Acta Cryst. (1993). D49, 108–119

Low-Resolution Real-Space Envelopes: Improvements to the Condensing Protocol Approach and a New Method to Fix the Sign of Such Envelopes

By S. Subbiah

Beckman Laboratories for Structural Biology, Department of Cell Biology, Stanford University School of Medicine, Stanford, CA 94305, USA

(*Received* 16 *June* 1992; *accepted* 14 *September* 1992)

Abstract

The condensing-protocol strategy for obtaining low-resolution real-space envelopes of macromolecules from native Fourier intensities alone was recently reported [Subbiah (1991). *Science*, **252**, 128–133]. The present work introduces three improvements to the original methodology. In terms of the j index, which was defined as the crude ratio of correctly placed scatterers to those incorrectly placed, these changes collectively contribute to j values that range from 8 to 28 and in one case even infinity. This compares with the previously reported values of 1.9 and 2.6. Nevertheless, under different conditions, both the original and the current improved condensing protocol can preferentially converge on the solvent or the matter region. Since such control is not absolute there still remains a certain amount of ambiguity in deciding whether the congealed gas represents bulk solvent or bulk matter. To overcome this ambiguity, a simple method based on the systematic placement of point scatterers on a grid lattice within and without the envelope is presented. This 'sign-fixing' method is shown to work with both perfect envelopes as well as the cruder envelopes produced by the improved condensing protocol. The new methodology is illustrated using four macromolecular examples.

Introduction

I recently reported on a method – the condensing protocol – that can be used to obtain low-resolution envelopes from native Fourier intensities alone (Subbiah, 1991). The method employs a random gas of hard-sphere point scatterers and alters their coordinates in prescribed ways while maximizing the correlation coefficient between its calculated amplitudes, F_c, and the available native amplitudes, F_o. The method preferentially converged on the bulk solvent regions of the unit cell rather than the region occupied by the bulk matter. Since this initial work, I have been able to refine this method further.

This paper will first focus on these improvements and will assume familiarity with my previous work and the terminology employed therein. Appropriately, I start with a slight change in terminology that has been suggested to me by Bill Bennett of EMBL. The term 'hard-sphere point scatterer' is a mouthful. Since these objects preferentially condense on the regions that lack ordered matter, the term 'notdots' is more accurate and certainly more interesting. Accordingly, if and when they converge on the matter fraction they can be called 'dots'. Prior to, and during, the condensation process the terminology can display neutrality in predisposition by being 'adots'.

This paper has a second and quite unrelated focus which concerns the aforementioned lack of absolute control over whether the congealed adots happen to be dots or notdots. Obviously, this uncertainty in the 'sign' of the resultant evelopes needs to be resolved. So in this second part I present a simple method – the sign-fixing method – that can successfully decide whether the resultant low-resolution envelope wraps bulk solvent or bulk matter. The method is shown to work well with both perfect calculated envelopes and, more importantly, with the cruder envelopes generated by the improved condensing protocol. The method appears to be general and works well with all space groups considered so far. It also appears to be insensitive to the solvent content; it has been tested over the wide solvent content range that typical macromolecular crystals adopt *i.e.* from 30 to 70%.

First, I present the methodology behind both the three improvements and the sign-fixing strategy. Then I demonstrate the improvements by predicting envelopes for four macromolecular cases. Next, the sign-fixing method is applied to the same cases using perfect calculated low-resolution envelopes as well as the cruder envelopes that are predicted by the improved condensing protocol. Finally I discuss the reasons behind the way in which the sign-fixing method works.

Methods

Improvements to the condensing protocol

With regard to my previous work, the improvements I report here are as follows:

0907-4449/93/010108-12$06.00

(1) Previously, during any particular microcycle step, adots were only allowed to move along one of the six directions that are defined by the unit-cell axes. Thus during a macrocycle (*i.e.* a collection of microcycles), the same step size was taken in any of these six directions regardless of the relative dimenions of the unit-cell edges. When the unit cell was particularly 'thin' along one of the cell edges, an adot could occasionally (*i.e.* particularly in early condensation cycles that involve very large step sizes) move a distance greater than the unit-cell length along this thin direction in a single step. Since this resulted in an incommensurate treatment of the three different unit-cell directions the largest value of the step size was selected to be smaller than the thinnest edge length. However, this led to the systematic exclusion of very large step sizes. In effect, this restricted the extent to which the lowest order reflections could influence the overall correlation coefficient, r. {As discussed in Subbiah (1991), the standard Pearson formulation is used for $r = [\sum(E_o - \langle E_o\rangle)(E_c - \langle E_c\rangle)]/[\sum(E_o - \langle E_o\rangle)^2\sum(E_c - \langle E_c\rangle)^2]^{1/2}$, where E_o is the normalized value obtained from the observed Fourier amplitudes, F_o, and E_c is calculated from the current distribution of point scatterers.} This relative under-weighting of the lowest order reflections was particularly significant in the critical early steps of the condensing protocol where step sizes are large.

In my improved scheme, each adot is moved in an arbitrary direction at all times. A random direction vector is chosen at each move. Then a step size consistent with the given macrocycle (as described in my previous work) is taken along this direction vector. This avoids restricting the allowed magnitude of the initial step sizes. Also, since all points in the unit cell can be accessed as opposed to the previous subset that was limited to the directions of the unit-cell edges, the unit-cell contents can be more finely modeled by this gas of condensing adots.

(2) The second improvement involves the manner in which the step sizes are reduced from supercycle to supercycle. Previously step sizes were reduced relatively slowly and in integral ångström units. Now, I reduce them faster and more smoothly. When a given supercycle using a step size of $X_{\rm old}$ terminates, the new supercycle employs a step size of $X_{\rm new}$, where

$$X_{\rm new} = (2/3)X_{\rm old}. \qquad (1)$$

In general for the low-resolution work considered here, this scheduling of step size allows for fewer intervening supercycles between two selected end points for the step-size range X_i to X_f. To allow for a smooth transition between the now more widely spaced supercycles, the actual step-size values applied to an adot in a given macrocycle are randomly picked from a small range centered on $X_{\rm new}$. More specifically, in the new supercycle, the above-mentioned value of $X_{\rm new} = (2/3)X_{\rm old}$ is used to define a small linear range from which the actual applied step sizes are randomly and uniformly selected. To avoid future confusion and provide continuity in terminology with my previous publication, the individual applied step sizes will always be labeled by $X_{\rm subscript}$. Accordingly, the parent mean step size for the new supercycle will henceforth be labeled $\mu_{\rm new}$. Therefore, under this new terminology, equation (1) reads

$$\mu_{\rm new} = (2/3)\mu_{\rm old}. \qquad (2)$$

Consequently, the range spanned by the step sizes under the old terminology, X_i and X_f, will be redefined as the equivalent mean step-size range μ_i to μ_f. So the improved condensing protocol schedules the mean step sizes from supercycle to supercycle, by starting with μ_i, and decreasing as per equation (2) until μ_f is reached. So for any given supercycle q the actual step sizes, X, are randomly selected from a small linear range centered on μ_q. This range is defined as follows,

$$(5/4)\mu_q < X < (5/6)\mu_q \qquad (3)$$

where μ_q is the mean step size for supercycle q. This linear range is designed to provide a smooth and continuous transition in the applied values of X from supercycle to supercycle with no overlap in range. Primarily, the net effect of this new scheduling of step sizes is an order of magnitude decrease in the computer effort required to achieve final condensation. Secondly, the smooth and continuous scheduling of step sizes, besides being more elegant, also allows more of the unit-cell volume to be accessible for exploration by the condensing gas of adots.

