were generated from the coordinates (1LPE) of apolipoprotein E (Wilson, Wardell, Weisgraber, Mahley & Agard, 1991). The map was then subjected to the iterative skeletonization procedure, using the amplitudes calculated from the coordinates as the F_{obs}. The weighted phase error,

$$\Sigma F_{obs}|\alpha_{calc} - \alpha_{true}|/\Sigma F_{obs},$$

at each cycle is shown in Fig. 2 (triangles). The phase error rapidly increased from 0 to 12° and then leveled out at about 18°. A similar test of the earlier skeletonization procedure gave a final phase error of approximately 27° (Wilson & Agard, 1993); thus the new algorithm is considerably more accurate than the algorithm used previously.

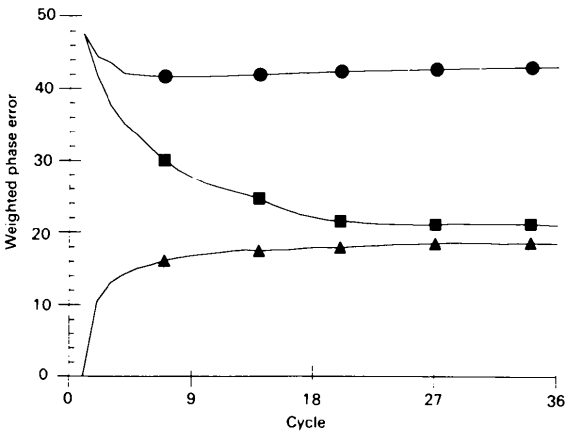

Fig. 2. Progress of phase refinement. Structure factors (F_{obs}, α_{true}) and a 'perfect' electron-density map were calculated at 2.5 Å resolution from the coordinates of apolipoprotein E (1LPE) translated into a $P1$ unit cell with $a = 28$, $b = 34$ and $c = 66$ Å. As illustrated in the accompanying paper, the procedure is space-group general; space group $P1$ was chosen for the test case to simplify visualization of the skeletons and maps. The iterative skeletonization procedure was then applied to the perfect starting map. The grid spacing was 0.67 Å. The values for the parameters minden, epden, maxden, β and mingraph were 1.2, 2.0, 2.5, 6.0 and 15, respectively. At each cycle a new map was calculated using Fourier coefficients $(2wF_o - F_c)\exp(i\alpha_{calc})$ where the α_{calc} are phases calculated from the skeletonized map generated in the previous cycle. The weighted phase difference $(\alpha_{calc} - \alpha_{true})$ is shown in the figure (line with triangles). In a second experiment, structure factors (F_{calc}, α_{calc}) were generated from a molecular-replacement model lacking residues 39–59 and 107–144 and side-chain atoms beyond C^β. The iterative skeletonization procedure was then applied to a map calculated using the amplitudes from the intact structure and phases from the partial model. The skeletonization parameters were identical to those listed above. The weighted phase error over 36 cycles of refinement is shown in the figure (squares). In a third experiment the F_{obs}, α_{calc} starting map described in the previous experiment was subjected to the standard solvent-flattening protocol. Envelopes were calculated using Leslie's reciprocal-space adaptation of Wang's algorithm with the solvent content set at 60% [using the formula volume (Å^3) = 1.3 × molecular weight (D), the fraction of solvent in the unit cell was calculated to be 65%]. The envelope was recalculated every two cycles. The weighted phase error over 36 cycles of solvent flattening is shown in the figure (circles). Very similar results were obtained when envelopes were calculated using a solvent content of 55%. For clarity, every eighth data point is represented by a symbol in the figure.

Table 1. *Skeletonization starting with molecular-replacement model*

The table shows the course of skeletonization during the first three cycles of refinement (Fig. 2, squares) of the phases from the molecular-replacement model. The skeletonization procedure may be divided into three stages. In the first stage, nodes are placed at all grid points above minden. In this case minden was 1.2 standard deviations above the mean, and nodes were placed at approximately 20000 (column 2) of the 200000 grid points of the map. In the second stage, nodes are considered in order of increasing density and are removed unless this would disrupt the connectivity of the chain. The tips of the skeleton are protected if they are above epden, here 2.0 standard deviations above the mean. Nodes above maxden, here 2.5 standard deviations above the mean, are also not removed. This thinning step reduced the number of nodes roughly threefold (column 3). In the final step, small graphs containing less than mingraph (here 15) nodes are removed. Approximately 100 nodes were removed in this pruning step (column 4). A convenient measure of the connectivity of the final skeleton is the fraction of remaining nodes which are in the largest graph. Using this measure, the connectivity of the skeletons increased from 64 in the first cycle to 98.5% in the third cycle (column 5).

		Number of nodes		
Cycle	Start	After thinning	After pruning	Connectivity
1	20121	6042	5921	64
2	20881	8105	7978	98
3	20829	8863	8762	98.5

Application of *PRISM* to a molecular-replacement test case

The iterative skeletonization procedure was applied to a challenging molecular-replacement (MR) test case using a starting model containing only $\frac{1}{3}$ of the atoms in the native protein. The N-terminal domain of apolipoprotein E is an elongated four-helix bundle of 144 residues and 1172 non-H atoms (Wilson *et al.*, 1991). The starting model was generated by truncating the four-helix bundle and removing all side chains beyond the C^β, leaving 85 residues and 421 atoms. A similar short polyalanine four-helix bundle was used originally in an attempt to solve the structure of apolipoprotein E through molecular replacement (Wilson *et al.*, 1991). A starting map was calculated using amplitudes from the intact protein and phases from the truncated MR model. The map was then subjected to 36 cycles of the iterative skeletonization procedure. The weighted phase error dropped from an initial value of 47.6° to 21.0° after 36 cycles (Fig. 2, squares). Comparison with the results of skeletonizing a perfect map (Fig. 2, triangles) shows that the final error is close to the limit set by the intrinsic errors of the skeletonization procedure.

