Acta Cryst. (1993). D**49**, 186–192

Uniqueness and the *Ab Initio* Phase Problem in Macromolecular Crystallography

BY DAVID BAKER, ANTON E. KRUKOWSKI AND DAVID A. AGARD

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

(*Received* 17 *June* 1992; *accepted* 14 *August* 1992)

Abstract

The crystallographic phase problem is indeterminate in the absence of additional chemical information. A successful *ab initio* approach to the macromolecular phase problem must employ sufficient chemical constraints to limit the solutions to a manageably small number. Here we show that commonly employed chemical constraints – positivity, atomicity and a solvent boundary – leave the phase problem greatly underdetermined for Fourier data sets of moderate (2.5–3.0 Å) resolution. Entropy maximization is also beset by multiple false solutions: electron-density maps are readily generated which satisfy the same Fourier amplitude constraints but have higher entropies than the true solution. We conclude that a successful *ab initio* approach must make use of high-resolution Fourier data and/or stronger chemical constraints. One such constraint is the connectivity of the macromolecule. We describe a rapid algorithm for measuring the connectivity of a map, and show its utility in reducing the multiplicity of solutions to the phase problem.

Introduction

Much attention has been devoted to the problem of reconstructing a three-dimensional structure solely from the X-ray diffraction pattern of a single macromolecular crystal. In order to recover the phases of the Fourier components, additional information must be provided. This information, whose source ultimately is chemistry, ranges from knowledge that the electron density must be real and bounded through to the detailed rules of stereochemistry.

Most recent work on *ab initio* phasing in macromolecular crystallography has focused on the search problem – the construction of efficient algorithms and numerical procedures for obtaining the electron density $\rho(x,y,z)$ from the diffraction data $|Fhkl|$ and some limited chemical information, usually positivity and/or atomicity. The question of the uniqueness of the solutions obtained by such algorithms has in contrast been relatively neglected. This is unfortunate since even the most sophisticated of algorithms is bound to fail when the true distribution $\rho(x,y,z)$ is only one of a very large number of distributions satisfying all of the constraints. The goal of this paper is to begin to map the degeneracy of the solution space with the hope of focusing attention on problems likely to have a manageably small number of solutions.

We begin by reviewing analytical results on the uniqueness of solutions to the related phase problem in optics, and then attempt to estimate the degree to which various amounts of chemical information reduce the multiplicity of solutions of the crystallographic phase problem by seeking solutions satisfying the given constraints in the neighborhood of random $\rho(x,y,z)$.

Uniqueness of phase recovery in optics

There is extensive literature on the uniqueness of solutions to the phase problem in imaging applications where one seeks to recover an object ρ from the amplitude of its Fourier transform. The following analysis shows that the phase problem in two or higher dimensions almost always has a unique solution.

Define

$$F = \sum_{y=-L/2}^{L/2} \sum_{x=-L/2}^{L/2} \rho(x,y)\exp[2\pi i(hx+ky)] \quad (1)$$

and

$$I = FF^* \quad (2)$$

where * indicates the complex conjugate. The phase problem is to recover the complex quantity F from the scalar intensities I. With the substitutions $z_1 = \exp(2\pi ih)$ and $z_2 = \exp(2\pi ik)$, F and I become polynomials in the complex variables z:

$$F = \sum_{y=-L/2}^{L/2} \sum_{x=-L/2}^{L/2} \rho(x,y) z_1^x z_2^y. \quad (1')$$

F and I are analytic functions throughout the complex plane provided $\rho(x,y)$ is bounded. If F can be factorized,

$$F = F_1 F_2 F_3 ... F_n \quad (3)$$

0907-4449/93/010186-07$06.00

then

$$F^* = F_1^* F_2^* F_3^* \ldots F_n^* \quad (4)$$

and

$$I = F_1 F_1^* \; F_2 F_2^* \; F_3 F_3^* \ldots F_n F_n^*. \quad (5)$$

Now consider the 2^{n-1} functions H obtained by replacing one or more of the n factors F_i in (3) with F_i^*. As H^* will contain F_i instead of F_i^*, the product $HH^* = FF^* = I$, and the phase problem thus does not have a unique solution. If, conversely, F cannot be factorized, the relation $I = FF^*$ uniquely determines F. If a general algorithm existed for factorizing polynomials in two or more dimensions, the phase problem in principle could be solved simply by decomposing I into its unique factors F and F^* [for more details see Fiddy (1987)].