(3) The condensing protocol is primarily a strategy to roughly ascertain the rough location and shape of the relative distribution of the bulk matter and bulk solvent. As most macromolecules of crystallographic interest are either globular proteins that are compact in shape or associate in clustered groups within the crystal environment, it is reasonable to constrain the condensing gas of adots further to converge in a 'compact' and clustered manner. This criterion was not included in the original work on the condensing protocol as it would have been difficult to avoid the criticism that such a constraint had 'somehow biased the answer' in a less than honest way. Given this and that the condensing protocol as previously presented could produce convincing final envelopes without benefit of an additional constraint enforcing compactness, I left my experiences with such an option undiscussed. Now I discuss the effects of such a compactness criterion on the condensing protocol which works particularly well in conjunction with the previously mentioned improvements. Based on my

original work, Chris Bystroff at the Unversity of California at San Francisco has also quite independently implemented and had experiences with such an idea of enforcing compactness.

After experimenting with several possible simple implementations of the notion of compactness, it appears that the simplest and most obvious implementation works as well or better than more complex ones. The simplest implementation requires that any and all steps taken by a condensing adot can do so if, and only if, the sum of all the Euclidean distances between it and all the other adots decreases. When assessing the Euclidean distance between any two adots all symmetry-related adots are inspected and the closest one is selected for use in the summation over all pairs of adots. This sum is calculated at each and every microcycle and used to reject non-compact moves. This in a very simple manner strictly forces the sum of Euclidean distances between all adot pairs to monotonically decrease throughout the condensation protocol. To put the notion of compactness in more formal mathematical terms, I define $C_{\rm omp}$ as a measure of compactness,

$$C_{\rm omp} = \sum_{i=1,N_{\rm hs}} \sum_{j=1,i-1} \{ \min_{m=1,n_{\rm sym}} [{\rm distance}(\mathbf{y}_i, S_m \mathbf{y}_j)]\} \tag{4}$$

where $\mathbf{y}_i$ is the coordinate of the ith adot, $N_{\rm hs}$ is the number of adots in the asymmetric unit, S_m is the mth space-group symmetry operator and $n_{\rm sym}$ is the total number of symmetry operators.

This additional requirement of compactness results in substantial improvement in the quality of the final envelopes. In particular the number of 'stray' and misplaced adots in the condensed gas is significantly reduced. Accordingly the j values – the ratio of correctly placed adots to incorrectly placed adots when the unit cell is crudely divided in half – are much higher than previously reported. Further, it appears that the general preference noted in the earlier work for adots converging to notdots (*i.e.* bulk solvent) rather than dots is no longer the case in the presence of compactness and the other improvements. This is not altogether unreasonable since at low resolution the bulk matter in a globular protein crystal is likely to be distributed as a single compact 'blob' while the bulk solvent is less likely to be so.

In this regard, it is worth pointing out that enforcing a compactness constraint in particularly non-compact protein cases (*e.g.* rod-like structures like myosin) may lead to artificially compact condensed envelopes. This difficulty can be easily avoided by carrying out the condensing protocol twice, with and without the compactness criterion, and comparing the results. A benefit of employing the compactness criteria involves missing or incomplete low-resolution data. Without the aid of the compactness criterion such data lead to poorer envelopes and lower j values. However, with compactness enforced the same poor data can lead to better envelopes. Nevertheless, as I shall demonstrate with an example, the best results are obtained when data are as complete as possible and compactness is enforced.

Before discussing some results, let me outline the preferred overall strategy I presently use for obtaining the best low-resolution envelope when presented with the native Fourier amplitudes in an unknown case.

(1) First, obtain a complete and high-quality low-resolution data set, collected preferably on a diffractometer with particular pains taken in obtaining the ultra-low resolution reflections *i.e.* 001, 010, 100 *etc.*

(2) Second, use the rules-of-thumb that I presented in my earlier paper to ascertain the approximate number and properties of the adots needed. Since it is not possible to guarantee with certainty the outcome of the condensation with regard to dots or notdots – particularly when the compactness is enforced – it is advisable to compute the number of adots so that they will 'fill' the smaller of the two volumes – bulk matter and bulk solvent.

(3) Third, perform the condensing protocol using this smaller number of adots from different random starts until complementary images are obtained. The point is to observe several different runs converge to a similar image, while others converge to the opposite image. It is worth noting that in space groups where there are origin ambiguities, judgement on the similarity between different runs should make an allowance for the inability of the condensing protocol to distinguish between different allowed origins. When such complementary images are seen the odds that one of the two images corresponds crudely to the location of the bulk matter are extremely high. In the unlikely instance that no 'inverse' images or worse any kind of strong partitioning of the unit-cell volume appears the protocol should be repeated varying the number of adots used by up to ± 30%. Another variant is to explicitly target for the larger of the two volumes – bulk solvent or bulk matter – rather than the smaller volume. The number of adots should then be chosen accordingly to fill this larger volume optimally. In any case, in my experience, the first few runs have almost always resulted in complementary images. Incidentally, if there is reason to suspect the macromolecule to be dramatically non-globular in shape, the same runs should be repeated without the compactness criterion and compared with the original results.

(4) Once a set of complementary images have been observed, the last step toward establishing a crude low-resolution envelope involves deciding which of the two complementary images corresponds to bulk matter. This step can be done by a new method I

have developed to 'fix the sign' of envelopes. This 'sign-fixing' method is presented next.

The sign-fixing method

Assuming a low-resolution image (either dots or notdots or a perfect image derived from the true atomic structure) of the unit-cell contents is available, the method requires the following five steps.

(1) The ultra-low resolution to which all calculations will be limited to, U, is chosen to include approximately the 20 lowest order reflections.

(2) The available image – dots/notdots or in the instance of a test case the true atomic structure – is used to compute a Fourier transform at this resolution U. Dots/notdots are treated as dummy C^{α} atoms for this transformation. The ultra-low resolution Fourier data set calculated in this manner is then back-transformed to obtain an ultra-low resolution electron-density map. To ensure that the standard Shannon criteria is not exceeded, this map should be sampled at a grid spacing less than $U/3$. I typically use about $U/10$.

(3) Adots – hard-sphere point scatterers – are placed at the grid points having the highest electron density. When the highest 10% of such grid points have been 'dotted' this list of adots is labelled the '$w = 0.1$ envelope contour'. The Fourier transform of this distribution of static point scatterers is then computed for all the ultra-low resolution reflections to U. These computed Fourier amplitudes are then used together with the observed Fourier amplitudes to calculate the usual Pearson correlation coefficient, r_u, exactly as in my previous work (Subbiah, 1991). This whole process is repeated for the series of w values = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and all corresponding values for r_u plotted on a graph of r_u against w. This curve will be referred to as the 'variation in r_u with positive envelope contours' or '+ve env' for short.

