Acta Cryst. (1994). D**50**, 132–138

Low-Resolution Real-Space Envelopes: the Application of the Condensing-Protocol Approach to the *ab initio* Macromolecular Phase Problem of a Variety of Examples

BY P. R. DAVID AND S. SUBBIAH

Beckman Laboratories for Structural Biology, Department of Cell Biology, Stanford University School of Medicine, Stanford, CA 94305-5400, *USA*

(*Received* 22 *June* 1992; *accepted* 14 *October* 1993)

Abstract

A recently reported method – the improved condensing protocol [Subbiah (1991). *Science*, **252**, 128–133; (1993). *Acta Cryst.* D**49**, 108–119] – for obtaining low-resolution macromolecular envelopes is applied to five varied test cases. These examples were chosen to illustrate the general applicability of the method to the wide range that typical macromolecular crystals adopt. The cases include small and large asymmetric unit volumes (4.7×10^4 to 1.17×10^6 Å^3), low and high symmetry (2 to 12 symmetry elements), small and large proteins (1570 to 12216 non-H atoms), orthogonal and non-orthogonal unit cells, a wide variety of space groups ($P2_1$ to $P6_322$), small and large solvent contents (33–80%), and a case of non-crystallographic symmetry (threefold). In all five cases the inherent ambiguity of the condensing protocol in differentiating between bulk matter and bulk solvent is then resolved by use of the recently reported sign-fixing method (Subbiah, 1993).

Introduction

This paper demonstrates the strength and utility of recently developed methods in the *ab initio* phasing of macromolecules. Recent work demonstrates that the consecutive application of the improved condensing protocol and the sign-fixing method can be used to obtain crude low-resolution envelopes of macromolecular structures (Subbiah, 1991, 1993). In this paper we apply these methods to five very different proteins and show their general applicability and effectiveness. The examples we present are bovine carboxypeptidase A (5CPA), human immunodeficiency viral (HIV) protease (4HVP), porcine citrate synthase (1CTS), bovine leucine aminopeptidase (1LAP) and influenza virus hemagglutinin (4HMG). By design, the cases include small and large asymmetric unit volumes (4.7×10^4 to 1.17×10^6 Å^3), low and high symmetry (2 to 12 symmetry elements), small and large proteins (1570 to 12216 non-H atoms), orthogonal and non-orthogonal unit cells, a broad range of space groups ($P2_1$ to $P6_322$), including the two most frequently encountered space groups in protein crystallography ($P2_1$ and $P2_12_12_1$), small and large solvent contents (33–80%), and a case of non-crystallographic symmetry (threefold). In all cases, using error-free calculated Fourier-amplitude data derived from the known atomic structures, we obtain crude but striking low-resolution macromolecular envelopes without the use of any prior phase information.

Method

For each protein, coordinates were extracted from the Brookhaven Protein Data Bank (PDB, Bernstein *et al.*, 1977). Structure factors were calculated by expanding the PDB coordinates to *P*1, calculating the structure factors in *P*1 using *GENSFC* (*CCP*4; SERC Daresbury Laboratory, 1979). The *P*1 structure-factor data set, F_o, was reduced to a compact asymmetric unit. This data was then presented to the program *GET-A-FIX*, which includes implementations of both the improved condensing protocol and the sign-fixing method. The improved condensing protocol was then carried out from different random starting collections of hard-sphere point scatterers (adots). The improved condensing protocol requires the following parameters to be supplied: Unit-cell constants, a, b, c, α, β, γ (Å and °); the high-resolution cut off for the supplied F_o data, K (Å); the number of reflections between infinity and this cut off, N_{ref}; the number of adots to place in each asymmetric unit, N_{hs}; the mean step size for the first supercycle, μ_i (Å); the mean step size for the last supercycle, μ_f (Å); and the hard-sphere radius of an adot λ_{hs} (Å). Once the value of K is chosen all other parameters can be determined by the rules-of-thumb reported earlier (Subbiah, 1991, 1993). For continuity, we shall discuss these briefly here. The condensing protocol relies on straightforward maximization of the fit between F_o and the equivalent data calculated for the current distribution of adots, F_c. For this to have any chance of correctly converging to the solution it is crucial that the many applied constraints cause the direct and