Uniqueness thus depends on whether F is factorizable. In one-dimensional problems, F is polynomial in a single variable, and the fundamental theorem of algebra guarantees that F can always be expressed as a product of first-order polynomials. Thus, the phase problem in one dimension generally does not have a unique solution. However, there is no equivalent of the fundamental theorem in higher dimensions, and in fact almost all polynomials in two or more variables are irreducible (Hayes, 1987). Thus, neglecting the effects of noise, the phase problem in two or more dimensions almost always has a unique solution. This has been borne out in practice by the development of simple iterative algorithms (Dainty & Fienup, 1987) for reconstructing an object $\rho(x,y)$ from only the magnitude of its transform.

Unfortunately, these arguments do not carry over to the crystallographic phase problem. For the continuous function I to be completely specified by its values at discrete points, it must be sampled at the Shannon limit at least. From (1) and (2) simplified to one dimension for convenience, I is seen to be the transform of the autocorrelation of $\rho(x)$:

$$I = \sum_{x=-L}^{L} \Big[\sum_{x'=-L/2}^{L/2} \rho(x')\rho(x+x') \Big] \exp(2\pi i h x). \quad (6)$$

The limits on the outer sum are twice that of the inner sum since with ρ non-zero for $-L/2 < x' < L/2$, the autocorrelation of ρ will be non-zero for $-L < x < L$. As the spectral width of I is twice that of F, the Nyquist spacing required for I $(1/2L)$ is one half that required for F $(1/L)$. However, the periodicity of the crystal restricts sampling to integer multiples of $1/L$, which suffices for F but not for I. The polynomial I is thus not uniquely specified by crystal diffraction data, and hence there is no equivalent of $I = FF^*$ (I and F polynomials) to uniquely determine F in crystallography.

The information that ρ is bounded thus leaves the crystallographic phase problem twofold underdetermined, even with data to infinite resolution. Additional chemical information is clearly required to uniquely define ρ, but the amount needed is unknown since currently there is no satisfactory analytical theory for uniqueness in the presence of even very simple chemical constraints such as positivity (Millane, 1990).

The question of uniqueness in the presence of chemical constraints can be probed using a crude computational approach. The multiplicity of solutions for a given set of constraints can be very roughly estimated by seeking solutions in the neighborhood of random points in phase space. This type of approach of course can only demonstrate (by construction) non-uniqueness, but it may be useful in identifying problems which are too poorly determined for the true (physical) solution to stand out from multiple solutions satisfying all the constraints.

Atomicity

The single most important piece of chemical information which has been applied to the crystallographic phase problem is atomicity, as evidenced by the enormous success of classical direct methods. The assumption that a crystal is made up of atoms randomly placed in the unit cell leads to joint probability distributions of the structure factors (Hauptman & Karle, 1953; Klug, 1958). Insertion of the observed values of structure-factor amplitudes into the appropriate joint probability distributions leads to conditional probability distributions for the associated phases.

In small-molecule crystallography, the 'atomic' hypothesis makes the structure-determination problem greatly overdetermined [number of measured $|Fhkl| \gg 3(N-1)$, where N is the number of atoms]. For proteins, however, this is generally not the case. As atomicity still lies at the base of several recent approaches to the macromolecular phase problem, including Bricogne's use of the saddle-point approximation to obtain improved conditional probability distributions of very large numbers of structure factors (Bricogne, 1984), it is useful to investigate the extent to which atomicity limits the number of possible solutions for macromolecular diffraction problems. The algebraic minimum of Fourier magnitudes [$3(N-1)$ observations] is clearly not enough to uniquely determine the positions of N atoms, even for the very simple problem of three atoms in one dimension (Hauptman & Karle, 1951).