(4) An exactly analogous procedure to step 3 is then conducted but this time the series in w proceeds with progressively larger fractions of the unit cell 'dotted' that include the grid points with the *lowest* electron density. Thus the $w = 0.1$ envelope contour would include a list of adots placed at grid points that have the lowest 10% of electron density and $w = 0.2$ would include adots placed at the lowest 20% of all grid points and so on. The corresponding correlation coefficients are calculated and plotted on the same plot used in step 3 and referred to as the 'variation in r_u with negative envelope contours' or '−ve env' for short.

(5) Inspection of the '+ve env' and the '−ve env' curves should indicate whether the available image represents bulk matter or bulk solvent. Of the two curves, the one which has the lower and more rapidly decreasing value of r_u at higher values of w corresponds to a series of increasingly generous envelopes about the true bulk matter. In other words, at higher values of w one of the two curves will exhibit a more rapid decrease in r_u with respect to the other curve. Typically, the onset of this decrease will occur after a value of w greater than the smaller of the two fractions – the bulk matter fraction or the bulk solvent fraction. If the curve that decreases more rapidly is the '+ve env' one the condensed adots are dots and represent the distribution of bulk matter. If this curve happens to be the '−ve env' one the condensed adots are notdots and represent bulk solvent.

Results

Applications of the improved condensing protocol

Now I illustrate the improvements in methodology discussed above using the same two examples as my previous work – the N-terminal binding domain of 434 repressor (PDB reference 1R69, Bernstein *et al.*, 1977) and the elastase molecule from *Pseudomonas aeruginosa* (ELA) (Flaherty, Pley, Benvegnu & McKay, 1992). I further illustrate this using two other examples – porcine citrate synthase (PDB reference 1CTS) and a fake crystal using only the first of the two domains that constitute the elastase molecule mentioned above (ELAHALF). In order to simplify the analysis of the results all origins have been fixed to be 0,0,0. The very low resolution, the very crude analysis criterion **j** and the tendency of the compactness criterion to compress shapes made it difficult in some cases to resolve between enantiomers. Fortunately, in these cases, this issue became irrelevant as the enantiomers happened to be highly overlapped in the first place.

The 434 *repressor test case.* The first test case involves the all-helical 1R69 molecule which crystallizes in $P2_12_12_1$ ($a = 32.8$, $b = 37.5$, $c = 44.6$ Å) with one 69-residue molecule per asymmetric unit. The estimated solvent content is between 30 and 35%. The improved condensing protocol, including the compactness requirement, was conducted using parameters as similar as possible to the previous work: K, the high-resolution cutoff = 6 Å; $N_{\rm ref}$, the number of reflections to this resolution = 177; $N_{\rm hs}$, the number of adots used = 43; μ_i, the mean step size for the first supercycle = 35.0 Å; μ_f, the mean step size for the last supercycle = 4.6 Å; and $\lambda_{\rm hs}$, the hard-sphere radius of an adot = 1.5 Å.

After about 1 min of central processor time on a VAX 8550 computer the adots converged as notdots onto the bulk solvent (Figs. 1*a*–1*c*). The corresponding numerical results were: initial r, correlation coefficient of random adots = −0.04; final r, correlation coefficient of converged notdots = 0.83; initial $C_{\rm omp}$,

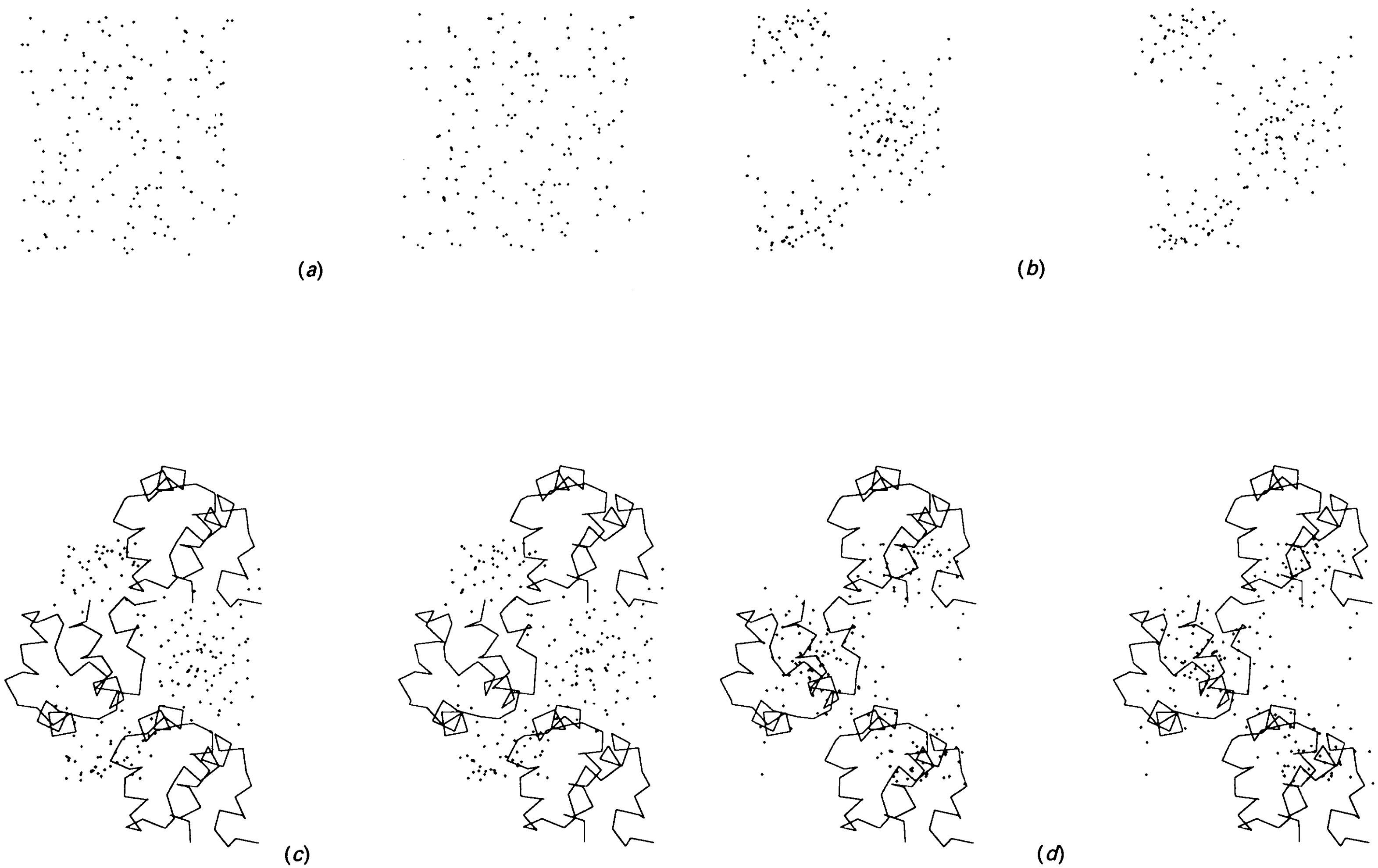

Fig. 1. (*a*) Results for the 1R69 case are illustrated. A random distribution of 43 adots is shown in stereo. The full $P2_12_12_1$ unit-cell contents are shown.(*b*) The random distribution shown in (*a*) condenses to this notdot distribution in the presence of the compactness criterion. (*c*) The notdot distribution of (*b*) is shown superimposed on the correctly packed unit cell of the 1R69 crystal. For clarity, only three of the four molecules found in the unit cell are displayed as C^{α} traces. (*d*) A dot distribution obtained with compactness enforced.