ISSN 0907-4449

robust condensation through the large space of possible answers. One way of improving the odds is by using only a small but sufficient number of adots to 'fill' and describe the target crude macromolecular envelope. Previous work has shown empirically that at ultra low resolution most macromolecules can be described well by 100 to 300 adots. For larger asymmetric units and larger macromolecules one can maintain this number of adots and still 'fill' the desired volume by increasing the radius of an adot, λ_{hs}, from the standard one of 1.5 Å used previously with small and average-sized problems (Subbiah, 1991). Since the closest distance between scatterers is then $2\lambda_{hs}$, it would be pointless supplying Fourier data much higher than $4\lambda_{hs}$. Therefore, K depends on deciding what λ_{hs} has to be in order to fill the required space with about 100–300 adots in the final condensed image. The space to be filled is simply that fraction of the asymmetric unit that has the lower volume (*e.g.* if the protein occupies 60% of the volume, then one need only fill 40% with adots). Therefore, one needs to have a reasonable estimate of the solvent content beforehand. (If this is unknown, a good estimate is obtained by multiplying the total number of amino acids in the asymmetric unit by the average volume of an amino acid, 140 Å^3, and then dividing this by the volume of the asymmetric unit to get the bulk matter fraction. Subtracting this from unity gives a good estimate of the solvent content.) Previous work (Subbiah, 1991) shows that random packing considerations cause this volume to be further reduced by a factor of 0.6. Additionally empirical Fourier mobility considerations require this to be further divided by 6. This volume is then divided by the number of adots that has been selected from the range 100 to 300. This number is then the final volume required of a spherical adot, $4/3\pi\lambda_{hs}^3$. This gives λ_{hs}. As described before, λ_{hs} then fixes the value of K. Knowing K, immediately gives N_{ref}. It is important to ensure at this point that N_{ref} (the number of knowns) is at least two or three times larger than N_{hs}, the number of unknowns. Otherwise, a small value is selected for N_{hs} and the calculations repeated. Of the parameters listed as required inputs to *GET-A-FIX*, only μ_i and μ_f remain. μ_i is selected to be two-thirds of the length of the diagonal of the standard asymmetric unit. Since only data to K Å resolution is to be supplied it is pointless to take moves that sample much less than K. So μ_f is arranged to be typically between K and $K/2$. Using these parameters, *GET-A-FIX* was run from different random starts until complementary images were seen. One type of image is then expected to reflect the crude distribution of bulk matter and the other should reflect the inverse distribution of bulk solvent. For these examples, this ambiguity was then resolved by performing the sign-fixing method (Subbiah, 1993), using *GET-A-FIX*. The required parameters for this are: the resolution that includes about 10–20 lowest order reflections, U; and the grid-spacing for creating a lattice description of the condensed adot images, t.

Table 1. *Summary of parameters and results*

Asym. cell vol. = volume of asymmetric unit cell in 1000 Å^3. K = upper resolution limit of data used in the condensing protocol. N_{ref} = number of reflections used in the condensing protocol. N_{hs} = number of hard spheres (adots) used. m_i/m_f = the mean starting step size for the first and last supercycles. U = resolution of sign-fixing calculations. t = grid spacing on the maps used in the sign-fixing calculations.

Protein	5CPA	4HVP	1CTS	1LAP	4HMG
Space group	$P2_1$	$P2_12_12_1$	$P4_12_12$	$P6_322$	$P4_1$
High resolution, K (Å)	7.	8.	8.	9.	10.
Asym. cell vol. (kÅ^3)	73	48	147	154	1,171
Solvent content (%)	33	50	58	55	80
No. of reflections, N_{ref}	485	250	774	571	2557
No. of adots, N_{hs}	114	110	200	145	407
Start/final step, m_i/m_f	55/3.8	60/3.5	75/4.3	97/5.5	141/8.2
Resolution, U (Å)	25	17	32	50	50
No. of reflections to U	12	17	18	24	23
Grid spacing, t (Å)	2.0	4.0	4.1	5.3	5.8