To investigate the power of atomicity in conjunction with varying amounts of diffraction data, solutions satisfying the Fourier constraints were sought in the neighborhoods of randomly generated distributions of atoms. The required manipulations were

Table 1. *Multiple distributions of atoms satisfying the same diffraction amplitudes*

This experiment utilized the 1.0 Å resolution structure of pancreatic trypsin inhibitor (5PTI, obtained from the Protein Data Bank) and the space group ($P2_12_12_1$) and unit-cell dimensions (74.1, 23.4, 28.9 Å) of native PTI crystals (Wlodawer, Walter, Huber & Sjölin, 1984). The 454 non-hydrogen atoms of PTI were placed randomly within the asymmetric unit of the crystal. The cost $C = \sum(|F_{true}| - |F_{calc}|)^2$ was then minimized with respect to the atomic positions using 120 cycles of *X-PLOR's* Powell conjugate-gradient minimizer. As the transformations $\rho(x) \rightarrow \rho(-x)$ and $\rho(x) \rightarrow \rho(x + \mathbf{T})$, where **T** is a vector between two distinct origins, change the phases but not the magnitudes of the structure factors, the phase error between the native and minimized structures was calculated for all possible choices of origin and enantiomorph. The phase errors listed in columns 3 and 4 are for the origin and enantiomorph giving the lowest overall phase error. $|F_{true}|$ and φ_{true} are amplitudes and phases calculated from the native distribution (PTI), $|F_{min}|$ and φ_{min} are calculated from the final minimized distributions, and φ_{start} are phases calculated from the starting random distributions. Phase errors are in °.

Resolution cut-off (Å)	$100\langle\lvert F_{true} - F_{min}\rvert\rangle / \Sigma F_{true}$	$\langle\varphi_{true} - \varphi_{min}\rangle$ (all reflections)	$\langle\varphi_{true} - \varphi_{min}\rangle$ (20.0–6.0 Å)	$\langle\varphi_{start} - \varphi_{min}\rangle$ (all reflections)	$\langle\varphi_{start} - \varphi_{min}\rangle$ (20.0–6.0 Å)	R.m.s. deviation (Å) (starting *vs* final coordinates)
3.0						
1	2.6	84.4	85.4	60.1	29.0	2.19
2	2.1	84.1	93.7	60.7	35.1	1.92
3	1.9	85.9	88.2	56.2	31.5	1.79
4	2.2	86.8	88.5	59.5	30.6	2.06
5	2.8	86.8	88.5	59.5	30.6	2.06
2.5						
1	12.9	85.3	83.9	67.6	29.8	1.90
2	12.2	87.4	88.0	69.3	30.2	1.86
3	12.8	86.7	85.3	65.5	35.0	1.93
4	12.1	87.4	86.6	67.5	34.2	1.83
5	11.8	86.0	80.4	65.1	27.8	1.94
2.0						
1	23.7	87.2	73.4	73.6	34.3	1.69
2	23.6	88.0	80.1	73.4	27.1	1.73
3	24.0	88.0	80.1	74.6	29.7	1.76
4	23.2	88.9	83.0	74.2	34.3	1.77
5	23.6	88.0	84.8	74.6	24.9	1.79
1.5						
1	31.6	88.4	86.2	83.4	25.7	1.64
2	32.1	88.3	90.4	80.6	23.1	1.54
3	31.9	88.8	89.8	82.3	27.6	1.67
4	31.4	88.7	81.8	92.8	28.0	1.65
5	32.0	88.1	85.8	83.1	25.7	1.64

conveniently carried out within the framework of Axel Brünger's program *X-PLOR* (Brünger, 1992). Random distributions were generated by placing the 454 atoms of the small protein pancreatic trypsin inhibitor (PTI) at random positions in the unit cell. The cost function $C = \sum(|F_{true}| - |F_{calc}|)^2$, where $|F_{true}|$ are Fourier amplitudes generated from the native configuration of atoms, was then minimized for each distribution. To investigate the relationship between the multiplicity of solutions and the amount of Fourier data, the procedure was performed with high-resolution cut-offs ranging from 3.0 to 1.5 Å.