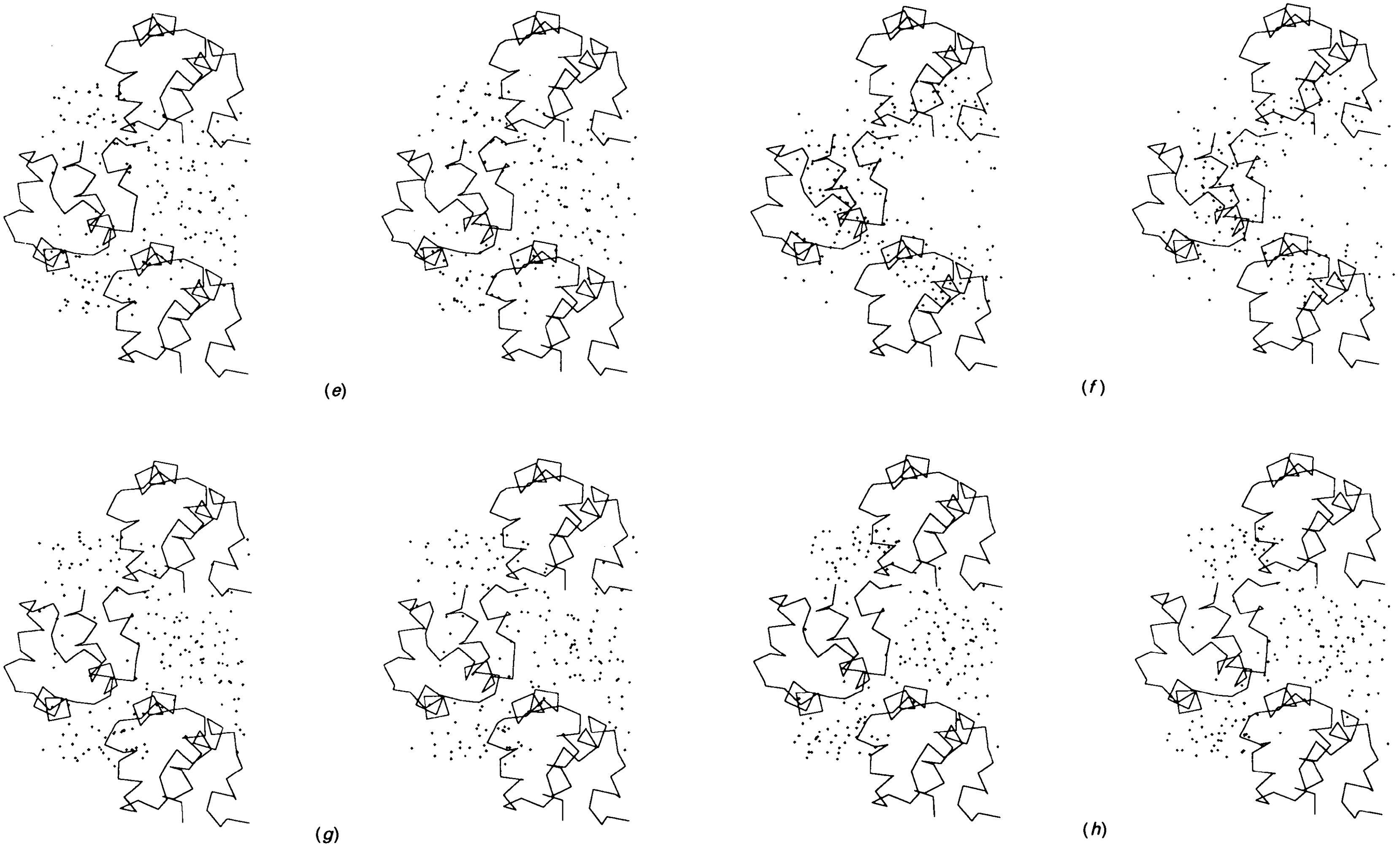

Fig. 1 (*cont.*) (*e*) A notdot distribution obtained with the compactness requirement relaxed. Note that relative to the distribution in (*c*) there are more notdots incorrectly lying within the macromolecule. (*f*) A dot distribution obtained with the compactness requirement relaxed. Clearly, relative to the distribution in (*d*) there are more dots incorrectly lying within the solvent void. (*g*) A notdot distribution obtained with the compactness requirement relaxed. However, the condensation was performed with 10% of the low-resolution data missing. Notice the poorer quality of the partitioning compared to both (*c*) and (*e*). (*h*) A notdot distribution obtained with the compactness requirement re-enforced and with 10% of the data missing. Note the relative improvement in the quality of the partitioning over that in (*g*). Relative to (*g*) there are less notdots incorrectly lying within the macromolecule.

compactness of random adots = 10 604; final $C_{\rm omp}$, compactness of converged notdots = 8439; initial *j*, ratio of adots in solvent to those outside = 1.2; and final *j*, ratio of notdots in solvent to those outside = 20.5.

As can be seen in Fig. 2, the *j* value is very high and corresponds to a very clear and correct partitioning of the bulk volume. For comparison, similar parameters resulted in a *j* value of 2.6 under the older implementation of the condensing protocol (Subbiah, 1991). When the same conditions were used in another run from a different random start the adots converged as dots onto the bulk matter (Fig. 1*d*) with the following numerical results: initial *r*, correlation coefficient of random adots = 0.02; final *r*, correlation coefficient of converged dots = 0.85; initial $C_{\rm omp}$, compactness of random adots = 10 624; final $C_{\rm omp}$, compactness of converged dots = 8596; initial *j*, ratio of adots in bulk matter to those outside = 1.1; and final *j*, ratio of dots in bulk matter to those outside = 7.6.

Incidentally, since the measure *j* behaves nonlinearly, the values 7.6 and 20.5 are not as different as the numbers appear to suggest. When compactness was not enforced and the other two new improvements were in force, different random starts gave notdots (Fig. 1*e*) and dots (Fig. 1*f*) with consistently poorer values of *j* = 7.6 and 2.6 respectively. Thus, relaxing the compactness requirement results in less distinct envelopes.