The details of skeletonization during the first three cycles of refinement are described in Table 1. The constraints of chain connectivity and linearity imposed on the density during the skeletonization procedure effectively reduce the number of parameters from the number of grid points in the map (as

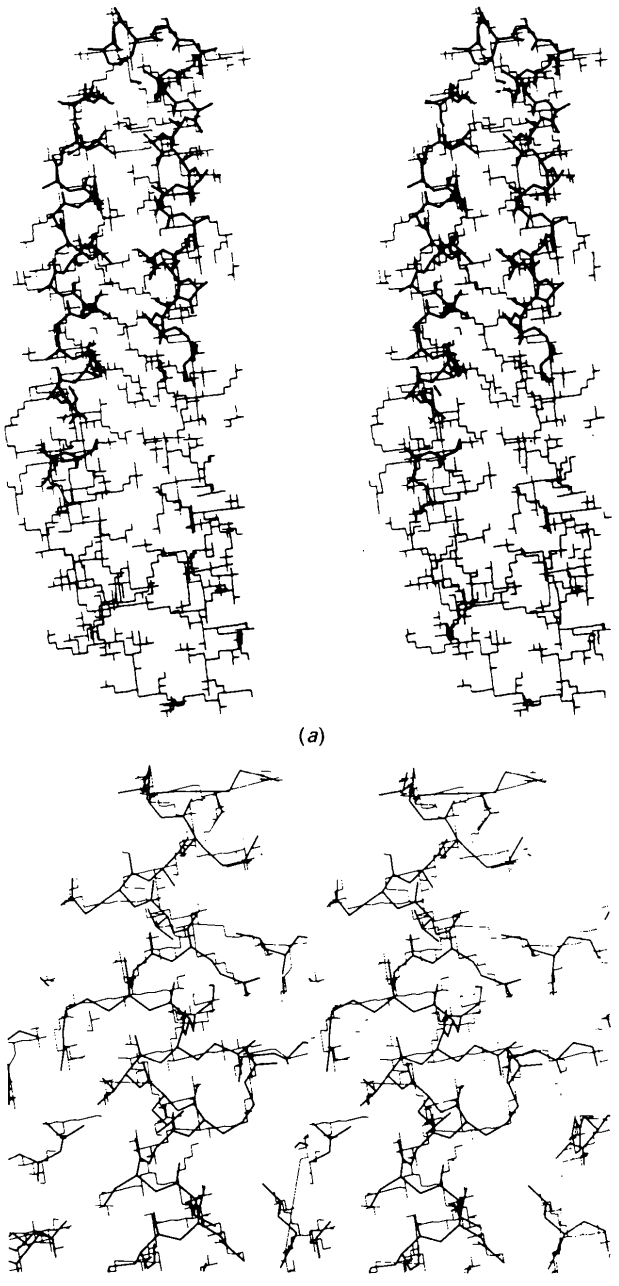

Fig. 3. Comparison of the final skeleton to the starting molecular-replacement model and to the intact structure. The *SKELETON* program optionally outputs the coordinates of the skeleton generated from the input map. The coordinates of the skeleton generated in the last cycle of the molecular-replacement test case (Fig. 2, squares) was converted to an *INSIGHT* user file with the program *PDB2SKEL* (available from the authors). (*a*) is a superposition of the final skeleton (thin lines) and the starting model (thick lines). Although skeletons obtained when diagonal edges are permitted are slightly more aesthetically pleasing, this did not improve the performance of the method. The skeleton in the region omitted from the starting model (the lower half of the figure) almost perfectly follows the coordinates of intact apolipoprotein E, as shown in greater detail in (*b*) (thin lines, skeleton; thick lines, 1LPE).

in conventional density-modification methods) to the number of nodes in the skeleton.

Fig. 3(*a*) shows a portion of the final skeleton generated after 36 cycles of iterative skeletonization starting with the molecular-replacement model (Fig. 2, squares). The coordinates of the starting model are also shown for comparison. The half of the four-helix bundle truncated in the starting model is regenerated in the skeleton as are most of the missing side chains. A more detailed view of the final skeleton in a region omitted in the starting model is shown in Fig. 3(*b*). The true solution – the coordinates of apolipoprotein E – is also shown. The skeleton accurately follows both the helical main-chain and the side-chain residues of the true structure, despite the fact that neither were included in the starting model.

Comparison to solvent flattening

The most widespread and successful density-modification procedure applied to the macromolecular diffraction problems is solvent flattening (Wang, 1985). The relative power of solvent flattening and skeletonization for this type of molecular-replacement problem was investigated by running the standard solvent-flattening protocol starting with phases from the molecular-replacement model, recalculating the envelope every two cycles. As shown in Fig. 2, solvent flattening (circles) led to only a slight reduction of the phase error, and was dramatically outperformed by the iterative skeletonization procedure (squares).