The sign-fixing method employs these parameters to generate two curves of correlation coefficients, $r_u = [\sum_u(F_o - \langle F_o\rangle)(F_c - \langle F_c\rangle)]/[\sum(F_o - \langle F_o\rangle)^2\sum(F_c - \langle F_c\rangle)^2]^{1/2}$ against various fractions, w, of the unit cell filled with adots (*i.e.* scatterers) placed on a grid of spacing t. The relative behavior of these two curves – '+ve env' and '−ve env' – reflect the nature of the condensed adots (when adots describe the bulk matter they are called dots and when they describe solvent they are call notdots). If the '+ve env curve' decreases more rapidly (*i.e.* with increasing w) from the maximum value of r_u than the '−ve env' curve at values of w near the estimated value of the expected macromolecular fraction, the condensed adots are predicted to represent bulk matter. Similarly, if the opposite is the case and the '−ve env' curve descends more rapidly relative to the '+ve env' curve at similar values of w, the condensed adots are predicted to represent bulk solvent. Since all the examples were test cases for which the true distribution of bulk solvent and bulk matter were known, all predictions could be verified. Moreover, a simple indicator of the convergence of the condensed adots to their targets was assessed by using the previously defined crude measure, j (Subbiah, 1991). j is the ratio of correctly placed adots over those incorrectly placed when the unit cell is very crudely divided into a bulk-matter half and a bulk-solvent half, irrespective of the true solvent content. For a solvent content of 50%, a value of 1 would imply a random distribution and, conversely, a value of infinity would imply perfect convergence into the target half. However, since the condensing protocol is insensitive to origin ambiguities, j values were not calculated in the $P2_1$ and $P4_1$

cases, which have translational ambiguities along one axis. Moreover no effort was made to adjust manually the origin of the condensed adots to make it coincident with the actual protein origin. So although we do not report j values for these cases, high values can be expected on the basis of the striking degree of correct convergence observed in projections along the appropriate axis of origin ambiguity.

Results

All parameters for the five test cases – general crystal information, information for the improved condensing protocol and information for the sign-fixing method – are tabulated in Table 1. These test cases were chosen to fill distinct needs. First, we wanted to cover a wide range of symmetries, both crystallographic and non-crystallographic, as well as including the most frequently encountered space groups. Second, a large range in the volume and dimensions of the asymmetric unit was sought to allay concerns about the applicability of this method only to large or small problems. Thirdly, it was also desirable that a broad a range of solvent content be represented. The selections were made from the Protein Data Bank with these criteria in mind. In the following cases the number of monomers per unit cell varies from 2 to 12, the number of monomers per asymmetric unit varies from 1 to 3, the numbers of amino