In order to study the effects of low-resolution reflections missing from the data set, 10% of the reflections from infinity to *K* Å were deleted. All 177 reflections were sorted according to resolution and every tenth member deleted. Using this incomplete data set and relaxing the compactness requirement, the same case discussed above with reference to Fig. 1(*c*) was repeated. The resulting notdots (Fig. 1*g*) had a very low value of *j* = 1.9. However, when the compactness requirement was turned back on the *j* value of the resulting notdots (Fig. 1*h*) improved to 6.2, despite the incompleteness of the data set. Thus, as summarized in Table 1, the compactness criterion can 'rescue' the frequently encountered situation where the available experimental data set is incomplete. *j* decreased even more signficantly when the missing reflections were chosen to be the 17 lowest order ones.

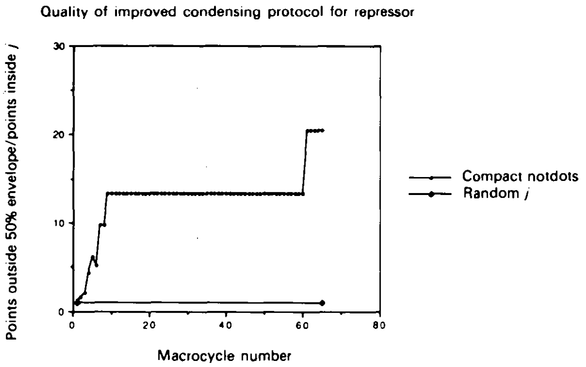

Fig. 2. The behavior of *j* for the 1R69 case as the condensing progressed from the adot distribution of Fig. 1(*a*) to the notdot distribution of Fig. 1(*b*). Note that as described in the text, owing to the definition of *j*, the rate of increase in *j* is inherently non-linear.

Table 1. *Compactness rescues incomplete data*

This table summarizes the *j* values associated with the condensed distributions depicted in Figs. 1(*c*)–(*h*). Inspection of the variation in *j* as one proceeds down the list illustrates the ability of the compactness criterion to partially compensate for the harmful effect of an incomplete low-resolution data set.

Conditions		Results		
Data quality	Compactness	Dots	Notdots	*j* value
Complete	Yes	Yes	—	7.6
Complete	No	Yes	—	2.6
Complete	Yes	—	Yes	20.5
Complete	No	—	Yes	7.6
Incomplete by 10%	No	—	Yes	1.9
Incomplete by 10%	Yes	—	Yes	6.2

The elastase test case. The second test case involves the ELA molecule which crystallizes in $P2_12_12_1$ ($a = 124.4$, $b = 51.5$, $c = 44.5$ Å) with one 298-residue molecule per asymmetric unit. The estimated solvent content is about 40%. The improved condensing protocol, including the compactness requirement, was conducted using the following parameters: *K*, the high-resolution cutoff = 7 Å; $N_{\rm ref}$, the number of reflections to this resolution = 537; $N_{\rm hs}$, the number of adots used = 199; μ_i, the mean step size for the first supercycle = 40.0 Å; μ_f, the mean step size for the last supercycle = 5.3 Å; and $\lambda_{\rm hs}$, the hard-sphere radius of an adot = 1.5 Å.

The condensation converged to notdots (Fig. 3*a*) with the following results: initial *r*, correlation coefficient of random adots = −0.01; final *r*, correlation coefficient of converged notdots = 0.80; initial $C_{\rm omp}$, compactness of random adots = 406 100; final $C_{\rm omp}$, compactness of converged notdots = 299 901; initial *j*, ratio of adots in solvent to those outside = 1.05; and final *j*, ratio of notdots in solvent to those outside = 9.0.

For comparison, the equivalent result for ELA from my previous work (Subbiah, 1991) was a final *j* of 1.9. Another run from a different random start condensed to dots with *j* = 3.4 and $C_{\rm omp}$ = 323 974 (Fig. 3*b*).

The citrate synthase test case. The third test case involves the 1CTS molecule which crystallizes in $P4_12_12$ ($a = 77.4$, $b = 77.4$, $c = 196.4$ Å) with one 437-residue molecule per asymmetric unit. The estimated solvent content is about 57%. The improved condensing protocol, including the compactness requirement was conducted using the following parameters: K, the high-resolution cutoff = 8 Å; $N_{\rm ref}$, the number of reflections to this resolution = 771; $N_{\rm hs}$, the number of adots used = 200; μ_i, the mean step size for the first supercycle = 70.0 Å; μ_f, the mean step size for the last supercycle = 6.1 Å; and $\lambda_{\rm hs}$, the hard-sphere radius of an adot = 1.5 Å.

The condensation converged to notdots (Fig. 4*a*) with the following results: initial r, correlation coefficient of random adots = 0.04; final r, correlation coefficient of converged notdots = 0.84; initial $C_{\rm omp}$, compactness of random adots = 519 081; final $C_{\rm omp}$, compactness of converged notdots = 378 968; initial j, ratio of adots in solvent to those outside = 1.1; and final j, ratio of notdots in solvent to those outside = 10.1.

Another run from a different random start condensed to dots with $j = 27.6$ and $C_{\rm omp} = 384605$ (Fig. 4*b*).

The artificial half-elastase test case. The fourth test case involved the artificial ELAHALF molecule that includes only the first 149 residues of the ELA case described above. This case was designed to test the

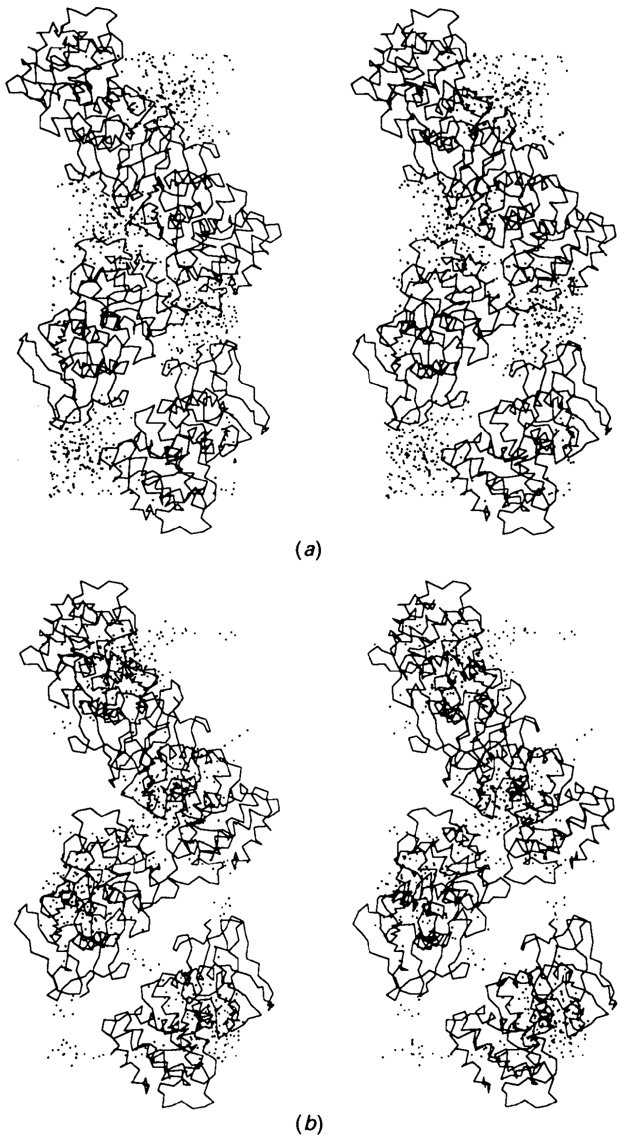

Fig. 3. (*a*) The notdot distribution for the ELA molecule is shown together with all four $P2_12_12_1$ copies of the C^α trace. The boxed unit cell has **c** coming out of the page and the origin at the lower right far corner. (*b*) A dot distribution obtained under conditions similar to that in (*a*) is shown.