The map produced by the skeletonization procedure, the final solvent-flattened map and the starting molecular-replacement map are compared in Fig. 4. The figure also shows the corresponding portion of the correct apoE coordinates. This region of the protein was not included in the starting model. As expected for such a large deletion, the starting map calculated with phases from the molecular-replacement model was quite disconnected (Fig. 4*a*). Solvent flattening somewhat reduced the noise in the map, but the density was still disconnected and essentially uninterpretable (Fig. 4*b*). However, as witnessed by the large drop in phase error shown in Fig. 2, the iterative skeletonization procedure led to a dramatically improved map (Fig. 4*c*). The electron density in the map produced by the skeletonization procedure was readily interpretable, in fact it almost perfectly followed the coordinates of the native structure.

Skeletonization restores missing amplitudes

The constraints of connectivity and linearity imposed by the skeletonization procedure effectively couple

the structure factors. The improvement in the phases during iterative skeletonization (Fig. 2) is a result of this coupling. The interactions among structure factors are potentially strong enough to restore not only missing phase information but also a limited amount of missing amplitude information.

As described in the methods section, the skeletonization algorithm requires several user-defined parameters. The optimal values of these parameters may vary from map to map. An objective criteria is needed to determine the optimal values of the parameters for each individual problem and in general to follow the progress of the phase refinement. In test cases, the crystallographic R factor ($\Sigma|F_{obs} - F_{calc}|/\Sigma F_{obs}$) was found to correlate poorly with the phase error. This is not surprising since at the limit of no density modification an R factor of 0.0 would be obtained (the F_{obs}'s are used in each round of map calculation) independent of actual phase error.

To investigate the possible use of a 'free R factor' (Brünger, 1992), determined for a fraction of the data not included in the map calculation, in assessing the progress of phase refinement, a perfect map was calculated with about 5% (chosen at random) of the reflections initially set to zero. The iterative skeletonization procedure was then applied, and after each cycle of density modification the output F_{calc}, rather than F_{obs}, was used in the map calculation for the 'free' set of reflections. As in Fig. 2, the phase error increased steadily over the first five cycles because of errors in the skeletonization procedure (Fig. 5, squares). The total R factor increased during the first cycle of skeletonization, and then dropped and ultimately leveled off, correlating poorly with the phase error (Fig. 5, triangles). Importantly, the R factor for the free set of reflections dropped from 100% (the F_{calc}'s for the free set are 0 prior to density modification) at cycle 0 to 23% after the first two cycles of skeletonization (Fig. 5, circles). Thus, the constraints of connectivity and linearity on the density define a complex interaction amongst the structure factors that effectively couples the missing reflections to the known amplitudes. The missing reflections recover from initial values of 0 to near to their true values. In contrast to the overall R factor, the free R factor rose after the second cycle, paralleling the phase error.

Effect of missing data

The sensitivity of the skeletonization procedure to the completeness of the data set is important both because real data sets are seldom 100% complete and because use of a free R factor for assessing the progress of the procedure necessitates the setting