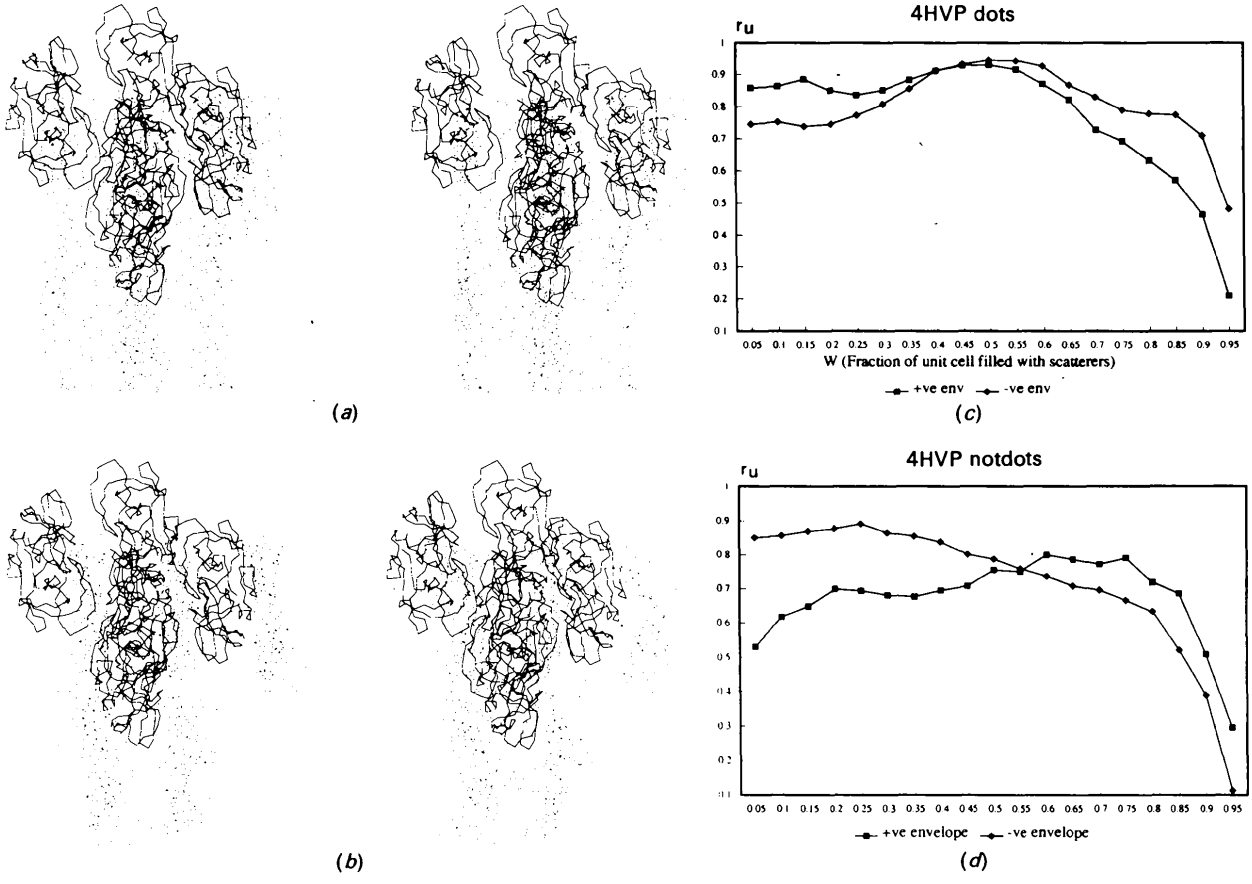
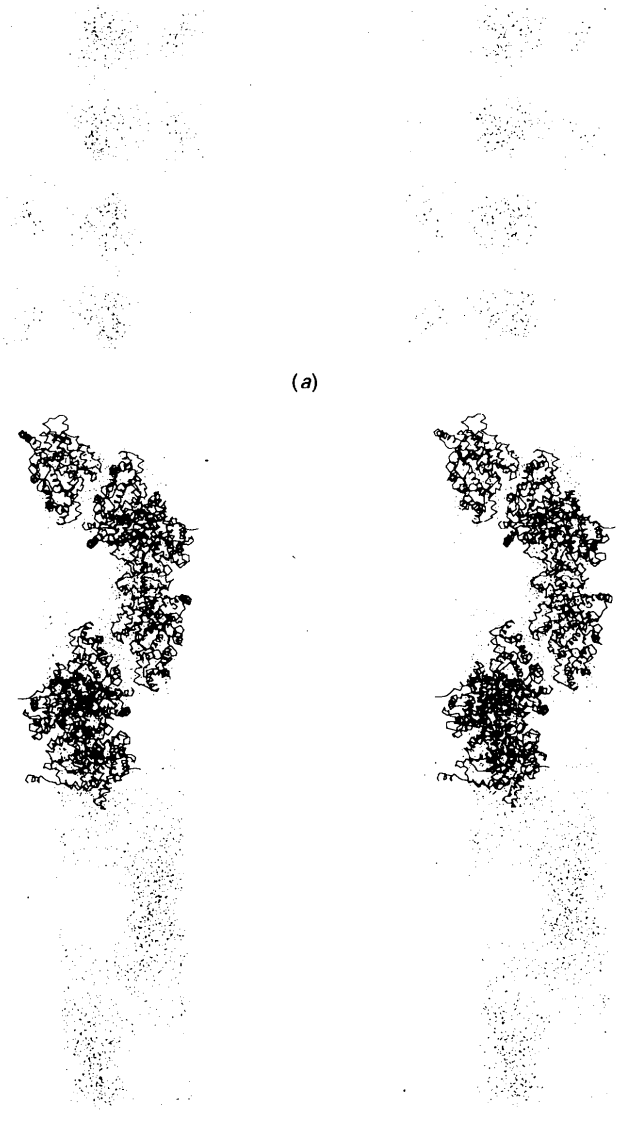