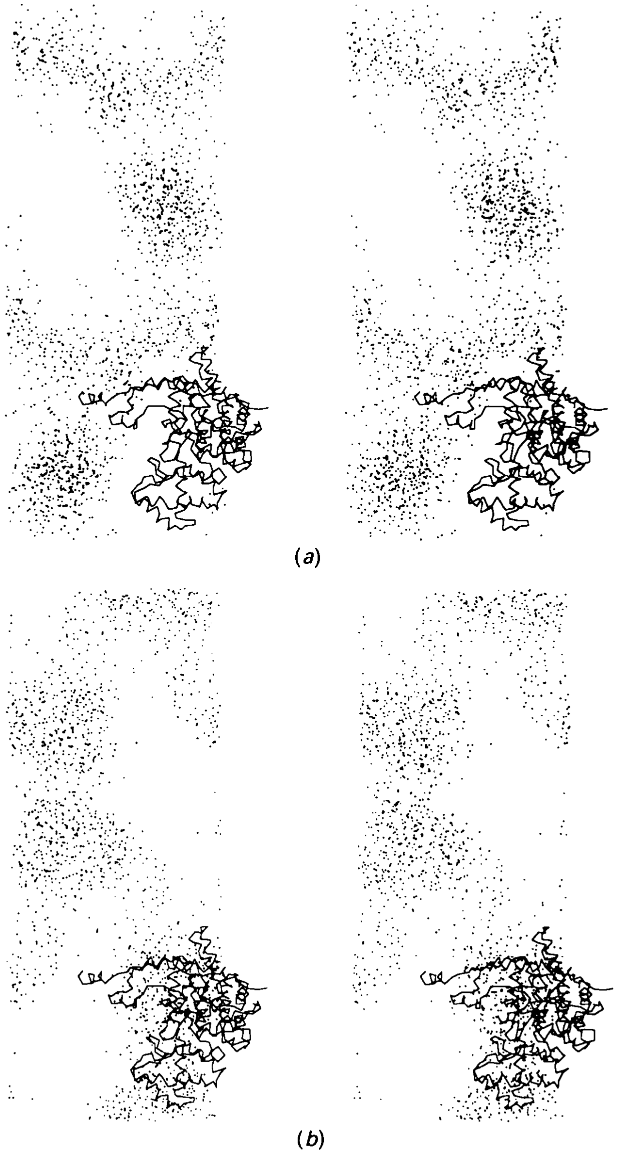

Fig. 4. (*a*) One of the eight 1CTS molecules in the $P4_12_12$ unit cell is shown superimposed on a notdot distribution obtained *via* the improved condensing protocol. The boxed unit cell is shown with **b** going into the page and the origin at the lower left near corner. (*b*) A dot distribution obtained in a manner similar to that in (*a*) is shown.

reliability of the improved condensing protocol at high solvent content levels. A mock Fourier data set was generated using exactly the same unit-cell information as in ELA. Except for N_{hs} being 150 all condensing parameters were the same as in the ELA case. This simply reflects the larger solvent content of 70%. Different random runs produced notdots (Fig. 5*a*) and dots (Fig. 5*b*). The *j* values were particularly impressive in this case. In fact in the case of the dots, since all the adots had moved to the correct half of the unit cell, the final value for *j* was infinity. The case of the notdots was only a little less impressive at 24.

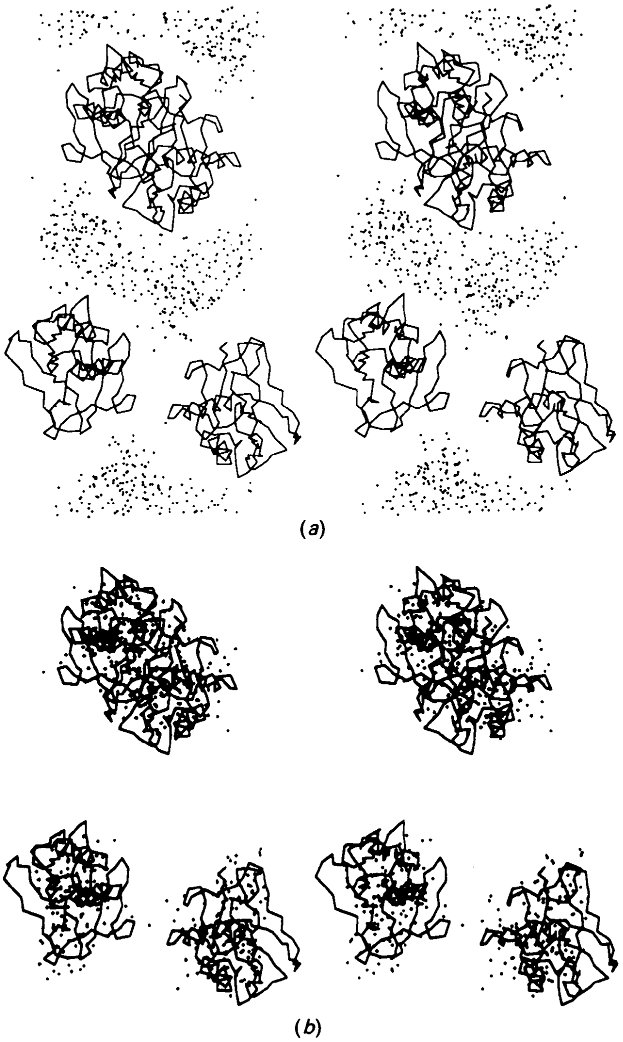

Fig. 5. (*a*) The notdot distribution for the artificial ELAHALF molecule is shown together with all four $P2_12_12_1$ copies of the C^{α} trace (two of the copies appear overlapped). The axes and origin of the unit cell are identical to those in Figs. 3(*a*) and 3(*b*). (*b*) A dot distribution obtained under conditions similar to those in (*a*) is shown.

Applications of the sign-fixing method

To illustrate the method, it was first used to correctly predict that a perfect electron-density map of the previously discussed citrate synthase molecule calculated at ultra-low resolution represented bulk matter and not bulk solvent. Such a perfect map was calculated to $U = 30$ Å using the corresponding 22 lowest order reflections. The grid spacing was chosen to be $U/7.75 \simeq 3.87$ Å. The corresponding '+ve env' and '−ve env' curves are shown in Fig. 6. The '+ve env' clearly decreases when large fractions of the unit cell are uniformly filled with point scatterers, while the '−ve env' remains high. According to the rule in step 5, this indicates that the available structure corresponds to bulk matter.