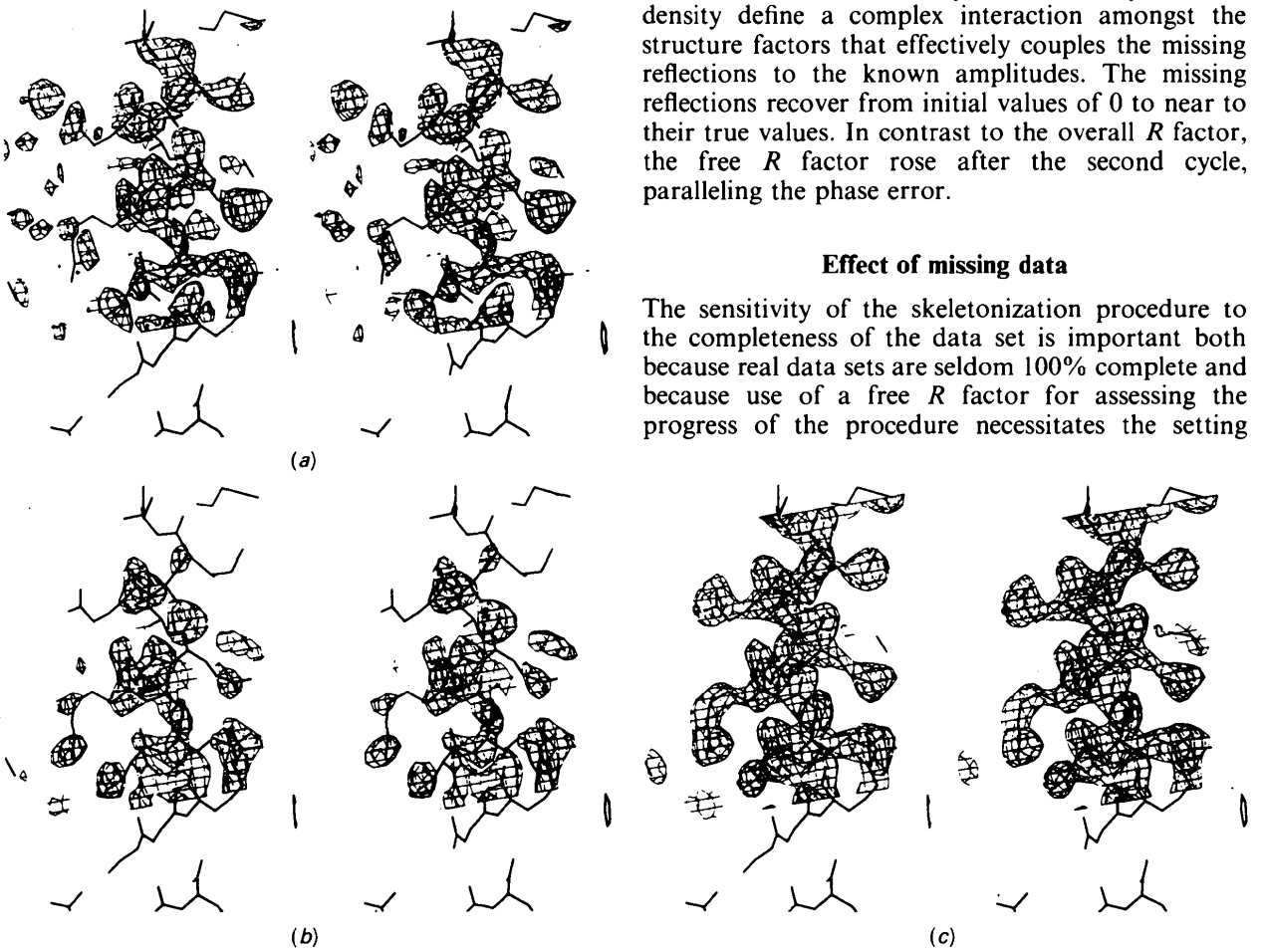

Fig. 4. Comparison of the maps produced by skeletonized and solvent flattening to the starting molecular-replacement map. Electron-density maps were displayed using the graphics program *FRODO*. A representative portion of the starting map (*a*), the map after solvent flattening (*b*), and the map after iterative skeletonization (*c*) are shown together with the corresponding region of the 1LPE PDB coordinates. This region was not included in the molecular-replacement starting model.

aside of a small fraction of the data. The effect of missing diffraction data was tested by randomly omitting 5% of the reflections and repeating the skeletonization procedure starting with phases from the molecular-replacement model. As shown in Fig. 6 (open circles), omitting data reduced both the rate of convergence and the accuracy of the final map, but there was still a dramatic drop in phase error. As the coupling among the diffraction amplitudes leads to partial restoration of missing data (Fig. 5), better results might be obtained by using $|F_{calc}|$'s from the latest skeleton rather than 0's for the missing reflections during map calculations after the first cycle. As shown in Fig. 6 (closed circles) this is indeed the case. Allowing the missing amplitudes to 'float' at the value calculated from the latest skeleton improved the performance of the method.

Low-intensity reflections are often missed during collection of diffraction data. To investigate the effect of missing low-intensity reflections, the molecular-replacement test case was repeated after omitting the weakest 17% of the reflections. As shown in Fig. 6 (triangles), omission of this data did not adversely affect the iterative skeletonization procedure. The weighted phase error in this case actually dropped lower than when all data were included (compare triangles to squares), presumably because the weighting does not completely compensate for the larger phase error associated with low-intensity reflections. Thus, a failure to measure a relatively large amount of low-intensity data is less disruptive to the skeletonization procedure than is a failure to collect higher intensity data in even a small section of reciprocal space.

Correlation of the free *R* factor with phase error

The resilience of the skeletonization procedure to the omission of a small amount of the diffraction data allows the use of a free *R* factor to monitor the progress of phase refinement. The ability of the free *R* factor to track the phase error was investigated for the molecular-replacement test case described above. An initial map was calculated using phases from the molecular-replacement model with 5% of the amplitudes (the 'free' set) set to zero. The iterative skeletonization procedure was then applied and the free *R* factor was calculated at each cycle. As shown in Fig. 7(*a*), there was an almost perfect correlation of the free *R* factor with the phase error.

The free *R* factor is thus a powerful tool for optimizing the method. As the skeletonization procedure is clearly sensitive to the degree of completeness of the data set, the free set must be as small as possible while retaining statistical significance.

The free *R* factor was also used to follow the progress of solvent flattening. 5% of the amplitudes were again set to zero and the solvent-flattening procedure described in the legend to Fig. 2 was

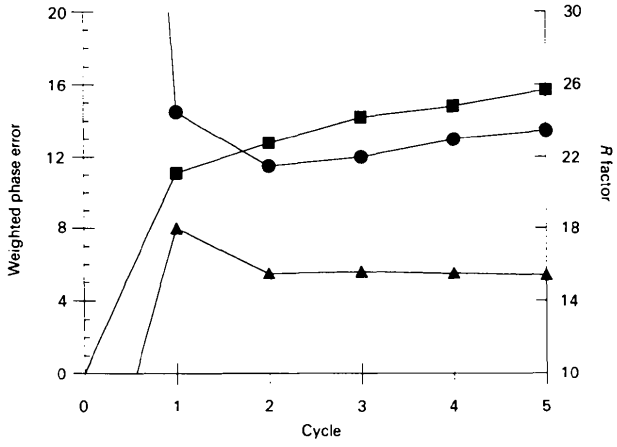

Fig. 5. Recovery of missing amplitudes during iterative skeletonization. A 'perfect' map was calculated from the apolipoprotein E PDB coordinates as described in the legend to Fig. 1 except that 5% of the Fourier data were randomly omitted. The map was then subjected to the iterative skeletonization procedure. The missing reflections were allowed to float as described in the text. The figure shows the weighted phase error (squares), the *R* factor for the reflections included in the map calculation (triangles) and the free *R* factor (circles) during the first five cycles of refinement. Cycle 0 corresponds to the perfect starting map – the *R* factor for the reflections included in the map calculation is 0% and the free *R* factor is 100% [$\Sigma(F_{true} - 0)/\Sigma F_{true}$] since the map has not yet been modified.