In the experiment described in Table 1, five random distributions were generated and minimized for each high-resolution cut-off. Minimization reduced the *R* factor $(\sum||F_{true}| - |F_{calc}||/\sum|F_{true}|)$ from a starting value of ~56% for each random distribution to between ~2% for the 3.0 Å trials and ~31% for the 1.5 Å trials (Table 1, column 2). The root-mean-square displacement of the refined atomic coordinates from the random start was only ~1.8 Å (Table 1, column 7). The phases calculated from the minimized coordinates correlated with those calculated from the starting coordinates, particularly at low resolution (Table 1, columns 5 and 6). Low *R*-factor solutions are thus found within a reasonably small neighborhood of each of the starting random distributions. To assess the similarity between the minimized distributions and the 'native' configuration of atoms, the phase error was calculated after translation to each of the eight origins of the space group ($P2_12_12_1$) with and without inversion of coordinates. Importantly, the phase error for all of the minimized distributions in all positions relative to the native coordinates was close to the random value of 90°. There was no similarity between the native and minimized phase sets even at low (20.0–6.0 Å) resolution (Table 1, columns 3 and 4). Finally, to determine whether the different minimized distributions were distinct solutions, the phase error (for all possible choices of origin and enantiomorph) was calculated between each pair of distributions. As shown in Table 2, the phase sets for six minimized distributions were completely different from each other and from the true distribution.

There are two striking features of the data presented in Table 1. (i) A low *R*-factor solution was found in the neighborhood of each of the random distributions generated when the Fourier data was cut off at 3.0 or 2.5 Å resolution. As the errors in macromolecular diffraction experiments are such that *R* factors for refined structures are often above 15%, the fit between these 'false' solutions and the diffraction data is as good as might reasonably be expected for the true solution. Thus, with a diffraction data cut-off of 2.5 Å resolution, there are a very large number of solutions satisfying atomicity and the Fourier amplitudes within reasonable experimen-

Table 2. *Phase error between distinct low R-factor solutions at* 2.5 Å *resolution*

Six random distributions of atoms were minimized to fit amplitudes calculated from native PTI as described in Table 1 using a high-resolution cut-off of 2.5 Å. The final R factors were 12.8, 12.0, 12.2, 11.9, 12.6, 12.0% for trials 1–6 respectively. The weighted phase error between each pair of distributions was calculated for all choices of origin and enantiomorph. The entries above the diagonal are averages over the entire resolution range, the entries below the diagonal, averages over the range 20.0–6.0 Å. As in Table 1, the errors are for the configuration having the lowest overall phase error.

	PTI	1	2	3	4	5	6
PTI	0.0	86.3	87.6	86.7	85.1	87.6	85.5
1	76.1	0.0	87.7	86.1	88.0	85.9	86.5
2	87.5	87.3	0.0	85.4	87.8	88.0	85.0
3	86.4	80.6	74.2	0.0	87.1	87.1	85.4
4	81.5	89.0	87.0	85.1	0.0	88.1	86.6
5	84.1	90.4	94.1	90.3	79.4	0.0	86.1
6	85.1	89.5	84.5	90.7	90.0	91.8	0.0

tal error. The joint probability distributions used in direct methods and their extensions will thus have a very large number of false maxima. (ii) The statistics for each of the five trials within a single resolution range were virtually identical. Minimization of 75 additional random distributions using a 2.5 Å resolution data cut-off resulted in R factors ($\simeq 12\%$) and mean coordinate shifts very similar to those of the five 2.5 Å trials listed in Table 1. The surface defined by the cost function C thus appears to be quite regular with local minima of constant depth within the vicinity of any random point in coordinate space. The important parameter describing the surface is presumably the ratio of data (the Fourier amplitudes) to free parameters (the atomic coordinates). For PTI, the ratio is 0.86, 1.45, 2.75 and 6.32 at 3.0, 2.5, 2.0 and 1.5 Å resolution, respectively. In a similar experiment with the larger protein ribonuclease (124 residues), the same dependence of the R factor after minimization on the resolution of the data was observed, with large numbers of false solutions having R factors less than 12% found when the Fourier data were truncated below 2.5 Å (at 2.5 Å resolution, the ratio of data to parameters is 1.47).

Solvent boundary

The information that protein crystals contain relatively featureless solvent regions has been succesfully exploited for phase refinement (Wang, 1985). The constraint of a solvent boundary may also be incorporated into the joint probability distributions of structure factors arising from atomicity (Bricogne, 1988). To investigate the power of the combined constraints of a solvent boundary and atomicity in reducing the degeneracy of the phase problem, random distributions of atoms were generated and minimized as described above except that the atoms were additionally constrained to be within an envelope generated from the native structure. For this experiment we used the structure and crystal form of apolipoprotein E (1LPE; Wilson, Wardell, Weisgraber, Mahley & Agard, 1991) because solvent regions comprise an appreciable portion ($\sim 50\%$) of the unit cell. As described in Table 3(*a*), low R-factor solutions were found for each of the nine starting random distributions using data to 3.0 Å resolution [(number of reflections)/$3(N-1) = 1.2$]. The true structure and one of the minimized distributions are compared in Fig. 1. The atoms of the native structure (×) and of the minimized distribution (o) fall within the same regions of the asymmetric unit, but inside these regions they are completely uncorrelated.