Next, in order to simulate the experimental situation, where the nature of the available condensed set of scatterers is uncertain, a series of mock test cases were conducted using the same four macromolecular examples discussed earlier. In each case, the condensed scatterer distributions that had been predicted earlier by the improved condensing protocol were subjected to the sign-fixing procedure. In all four cases the procedure was able to resolve the inherent sign ambiguity and the respective sets of condensed scatterers were correctly declared to be either dots or notdots.

The citrate synthase test case. The dots shown in Fig. 4(*b*) were used to calculate the 22 lowest order

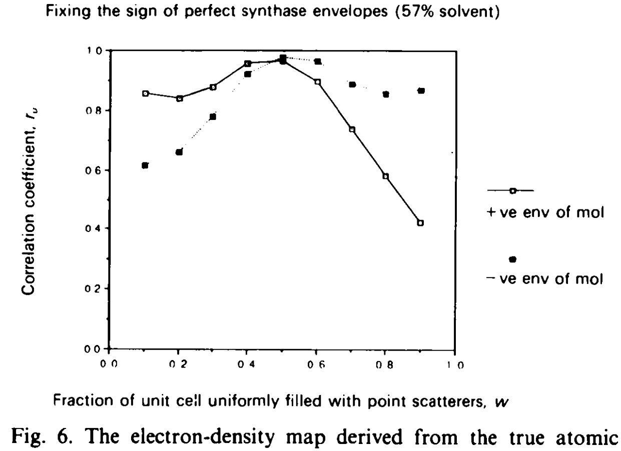

Fig. 6. The electron-density map derived from the true atomic structure of the synthase molecule was used to generate the '+ve env' and the '−ve env' curves shown. The rapid decrease of the '+ve env' curve as *w* exceeds 0.5 relative to the minor decrease in the '−ve env' curve correctly predicts the fact that the electron-density map represents bulk matter. To ensure consistency in symbol usage, curves representing bulk matter (*i.e.* dots) always use squares. '+ve env' curves of bulk matter are represented by open squares connected by solid lines. '−ve env' curves of bulk matter are represented by solid squares connected by dotted lines. The curves describing bulk solvent (*i.e.* notdots) are represented similarly except for the open/solid squares being replaced by open/solid triangles.

Fourier amplitudes. As before, a grid spacing of 3.87 Å was used in generating the '+ve env' and '−ve env' curves shown in Fig. 7(*a*). The condensing protocol is correctly predicted to have converged as dots, since at high *w* the '+ve env' curve decreases while the '−ve env' one remains high. The same calculation was repeated using the notdot distribution of Fig. 4(*a*). Here too, the presence of notdots is correctly predicted since at high *w* the '−ve env' curve is decreasing while the '+ve env' curve remains high (Fig. 7*b*).

The 434 *repressor test case.* Using $U = 15$ Å and a grid spacing of $U/7.3 \simeq 2.05$ Å, the '+ve env' and '−ve env' curves were calculated for both the dots and the notdots shown in Figs. 1(*d*) and 1(*c*), respectively. The two pairs of curves plotted in Figs. 8(*a*) and 8(*b*) clearly predict both the dot and the notdot situation correctly. This particular example demonstrates the viability of the method at the lower end of the solvent content range typically seen with macromolecular crystals.

The elastase test case. The notdots of Fig. 3(*a*) were subjected to the sign-fixing procedure. The resultant curves in Fig. 9 were calculated using the 16 lowest order reflections to $U = 25$ Å and a grid spacing of $U/13.7 \simeq 1.82$ Å. Once again the converged adots are correctly predicted to be notdots.

The artificial half-elastase test case. This example is particularly suited to assessing the viability of this method at very high values of the solvent content. The dots of Fig. 5(*b*) could again be correctly identified by plotting the two curves (Fig. 10). Note that the discrimination, particularly at very high *w* – corresponding to a grossly incorrect estimation of the true solvent content – is not as high. Studies with

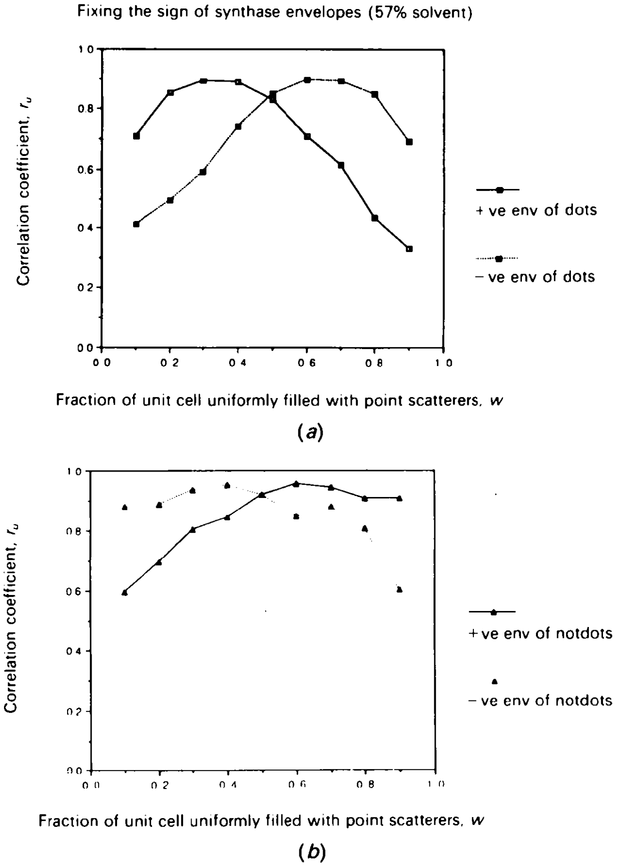

Fig. 7. (*a*) The dots resulting from an application of the improved condensing protocol to the synthase case are correctly predicted to represent the bulk matter. As in Fig. 6, the '+ve env' curve decreases more rapidly than the '−ve env' one at large values of *w*. (*b*) The notdots resulting from an application of the improved condensing protocol to the synthase case are correctly predicted to represent the bulk solvent. Here, in contrast to (*a*) the '−ve env' curve decreases more rapidly than the '+ve env' one at large values of *w*.