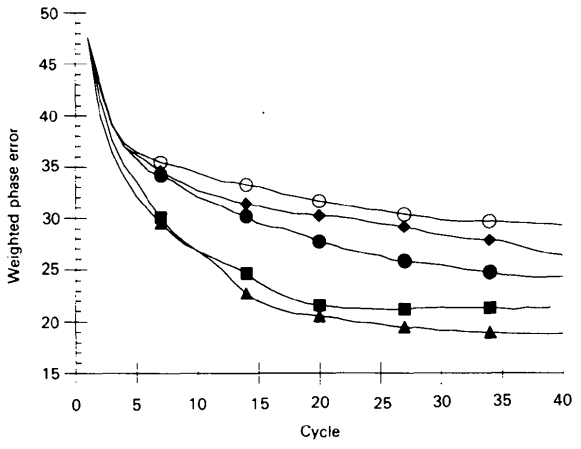

Fig. 6. Iterative skeletonization using incomplete and noisy data. The molecular-replacement problem described in Fig. 2 was repeated using incomplete error-containing Fourier data. The figure describes the results of omitting 5% of the reflections at random (circles), omitting the lowest intensity 17% of the reflections (triangles), and of omitting 5% of the reflections and also adding 10% random errors to the remaining reflections (diamonds). The results with complete and error-free data (Fig. 2) are replotted (squares) for comparison. The missing amplitudes were either replaced by F_{calc}'s from the latest skeletonized map (diamonds and closed circles) or were set to zero throughout the run (triangles and open circles).

repeated. The free R factor again correlated well with the phase error, dropping in the first five cycles and then very slowly increasing (Fig. 7*b*, circles). In contrast, the overall R factor fell continuously throughout the 40 cycles of refinement, reaching a final value of less than 10% (Fig. 7*b*, triangles). Not surprisingly, the overall R factor is thus a very poor measure of the success of solvent flattening. These results suggest that a free R factor calculation should be included in the standard solvent-flattening protocol.

Optimization of parameters using the free R factor

The parameters required by the skeletonization algorithm are minden, the minimum density for a node to be added at a grid point, maxden, the density above which nodes will not be removed, mingraph, the size of the smallest graph which will be retained after skeletonization, and β which determines the fall-off of density with distance from the skeleton in the output map. The molecular-replacement test case described above was repeated four times; in each run one of the parameters was allowed to vary and the others were held fixed. The phase error and the free R factor after the fourth cycle of refinement are plotted for each combination of parameters in Fig. 8. The correlation between phase error and free R factor was reasonable in all cases, thus the free R factor appears be an extremely useful target function for optimizing parameters in real cases (in which the phase error cannot be calculated). In some applications it may be useful to calculate the initial map with the free reflections set to the average values within their respective resolution shells instead of to zero. This would provide a more accurate initial guess for their true values and should result in an improved starting map.

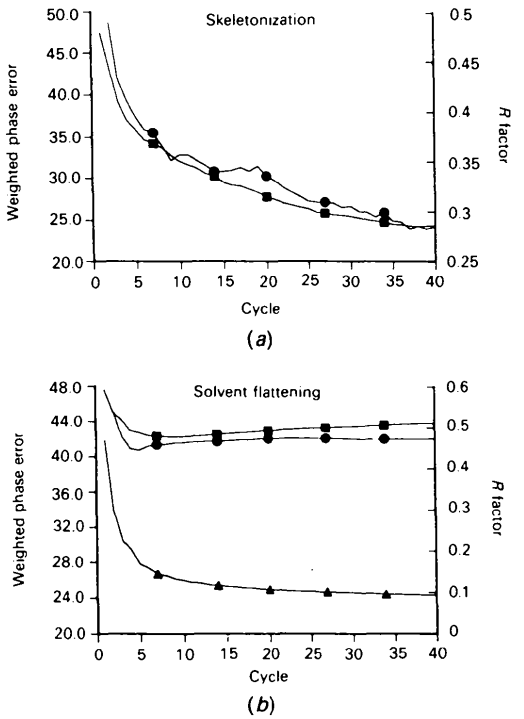

Fig. 7. Use of the free R factor to monitor (*a*) skeletonization and (*b*) solvent flattening. Skeletonization and solvent flattening were applied to the molecular-replacement problem as described in Fig. 2, except that 5% of the reflections were reserved for free R factor calculation. Squares, weighted phase error; circles, free R factor; triangles, overall R factor.

It is instructive to examine the dependence of the skeletonization procedure on the parameter values (Fig. 8). The lowest density at which a grid point can contribute to the skeleton is set by minden which ideally should be well above the background solvent density but below the weakest link in the protein chain. Too low a value of minden results in a completely connected skeleton protruding into the solvent region and with many incorrect connections, while too large a value leads to omission of parts of the protein from the final skeleton. The optimal value for minden in the rather diverse problems we have studied has been close to 1.2σ above the mean density in the map.

The extent of skeletonization is determined by maxden, an upper-density cutoff above which a node will not be removed even if it is not required to maintain connectivity of the skeleton. Too low a value of maxden overly limits the amount of density modification, while too large a value leads to the amplification of errors in the starting map. In practice, the optimal value of maxden has varied from 2.5 to 3.5σ above the mean.