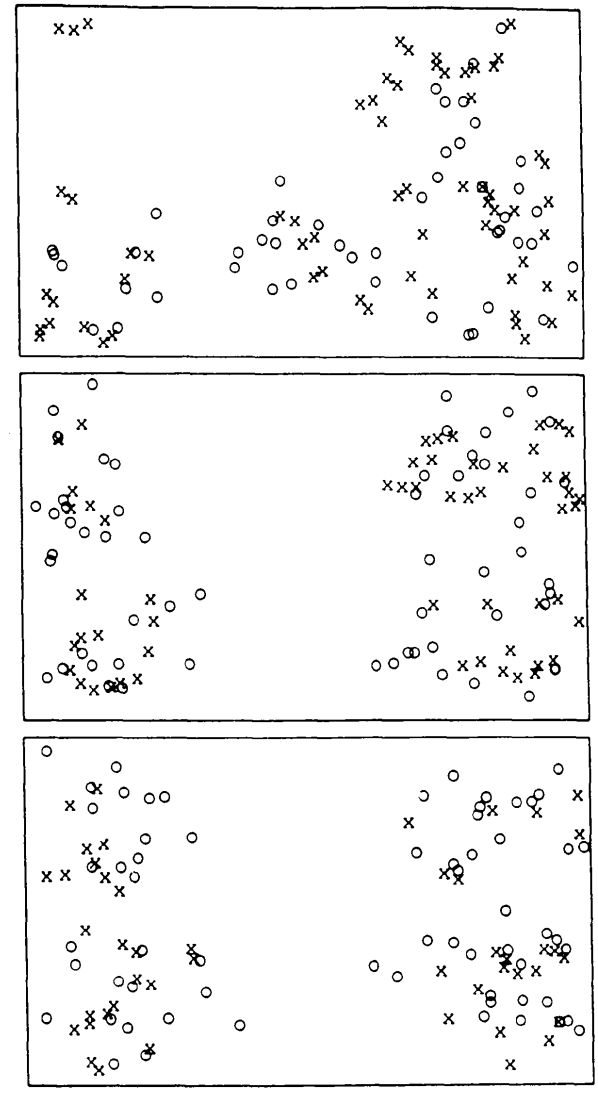

Fig. 1. Comparison of native and non-native distributions of atoms. Three representative 1 Å thick sections of the native (1LPE) distribution and the low R-factor minimized distribution described in line 3 of Table 3 are shown. × represents the position of an atom in 1LPE and o represents an atom in the non-native minimized distribution.

Table 3. *Multiple distributions of atoms fitting the native Fourier amplitudes and solvent boundary*

This experiment utilized the structure and crystal form ($P2_12_12_1$, $a = 40.65$, $b = 53.96$, $c = 85.45$ Å) of the N-terminal domain of apolipoprotein E (1LPE), an elongated four-helix bundle of 144 residues (Wilson *et al.*, 1991). An envelope was generated from a map of the native protein using Wang's algorithm (Wang, 1985) with a solvent content of 50%. The 1172 heavy atoms of apolipoprotein E were placed randomly inside the envelope. The random distributions were then minimized to fit the native Fourier amplitudes as described in the legend to Table 1. To retain the atoms inside the envelope, the cost function C was supplemented with a term $\sum|\varphi_{true} - \varphi_{calc}|$ where the sum is over reflections below 14 Å resolution. Specification of the envelope removed the ambiguity in origin and enantiomorph simplifying the calculation of the phase difference between transforms of the true and minimized distributions.