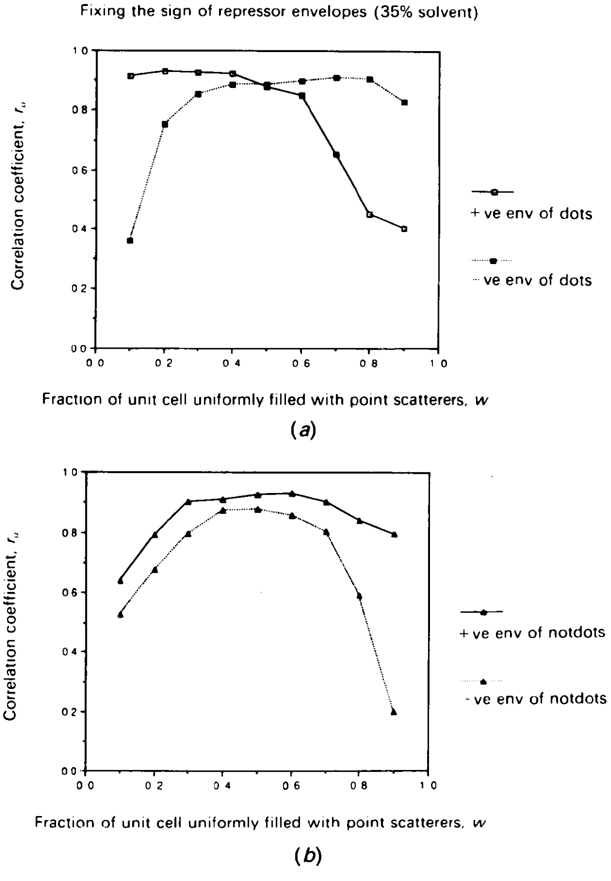

Fig. 8. (*a*) The dots resulting from an application of the improved condensing protocol to the repressor case are correctly predicted to represent the bulk matter. The '+ve env' curve decreases more rapidly than the '−ve env' one at large values of *w*. (*b*) The notdots resulting from an application of the improved condensing protocol to the repressor case are correctly predicted to represent the bulk solvent. The '−ve env' curve decreases more rapidly than the '+ve env' one at large values of *w*.

other examples with very low or very high solvent contents confirm this behaviour of somewhat decreased discrimination at extreme values of the solvent content. Nevertheless, as seen here, in all such cases the discrimination is sufficient for the task at hand.

Discussion

The improvements to the condensing protocol are all of a simplistic nature and require little by way of explanation. The only caveat is that care has to be demonstrated in employing the compactness criterion when a non-compact macromolecule is suspected.

Fixing the sign of elastase envelopes (40% solvent)

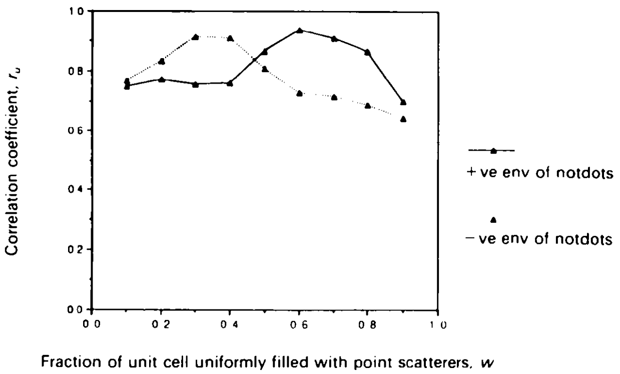

Fig. 9. The notdots resulting from an application of the improved condensing protocol to the elastase case are correctly predicted to represent the solvent. The '−ve env' curve decreases more rapidly than the '+ve env' one at large values of w.

Fixing the sign of half-elastase envelopes (70% solvent)

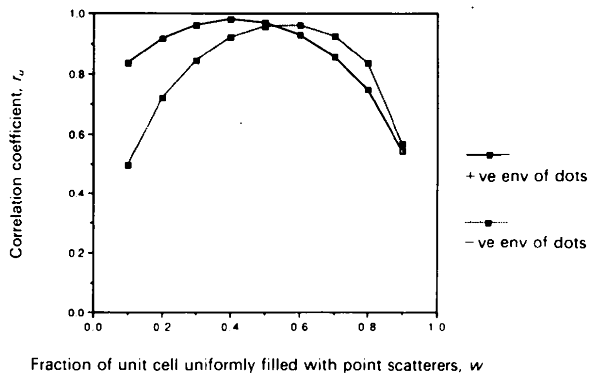

Fig. 10. The dots resulting from an application of the improved condensing protocol to the half-elastase case are correctly predicted to represent the bulk matter. The '+ve env' curve decreases more rapidly than the '−ve env' one at large values of w.

On the other hand, the sign-fixing method requires more explanation. This method exploits a fundamental difference between the bulk solvent and bulk matter found in typical crystals of biological macromolecules. In terms of its Fourier diffraction properties, the bulk solvent is essentially featureless. This is not true of the bulk matter which contains internal variations in density, particularly when modeled by a generous ultra-low resolution envelope. Macromolecules typically have internal cavities and surface indentations. Since the point scatterers used in creating the grid-lattice representation of the envelope are themselves featureless, they can, in general, better approximate the diffracting properties of the bulk solvent as compared to the bulk matter. Thus when the envelope encapsulating bulk solvent is made more generous than the correct solvent contour, the additional point scatterers that incorrectly protrude into the bulk matter volume contribute to a decrease in the correlation coefficient, r_u. However, when generously modeling the bulk matter in an analogous manner, the layer of point scatterers that incorrectly protrude beyond the true bulk matter contour into the bulk solvent volume cause an even greater decrease in the ultra-low resolution correlation coefficient, r_u. Put another way, overly generous point-scatterer-filled envelopes that model bulk matter cause a greater decrease in r_u relative to a situation where the bulk solvent is similarly overmodeled. This relative effect tends to be amplified when perfect envelopes are not available. Given the same contour level for comparison, in crude envelopes even more point scatterers incorrectly protrude beyond the true contour levels. Thus with the imperfect envelopes generated by the condensing protocol, the imperfections cause a relatively larger discrepancy in the behaviour of the ultra-low resolution correlation coefficient when compared to perfect envelopes. In effect, the sought-after signal is amplified when dealing with the real-life situation of correctly deciding the nature of the condensed scatterers produced by the condensing protocol.

Concluding remarks

In conclusion it appears that the three new improvements to the condensing protocol result in much clearer and quicker partitioning of the unit-cell volume into bulk matter and bulk solvent. Further the new compactness requirement can successfully rescue the experimental situation where a small but significant percentage of the low-resolution data is missing. However, care needs to be taken with this requirement when dealing with molecules that are far from globular in shape. Finally, these improvements affect the outcome of the condensation process in such a way that dots occur more frequently than with the original method. This ambiguity in sign can be resolved with the aid of the new and simple sign-fixing method. Thus the crude envelopes generated by the improved condensing protocol approach can always be correctly catalogued as representing either bulk matter or bulk solvent. This is a necessary ingredient of any future attempt to use the crude spatial information generated by the improved condensing protocol as a starting point to extract low-resolution *ab initio* phases and extend them to higher resolution.

I am deeply indebted to T. B. Splott for the stuff that dreams are made of, the children at Stanford Childrens Hospital for the inspiration to dream and my supervisor M. Levitt for the freedom to dream. G. Petsko, a Damon Runyon–Walter Winchell Cancer Research Postdoctoral Fellowship (DRG 1019) and an NIH grant No. GM-41455 awarded to M. Levitt made this work possible. I also thank Dr C. Bystroff, Dr P. David, Dr W. Hendrickson, P. Gosh and K. Flaherty for valuable scientific interaction.

References

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B. MEYER, E. F. JR, BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.

FLAHERTY, K. M., PLEY, H., BENVEGNU, N. & MCKAY, D. (1992). Unpublished results.

SUBBIAH, S. (1991). *Science*, **252**, 128–133.