Fig. 1. (*a*) The C^{α} trace of the 4HVP dimer is shown superimposed upon condensed dots. Here the dots are clearly centered on the protein core. (*b*) The C^{α} trace of 4HVP dimer is shown with the layers of notdots. The notdots are clearly separated from the volume occupied by the protein. This figure differs from (*a*) solely by the randon number used to seed the initial adot locations. (*c*) This figure shows the result of applying the sign-fixing procedure to the dots in (*a*). As described elsewhere (Subbiah, 1993), two curves (one positive and one negative) of r_u *versus* w are shown. The sign-fixing method depends on the behavior of these two curves near $w = (1 -$ the solvent fraction). In this example, (1 − solvent fraction) is 0.5, and therefore we examine the behavior of the '+ve envelope' curve at 0.5 and above. As w increases from this value (1 − solvent fraction) the '+ve envelope' curve decreases more quickly than the '− ve envelope' curve. Because the '+ve envelope' decreased more rapidly, the sign-fixing method predicts that the condensed adots of (*a*) to be dots (Subbiah, 1993). (*d*) This figure shows the result of applying the sign-fixing procedure, as in (*c*), but now to the notdots of (*b*). Here, the '− ve envelope' curve decays more quickly near $w = 0.5$. This scenario is opposite to that seen in (*c*). Therefore, the method correctly predicts the condensed adots of (*b*) to be notdots.

acids range from 198 to 1527, the solvent content ranges from 33 to 80%. Porcine citrate synthase (PDB-1CTS) was chosen for continuity with previous work (Subbiah, 1991). Bovine carboxypeptidase A (PDB-5CPA) was chosen as an example of a small protein in a common, non-orthogonal space group with low solvent content. HIV protease (PDB-4HVP) was chosen as a protein of normal solvent content, in a very common space group. Bovine leucine aminopeptidase (PDB-1LAP) was chosen as a larger protein in a non-orthogonal space group with a normal solvent content. Hemagglutinin (PDB-4HMG) has a high molecular weight and volume per asymmetric unit as well as a relatively high solvent content.

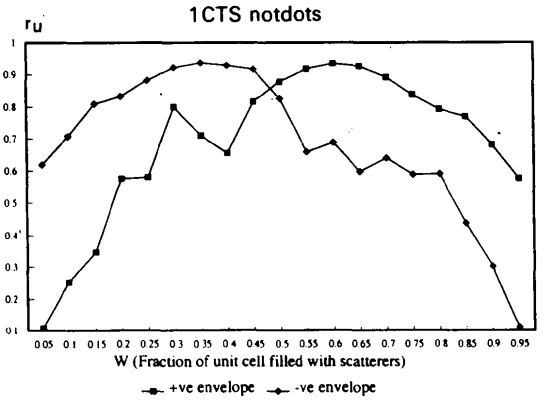
HIV protease (PDB-4HVP)

This protein was chosen as an example of a small protein with normal solvent content in a very common space group, $P2_12_12_1$. Two runs of the improved condensing protocol differing only in the random number seed resulted in dots (Fig. 1*a*) and notdots (Fig. 1*b*). The respective *j* values are shown in Table 1. In both cases, the striking feature of the unit cell is the alternating layers of density given by the layers of twofold axes in this space group $P2_12_12_1$. The positions of the remaining molecules are obvious and lie in the remaining open spaces either side of the pictured molecules. Here, the fixed origin allows a simple application of the sign-fixing method to the condensed adot distribution to determine whether the distribution is of dots or notdots. The sign-fixing method correctly shows the '+ve env' curve to descend more rapidly relative to the '−ve env' curve or values of *w* higher than about $w = 0.5$ (Fig. 1*c*). At $w = 0.5$, half the unit cell has been filled with adots, leaving the other half empty. This scenario corresponds to the estimated solvent content of 50%. When sign-fixing was carried out on the notdots, at about $w = 0.5$, the '−ve env' curve



(*a*)

(*b*)



(*c*)

Fig. 2. (*a*) Here notdots are shown without the C^α trace of the citrate synthase dimer. (*b*) Here two adjacent cells of notdots are shown clearly distinct from the C^α trace of the 1CTS monomer in a view parallel to *c* and bisecting *a* and *b*. The volume occupied by the remaining dimers is clearly visible as the holes above the C^α trace, with the last dimer split between the top of the figure and the bottom. Citrate synthase forms a crystallographic homodimer, formed by the interdigitating ends of the disparate helices of each monomer that overlap in the core of the unit cell, forming the eight crystallographic monomers into four dimers. (*c*) This figure shows the result of applying the sign-fixing procedure to the notdots in (*a*). Here the somewhat ragged plot of r_u illustrates the effect of using too few reflections in the sign-fixing process. Nonetheless, the signal is still strong, and starting at about $w = 0.42$ the '−ve envelope' clearly decays more quickly indicating that these are notdots.

was seen to descend more rapidly than the '+ve env' curve (Fig. 1*d*). This correctly predicts their notdot nature (Fig. 1*d*).