The level of noise reduction is set by mingraph, the size of the smallest graph retained in the final skeleton. The noise in the map may be reduced by removing small disconnected graphs, but with mingraph too large, not-yet-connected side chains and backbone fragments may be incorrectly pruned. The optimal value of mingraph has been found to be between 4 and 15 nodes.

The parameter β, essentially an anisotropic B factor, determines the fall-off in electron density with distance from the skeleton. Initially, pseudo atoms (corresponding to a β of 0) were inserted at each point in the skeleton, but much better performance was obtained with a non-zero β. This is probably because accurate skeletonization requires a grid spacing considerably less (about $\frac{1}{4}$ the diffraction limit) than the carbon–carbon bond distance: converting a linear piece of the skeleton into density by placing C atoms at 0.6 Å intervals results in much too sharp a fall-off in the direction perpendicular to the skeleton. At the other extreme, too large a value of β blurs out the skeleton entirely.

Fortunately, the optimal values of the parameters have not varied substantially in the quite diverse problems we have studied: the optimal values of minden, maxden, β and mingraph range between 1.0 and 1.4, 2.5 and 3.5, 6.0 and 10.0, and 4 and 15, respectively (minden and maxden are in σ above the mean, β is in Å^2, and mingraph is in number of nodes).

Use of a solvent envelope

It is apparent from the above considerations that the optimal value of mingraph would probably differ for the protein and solvent regions of the asymmetric unit. In the solvent region few nodes should be put in (high minden) and all noise should be flattened (high mingraph), but large values of these parameters result in an unacceptable level of errors within the protein region. These competing requirements can be uncoupled if a reliable molecular envelope is available – all grid points outside of the envelope can be flattened and the skeletonization procedure applied only within the envelope. As expected, the optimal value of mingraph decreases with the use of an envelope (Fig. 8*c*, dashed lines). The parameters mingraph and maxden are also coupled since the average graph size decreases with increasing maxden. With the lower mingraph for skeletonization with an envelope, the deterioration in performance for maxden greater than 2.5 was considerably slowed (Fig. 8*d*, dashed lines). This implies that the most damaging errors in the skeletonization protocol are in the pruning and not the thinning step. This is despite the fact that many more nodes are removed during thinning than during pruning (Table 1).

Effect of errors in the data

To investigate the effect of random errors in combination with missing data, 5% of the reflections were set to zero and random errors were added to the remaining reflections to produce a final R factor of 10%. The missing reflections were allowed to float at the value calculated from the latest skeleton. The skeletonization procedure proved remarkably stable to errors in the data (Fig. 6, diamonds). Thus,