(*a*) Comparison of nine low R-factor solutions with the true solution and the starting random distributions

Trial	$100\langle\lvert F_{true} - F_{min}\rvert\rangle / \sum F_{true}$	$\langle\varphi_{true} - \varphi_{min}\rangle$ (all reflections)	$\langle\varphi_{true} - \varphi_{min}\rangle$ (20.0–14.0 Å)	$\langle\varphi_{start} - \varphi_{min}\rangle$ (all reflections)	$\langle\varphi_{start} - \varphi_{min}\rangle$ (20.0–14.0 Å)	R.m.s. deviation (Å) (starting *vs* final coordinates)
1	14.1	86.6	18.8	63.7	17.5	2.18
2	13.8	88.0	19.1	67.5	15.8	2.45
3	15.2	89.0	22.2	63.7	12.7	2.31
4	14.1	89.2	22.8	66.9	14.8	2.42
5	12.7	88.4	22.5	66.7	15.8	2.48
6	13.1	88.3	16.5	65.8	20.3	2.51
7	13.8	88.1	15.2	65.5	10.3	2.49
8	16.3	85.9	16.4	62.4	13.3	2.00
9	15.0	89.0	22.4	63.7	15.1	2.21

(*b*) Overall phase error between the nine low R-factor solutions

	2	3	4	5	6	7	8	9
1	86.5	86.9	86.7	88.1	87.1	89.0	86.8	86.7
2	—	86.5	87.6	88.6	90.1	87.7	89.9	86.4
3	—	—	86.9	87.2	86.3	88.3	87.2	85.8
4	—	—	—	85.5	88.0	88.2	86.3	86.5
5	—	—	—	—	88.9	87.0	89.3	87.3
6	—	—	—	—	—	87.0	89.2	86.9
7	—	—	—	—	—	—	84.8	87.5
8	—	—	—	—	—	—	—	88.3

Table 4. *Multiple low R-factor maps having higher entropy than the true solution*

Input maps were generated using amplitudes (to 2.5 Å resolution) calculated from PTI and phases calculated from either PTI or one of the six minimized random structures described in Table 2. Maps were modified according to equation (7) and new phases were calculated by inverse Fourier transformation. These phases were combined with PTI amplitudes to generate a new map which was again modified using (7). Convergence was reached in 8–12 cycles. The R factor and relative entropy $[-\sum\rho\ln(\rho/\mu)$ where μ is the uniform distribution and ρ is normalized such that $\sum\rho = 1]$ after the 12th cycle are listed in the table.

Input map	Starting entropy	Final entropy	Final R factor (%)
PTI	−0.457	—	0.6
MIN1	−0.474	−0.443	4.9
MIN2	−0.477	−0.440	5.0
MIN3	−0.478	−0.447	4.9
MIN4	−0.477	−0.446	4.8
MIN5	−0.483	−0.444	5.0
MIN6	−0.481	−0.441	5.0

The envelope constraint led to a low phase error at low resolution, but at medium and high resolution the phases of the true and minimized distributions were uncorrelated (Table 3*a*, columns 3 and 4). The phase sets for the different low R-factor solutions were completely unrelated at all but very low resolution (Table 3*b*). Thus, the combination of atomicity and a solvent boundary does not suffice to remove the phase ambiguity for moderate resolution data sets.

Entropy

Several recent approaches to the phase problem have been based on maximum-entropy methods. The maximum-entropy formalism is convenient as it ensures an everywhere-positive electron density. It is not always obvious whether additional chemical information (beyond positivity) is being applied in these methods. Two exceptions are the previously alluded to approach of Bricogne (Bricogne, 1984), in which maximum entropy enters as a consequence of atomicity, and the approach of Prince & Sjölin (Sjölin, Prince, Svensson & Gilliland, 1991) in which only the non-negativity property is utilized. A vague notion of smoothness may be implicit in some of the other applications of the entropy principle to the phase problem.

The maximum-entropy map may be the most probable map given a limited amount of experimental data, but for the approach to be successful, enough data must be used to make it reasonably probable that the maximum-entropy map resembles the true map. The power of the 'entropy principle' was investigated using an approach similar to that used to investigate atomicity above. Maps were generated using phases calculated from the random atomic distributions described in Table 2 and the 'native' diffraction amplitudes $|F_{true}|$. The entropy of each of the maps was maximized using the very simple iterative density-modification procedure described by Harrison (Harrison, 1989). A single cycle of this procedure consists of replacing the electron density ρ by the modified Newton's method update:

$$\rho' = \rho - \min[\rho, \langle\rho\rangle/e + (\rho - \langle\rho\rangle/e)/8][\ln(\rho/\langle\rho\rangle) + 1], \quad (7)$$