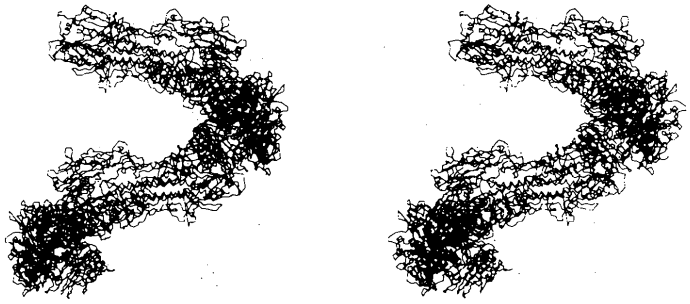
Porcine citrate synthase (PDB-1CTS)

This case was chosen for continuity with previous work (Subbiah, 1993). For 1CTS, there are eight copies of the monomer in the unit cell, which contains four homodimers in all. The space occupied by the other monomer of the dimer is strikingly obvious just to the left of the C^{α} traces shown. The origin is fixed in this rather long unit cell. It has relatively high symmetry ($P4_12_12$), a larger than average molecular weight and a larger than average solvent content. A run of the condensing protocol that resulted in notdots is shown in Fig. 2(*a*). The *j* value (Table 1) is particularly high. The sign-fixing method was then applied to these notdots and the resulting curves (Fig. 2*c*) correctly predict that these were indeed notdots (*i.e.* represent bulk solvent). As expected from the estimated solvent content of 58%, the '−ve env' curve decreases rapidly relative to the '+ve env' curve at about $w = 0.42$.

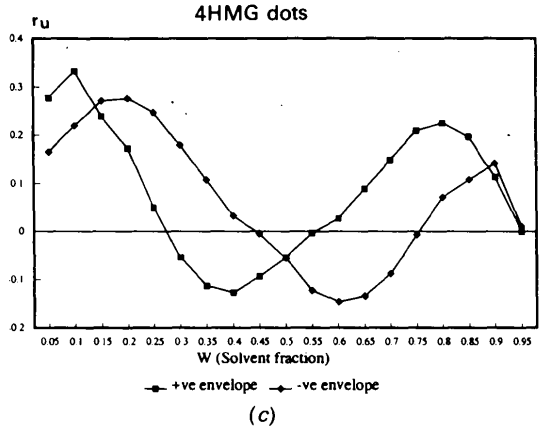
Hemagglutinin (PDB-4HMG)

This example poses three problems simultaneously; first, the solvent content is quite high for proteins, 80%+, second, the asymmetric unit is formed of a non-crystallographic trimer of three chains of 503 glycosylated amino acids each, totalling 12216 atoms, and third, there is an origin ambiguity due to the 4_1 screw axis. 4HMG is one of the largest proteins available in the PDB, with a very high molecular weight, a very high asymmetric unit volume, a very high solvent content and a threefold non-crystallographic symmetry. The dots resulting

(*a*)



(*b*)



(*c*)

Fig. 3. (*a*) The hemagglutinin trimers are viewed side on in this view along the *c* axis. The origin ambiguity is projected into the page and the reader can see that the dots have strongly condensed on the perimeter, where the 4HMG trimers pack. (*b*) Illustrating the spiral packing of the 4HMG trimers, this figure shows the right-handed spiral along *c*. The dots have fortuitously condensed with very little offset from the 4HMG C^{α} trace, but $P4_1$ does have an origin ambiguity along the 4_1 screw axis. (*c*) This figure shows the result of applying the sign-fixing method to the dots in (*a*). Due to the very high solvent content, 80%, of the 4HMG crystals the r_u plot peaks very early at the expected value of $w = 0.2$. The '+ve envelope' then decays rapidly relative to the '−ve envelope' with increasing *w*, confirming that the adots have condensed to dots.

from an application of the condensing protocol are shown in a view parallel to the z axis (Fig. 3*a*). Since there is a translational origin ambiguity along the z axis in this space group, $P4_1$, the adots may not necessarily condense in the vicinity of the protein. In this example, the adots fortuitously condensed almost exactly on the protein space. For clarity, we include another view that is perpendicular to the z axis (Fig. 4*b*). In this view the spiral staircase pattern of the stacking of the protein trimers is evident in the right-handed upward spiral in this figure. The results of the sign-fixing method clearly indicate that at about $w = 0.2$, corresponding to the estimated bulk solvent of 80%, the '+ve env' curve decreases more rapidly with increasing w than the '−ve env' curve.