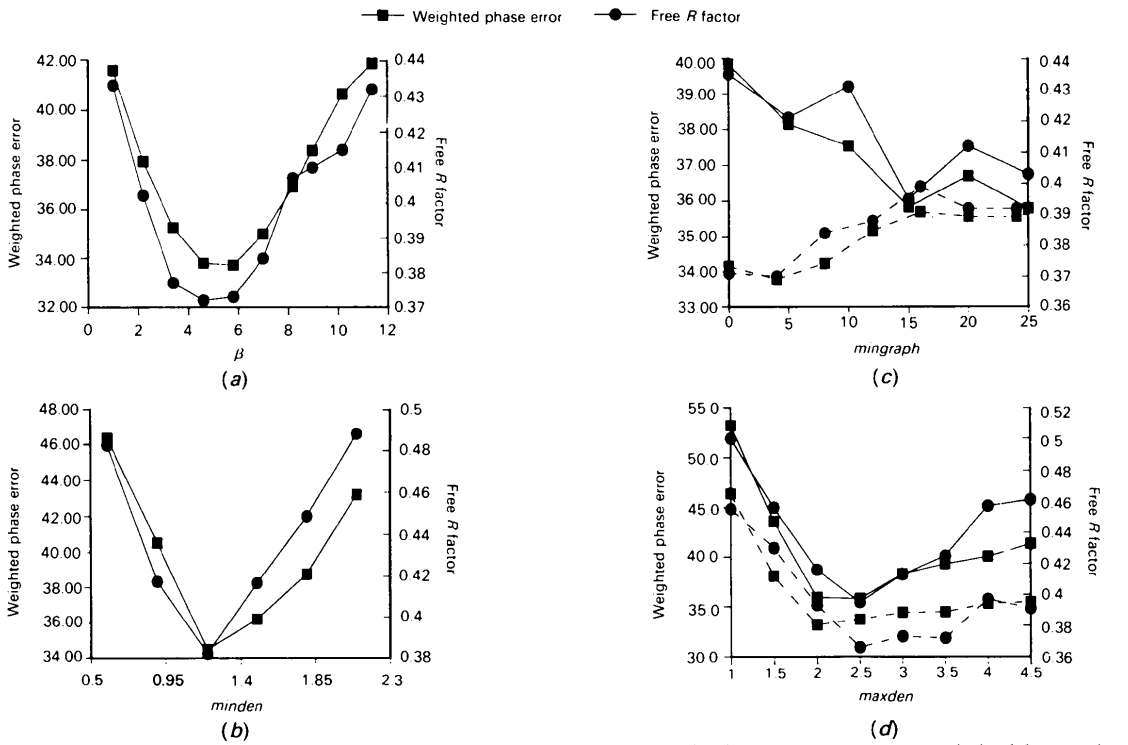

Fig. 8. (*a*)–(*d*) Optimization of parameters using the free R factor. The skeletonization parameters were optimized by varying each parameter independently. For each set of parameters, the skeletonization procedure was applied to the molecular-replacement problem as described in the legend to Fig. 7 except that only four cycles of refinement were carried out. The phase error (squares) and free R factor (circles) after the fourth cycle are shown in the figure (solid lines). All parameters except for that being varied were held fixed at the values listed in the legend to Fig. 1. The optimization of mingraph and maxden was repeated (dashed lines) using a solvent calculated after the eighth cycle of a run with the standard parameter set. Density in the solvent region was flattened prior to skeletonization. The basic parameter set for refinement with a solvent envelope was the same as in Fig. 1 except that mingraph was set to 6 rather than 15. Minden and maxden are expressed as numbers of standard deviations above the mean. The units for β are Å^2.

perfect data are not required for the success of the skeletonization procedure; large improvements can be achieved with incomplete data sets containing random noise.

Dependence of skeletonization on low-resolution data

We next investigated the sensitivity of the skeletonization procedure to low- and high-resolution Fourier data cutoffs. Commonly, the very low resolution reflections are not measured in X-ray diffraction experiments. To simulate this, the molecular-replacement test case was repeated with a low-resolution cutoff of 15 Å. Omission of the low-resolution data significantly reduced the performance of the skeletonization procedure (Fig. 9, diamonds). The requirement for low-resolution data probably results from the fact that absolute and not local contrast is used to determine the position of the nodes in the skeleton and the order in which they are considered for removal (see *Methods*).

Fortunately, the power of the skeletonization procedure can be largely restored through use of a solvent envelope which effectively compensates for the loss of low-resolution information. A reasonable molecular envelope can be rapidly calculated using Leslie's reciprocal-space adaption of Wang's procedure (Leslie, 1987). However, the envelope calculated using phases from a starting molecular-replacement model tends to cut off large portions of the missing density and is thus quite poor. A better envelope is obtained after the iterative skeletonization procedure during which much of the missing structure is regenerated. An envelope was generated from the map produced after 20 cycles of skeletonization in the absence of the very low resolution data (Fig. 9, diamonds). The iterative skeletonization procedure was then repeated starting with phases from the molecular-replacement model, adding nodes only within the envelope. The envelope was recalculated every four cycles. The drop in phase error was increased by almost 50% through the use of the envelope (Fig. 9, circles).

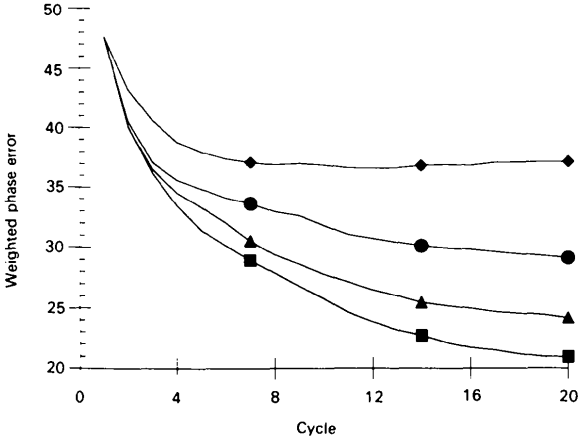

Fig. 9. Stability of skeletonization to missing low-resolution data. The molecular-replacement test problem (Fig. 2, squares) was repeated after omitting all data below 15 Å resolution (diamonds). An envelope was calculated from the map generated after 20 cycles and the procedure was repeated beginning with the starting map and flattening all density outside of the envelope (circles). The refinement procedure was again repeated, using the envelope, except that the missing low-resolution reflections were allowed to float at the value calculated from the latest skeletonized map (triangles). The refinement procedure was repeated a final time, using an envelope and including all low-resolution data (squares). The parameter set was the same as in Fig. 2, except mingraph was changed to 6 when an envelope was used.

As discussed above, the density constraints enforced by the skeletonization procedure lead to coupling of the structure factors. To determine whether this coupling is sufficient to restore missing low-resolution data, the previous experiment was repeated using $|F_{calc}|$'s from the latest skeleton instead of zeros for the low-resolution reflections during calculation of the electron-density maps after the first cycle. After 20 cycles of skeletonization, the $|F_{calc}|$'s for the missing reflections were within 5.2% of the true values. Not surprisingly, allowing the low-resolution missing reflections to 'float' significantly improved the refinement procedure (Fig. 9, triangles). Inclusion of low-resolution reflections presumably reduces fluctuations in the density along the protein chain from one region of the envelope interior to another. The combination of a solvent envelope and allowing the missing reflections to float restored the skeletonization procedure to nearly its power for complete data sets (Fig. 9, compare triangles and squares).

Dependence of skeletonization on high-resolution data

Two effects make the iterative skeletonization procedure potentially sensitive to the high-resolution limit of the data set. First, the accuracy of the skeletonization procedure depends on the resolution of the map. The skeletonization algorithm is biased towards the shortest path through the density. α-Helices present a particularly difficult problem: the shortest path through density at high resolution is along the helical backbone, but at low resolution may be through hydrogen bonds along the helical axis. Second, the phase-improvement problem becomes more poorly determined with the omission of increasing amounts of high-resolution data since this significantly reduces the ratio of observations (Fourier amplitudes) to parameters (nodes in the skeleton).