calculating new phases, and combining them with the $|F_{true}|$ to produce a new electron-density map. As shown in Table 4, this simple algorithm was quite effective in increasing the entropy of all of the maps while keeping the R factor below 5%. The entropies of the six 'false' maps prior to maximization were all lower than the entropy of the true map, but after maximization all of the false maps had higher entro-

pies than the true solution. The entropy is thus a very poor measure of the quality of a particular phase set for diffraction data at 2.5 Å resolution containing even very small amounts of error. However, at 1.25 Å resolution application of the entropy-maximization algorithm failed to generate maps with higher entropies than the true map, presumably due to the lack of high frequency noise in the true map.

Positivity

A similar approach was used to investigate the power of non-negativity. This constraint becomes more powerful at higher resolution since the negative ripples in the true map caused by Fourier series truncation effects are significantly reduced. 1.25 Å resolution maps were generated using phases calculated from the same six random distributions and the 'native' $|F_{true}|$. The F_{000} term was chosen such that the lowest density in the true map was 0.0. The standard 'error reduction' algorithm used in optics – alternate projection onto the space of positive maps and the space of maps satisfying the Fourier constraints – was then applied to each of the maps. This algorithm and slightly modified versions of it have had spectacular success in reconstructing images from Fourier amplitudes and random starting phases (Dainty & Fienup, 1987). As described in Table 5, everywhere-positive maps having low R factors were found in the neighborhood of each of the six random starting positions. Again, the phases of each of these positive solutions showed no correlation with the true phases. Thus positivity alone is a relatively poor constraint, even at 1.25 Å resolution.

These experiments suggest that at 2.5 Å resolution there are an extremely large number of distinct maps satisfying the Fourier data and the constraints of positivity and atomicity and having entropies higher than the true solution. The non-convexity of the Fourier amplitude constraint necessitates a multisolution approach to the phase problem (Bricogne, 1984; Luenberger, 1984). However, in order to be workable, the number of different solutions tracked by a branching algorithm must be kept to a manageably small number.

Clearly, a successful *ab initio* approach to the phase problem must make use of Fourier data to well beyond 2.5 Å resolution and/or more chemical information. The common strategy of slowly adding on reflections to a small low-resolution basis set has a high probability of ending up in one or several of the large number of false solutions at intermediate (2.5 Å) resolution. Bricogne's concept of likelihood (Bricogne, 1984, 1988; Bricogne & Gilmore, 1990) provides a way of utilizing high-resolution Fourier data even at early stages. The likelihood of a particular assignment of phases within a basis set is proportional to the conditional probability, given these assignments, of the observed amplitudes outside of the basis set. Additional chemical information that can be brought to bear on the problem includes partial structure information, non-crystallographic symmetry and knowledge of the density distribution in the form of a standard map histogram. The first two types of information may be incorporated into the likelihood formalism (Bricogne, 1988), but they are not always available in *ab initio* problems.

Table 5. *Everywhere-positive* 1.25 Å *resolution maps fitting native Fourier amplitudes*

Input maps were subjected to an iterative density-modification procedure similar to that described in the legend to Table 4 except that the high-resolution cut-off was 1.25 Å (instead of 2.5 Å) and the density-modification step consisted solely of setting negative density to 0.0. The R factor after ten cycles is listed in the table. Connectivity (here the fraction of nodes in the largest graph) was calculated as described in the text using a computer program available from the authors.

Input map	Final R factor (%)	Connectivity
5PTI	2.1	0.97
MIN1	6.1	0.05
MIN2	6.5	0.08
MIN3	6.3	0.04
MIN4	6.6	0.05
MIN5	6.8	0.03
MIN6	6.6	0.05

Connectivity

The macromolecular phase problem becomes greatly overdetermined once it is possible to trace an atomic model through the density and thus make use of the detailed rules of stereochemistry. Unfortunately, in general this is only possible at a relatively late stage in the solution of a macromolecular structure. However, the information that proteins are made up of connected chains of atoms can be exploited much earlier. We have developed a simple rapidly calculable (a fraction of a second on a VAX 9000) measure of the connectivity of a map. Grid points having density greater than 1.4 standard deviations above the mean are connected by edges if they are nearest neighbors (one grid unit away in either x, y or z). Two grid points belong to the same graph if they are connected by a continuous set of edges. Connectivity is measured as the fraction of grid points above the density cut-off which belong to the largest graph or, alternatively, as the number of distinct graphs.