Bovine carboxypeptidase A (PDB-5CPA)

This was chosen as an example of a typical protein in a common space group, $P2_1$, with a low solvent content. The notdots (Fig. 4*a*) show striking convergence to a compact group and in a non-conflicting site. No manual adjustment of the origin has been made to compensate for origin ambiguity along the y axis. The view has been chosen to be perpendicular to the y axis. j values were not computed. However, given the clear separation of the notdots and protein in projection, this suggests a high value of j. The sign-fixing method only indicates the 'notdot' prediction by a very slim margin near $w = 0.67$, corresponding to the estimated solvent content of 33% (Fig. 4*b*).

Bovine leucine aminopeptidase (PDB-1LAP)

This example was chosen for its high symmetry and hexagonal space group $P6_322$. The 12-fold symmetry is further complicated by the thin triangular nature of the normal asymmetric unit. The advantage of the more uniform sampling of the asymmetric unit can clearly be seen here. Under the previous protocol, the step size would have been limited to 14 Å along the a, b or c axes instead of the current step size of 97 Å along any vector. Thus, compared to before (Subbiah, 1991), the lowest resolution F_o data is now much better sampled. This condensing-protocol run converges to dots (Fig. 5*a*). The sign-fixing method correctly predicts these condensed adots to be dots (Fig. 5*b*), because near the estimated solvent content of 55% ($w = 0.45$) and with increasing w, the '+ve env' curve descends more quickly.

Discussion

The original work on the condensing method showed the possible utility of the method on two typical proteins in $P2_12_12_1$ (Subbiah, 1991). Since then, the method has improved considerably (Subbiah, 1993). The five cases presented here demonstrate the general utility of the improved condensing protocol and the sign-fixing method, as well as an increase in the maximum resolution. The cases discussed here span a large range of volumes, molecular weights, crystallographic and non-crystallographic symmetries, solvent contents and space groups. These cases were

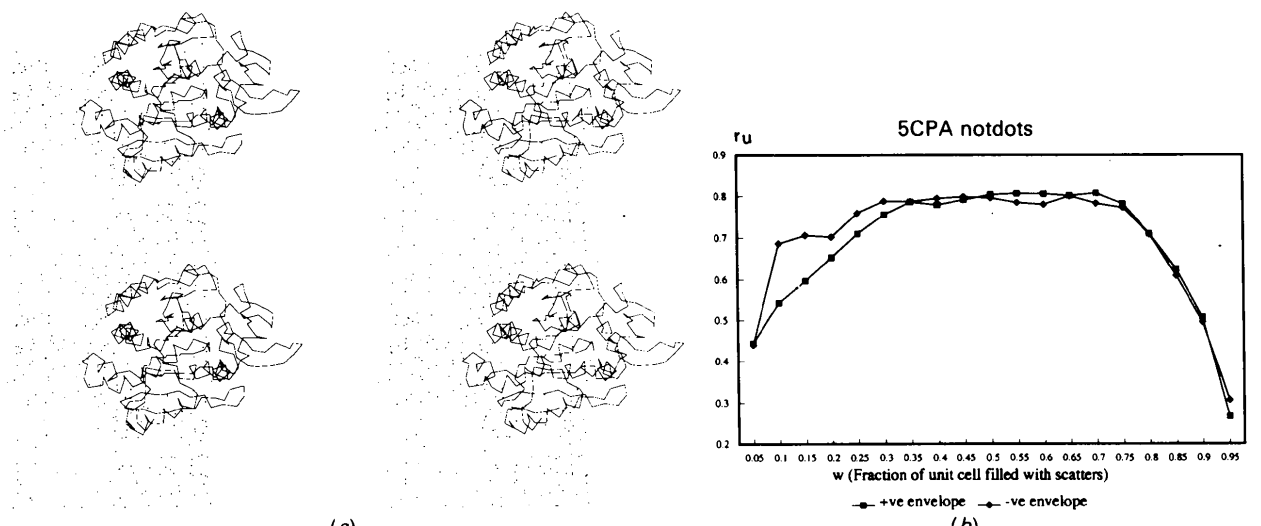


Fig. 4. (*a*) The notdots have coalesced to a very compact form well away from the boundaries of the true protein volume, here represented by the C^{α} trace of 5CPA. No adjustment of the relative origins was made. Notice the volume in the lower right for the second 5CPA monomer. (*b*) This figure shows the result of applying the sign-fixing method to the notdots in (*a*). Near the solvent content of 33% ($w = 0.67$), the '−ve envelope' decays marginally quicker than the '+ve envelope' corresponding to a prediction of notdots. However, here the discrimination is weak.