To investigate the sensitivity of the iterative skeletonization procedure on the extent of high-resolution

data, the experiment described in Fig. 6 was repeated using different high-resolution cutoffs. As shown in Fig. 10, a dramatic deterioration in performance was observed when the high-resolution limit was reduced from 2.5 (squares) to 3.0 (circles) to 3.5 Å (triangles). With a high-resolution cutoff of 3.5 Å, the procedure failed to improve the phases even slightly. Reducing the number of parameters by increasing the grid spacing did not improve the performance of the procedure at 3.5 Å resolution (Fig. 10, open triangles); the increase in the ratio of data to free parameters is presumably offset by an increase in the errors during skeletonization. Apolipoprotein E is an entirely helical protein and the problem of 'short circuiting' down the helix axis is particularly acute; the method may be more tolerant to the absence of high-resolution data in the case of a β-sheet protein.

Concluding remarks

The phase-refinement strategy described here should be a useful addition to the arsenal of tools available for solving macromolecular structures from diffraction data. The powerful constraints of linearity and connectivity reduce the multiplicity of solutions to the phase problem (Baker *et al.*, 1992). The free *R* factor provides a means to optimize the *PRISM* method for the particular problem at hand.

The use of connectivity as a restraint for phase refinement was pioneered by Bhat & Blow (1982). In their method, a starting model is extended through regions of contiguous, well connected high density to produce a modified and hopefully improved electron-density map. Our approach is similar in concept but offers several important advantages. First, the skeletonization procedure is efficient even in the absence of a starting model and enforces the additional chemical constraint of chain linearity. Second, one of the problems of the earlier approach – scaling the modified density to the starting molecular-replacement model – is resolved by a reciprocal-space scaling of the sum of the structure factors from the model and the skeletonized map to the observed amplitudes. Third, the parameters required for the density modification can be optimized using the free *R* factor. Our method appears to have a broader radius of convergence since the phases continue to improve through 40 cycles of refinement, while convergence was reached in one or two cycles in the applications described in the earlier work. More detailed comparisons are not possible since the two methods have not been applied to the same test case.

The *PRISM* method absolutely depends on the availability of sufficient phase information, either from molecular replacement or isomorphous replacement, to generate an initial map which contains stretches of connected density. There is often a large gap between obtaining such a map and being able to trace an atomic model through the density. The apparent very large radius of convergence of the *PRISM* iterative skeletonization procedure promises to make it a powerful method for improving an otherwise uninterpretable map to a point at which it may be readily interpreted.

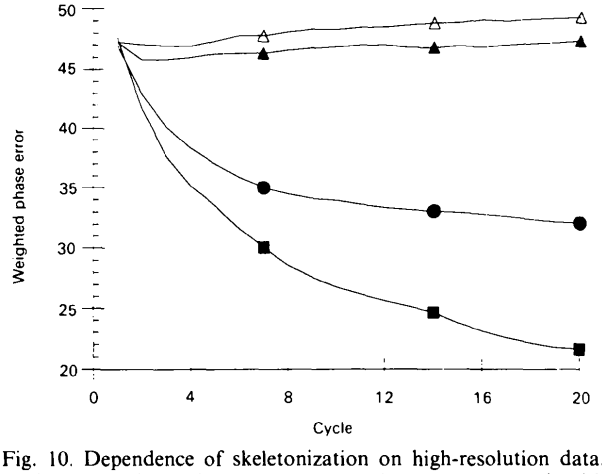

Fig. 10. Dependence of skeletonization on high-resolution data. The molecular-replacement test problem was repeated with high-resolution cutoffs of 2.5 (squares), 3.0 (circles) or 3.5 Å (triangles). The grid spacing was either 0.67 (closed symbols) or 1.0 Å (open symbols).

We wish to thank Shell Chen for programming assistance. This research was supported by funding from the Howard Hughes Medical Institute (DB and DAA) and through National Institutes of Health grants DK39304 and DK26081 (CB and RJF). DB is an HHMI LSRF Postdoctoral Fellow.

References

BAKER, D., KRUKOWSKI, A. E. & AGARD, D. A. (1993). *Acta Cryst.* D**49**, 186–192.
BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F. JR, BRICE, M. D., ROGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.
BHAT, T. T. & BLOW, D. M. (1982). *Acta Cryst.* A**38**, 21–29.
Biosym Technologies (1991). *INSIGHT*II. Biosym Technologies, San Diego, California, USA.
BRÜNGER, A. (1992). *Nature (London)*, **355**, 472–474.
GREER, J. (1985). *Methods Enzymol.* **115**, 206–226.
JONES, T. A. (1985). *Methods Enzymol.* **115**, 157–170.
LESLIE, A. W. (1987). *Acta Cryst.* A**43**, 134–135.
MAIN, P. (1979). *Acta Cryst.* A**35**, 779–785.
SERC Daresbury Laboratory (1986). *CCP*4. *A Suite of Programs for Protein Crystallography*. SERC Daresbury Laboratory, Warrington, England.
SIM, G. A. (1960). *Acta Cryst.* **13**, 511–512.
WANG, B. C. (1985). *Methods Enzymol.* **115**, 90–111.
WILLIAMS, T. (1982). PhD thesis, Univ. of North Carolina, Chapel Hill, USA.
WILSON, C. & AGARD, D. A. (1993). *Acta Cryst.* A**49**, 97–104.
WILSON, C., WARDELL, M., WEISGRABER, K. H., MAHLEY, R. W. & AGARD, D. A. (1991). *Science*, **252**, 1817–1822.