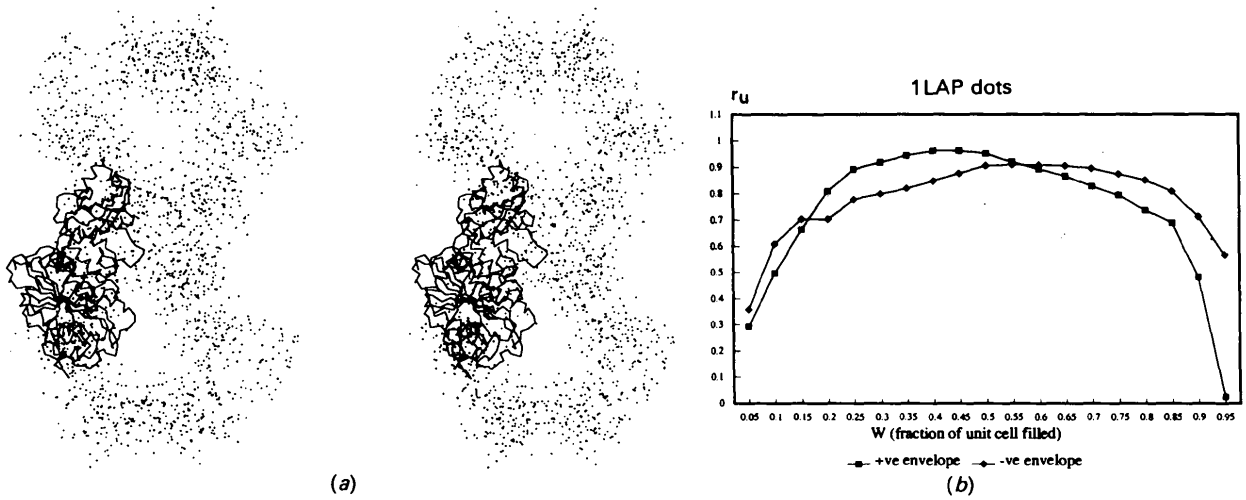


Fig. 5. (*a*) The two hexameric molecules of leucine aminopeptidase are shown side by side and are viewed along their internal threefold axes in this figure. One C^{α} trace of a monomer is shown for comparison. The dots have clearly condensed into the protein region and away from the unoccupied corners of the unit cell. Since each monomer contacts at least six other monomers, both within the unit cell and in adjacent cells, the dots appear more erratic in two dimensions than they really are. Only about one dot in six is misplaced as demonstrated by the *j* value in Table 1. (*b*) This figure shows the result of applying the sign-fixing method to the dots in (*a*). This plot has relatively more reflections included (24 *versus* 18) and the smoothness of the plot should be compared with Fig. 2(*b*). Near the solvent fraction of 55% ($w = 0.45$) the '+ve envelope' is decreasing more quickly. This correctly predicts the adots to have condensed as dots.

specifically selected to push the methods to perform in extreme situations and are representative cases, not our best five examples. Although we do not present the work here, in our hands, the methods have worked as well or better on many tens of different examples. We therefore believe that these techniques are generally applicable to the determination of low-resoution envelopes of macromolecules.

We thank Professor M. Levitt for his support and for creating a stimulating atmosphere. SS is supported by a postdoctoral fellowship from the Daymon-Runyon/Walter-Winchell Foundation (DRG 1019). This work was supported in part by a grant (GM–41455) to ML from the National Institutes of Health.

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.

SERC Daresbury Laboratory (1979). *CCP4. A Suite of Programs for Protein Crystallography*. SERC Daresbury Laboratory, Warrington WA4 4AD, England.

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

SUBBIAH, S. (1993). *Acta Cryst.* D**49**, 108–119.