The use of connectivity as a figure of merit breaks the multiple-solution ambiguity in the problems described above. For example, the everywhere-positive but false solutions at 1.25 Å have scores ranging from 0.03 to 0.08, while the true map has a score of 0.97 (Table 5). The same sharp differentiation between the true and false solutions also occurs at 2.5 Å resolution and for the distributions of atoms

within the native envelope described in Table 3 and Fig. 1. In order to be generally useful as a figure of merit, the connectivity must decrease smoothly with increasing phase error; a partially degraded version of the true map shoud be partially connected. As shown in Fig. 2, addition of increasing amounts of random phase error to a perfect map resulted in a monotonic decrease in the connectivity. This correlation of connectivity with phase error was observed even at very low (12 Å) resolution (data not shown).

Unfortunately, the inability to calculate derivatives of the connectivity rules out the use of efficient search algorithms for varying the phases to maximize the connectivity of a map. A Monte Carlo approach ran into difficulties due to the size of the search space at all but very low resolution ranges. A more promising application of connectivity may be as a figure of merit for evaluating and pruning the branches of a multisolution search tree.

Measurement of the connectivity of the electron density may also be a powerful means to identify the correct solution from a large number of phase sets generated using classical direct methods. Woolfson & Yao (1990) found that application of the program *SAYTAN* to random phase sets for a small protein led to mean phase errors of less than 45° in six out of 1000 trials. However, standard figures of merit, such as those used in *MULTAN* and *SAYTAN*, were not effective in recognizing the good phase sets. They concluded, 'The full exploitation of direct methods to solve protein structures awaits the discovery of a new figure of merit more effective in ranking trial phase sets.' The possibility that the connectivity of the electron density is such a figure of merit should be investigated.

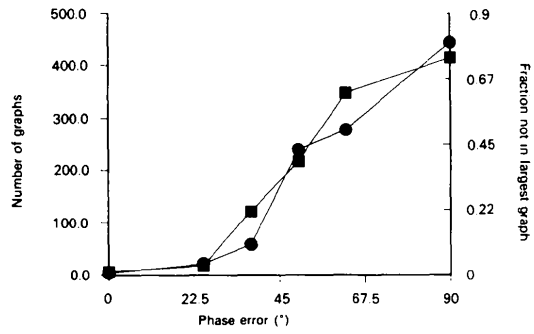

Fig. 2. Correlation between phase error and connectivity. Increasing amounts of random error were introduced into phases calculated from the PDB coordinates of apolipoprotein E (1LPE). Maps were generated using amplitudes calculated from the PDB coordinates and the random-error-containing phases. The connectivity of each of the maps was assessed using the algorithm described in the text. Square symbols represent the number of graphs, and circles, the fraction of nodes in the largest graph. Similar results were obtained for all resolution ranges tested; the results shown are with low- and high-resolution cut-offs of 20.0 and 3.0 Å respectively.

Concluding remarks

The chemical constraints employed by current approaches to the *ab initio* phase problem in macromolecular crystallography are not sufficient to limit the solutions to a manageably small number for data sets of moderate resolution. Stronger chemical constraints, such as connectivity, will be required to compensate for the low ratio of data to free parameters intrinsic to the macromolecular phase problem.

This work was supported by funding from the Howard Hughes Medical Institute. DB is an HHMI Fellow of the Life Sciences Research Foundation.

References

BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410–445.
BRICOGNE, G. (1988). *Acta Cryst.* A**44**, 517–545.
BRICOGNE, G. & GILMORE, C. J. (1990). *Acta Cryst.* A**46**, 284–297.
BRÜNGER, A. (1992). *X-PLOR Manual.* Verson 3.0. Yale Univ., USA.
DAINTY, J. C. & FIENUP, J. R. (1987). In *Image Recovery: Theory and Application*, edited by H. STARK. New York: Academic Press.
FIDDY, M. A. (1987). In *Image Recovery: Theory and Application*, edited by H. STARK. New York: Academic Press.
HARRISON, R. W. (1989). *Acta Cryst.* A**45**, 4–10.
HAUPTMAN, H. & KARLE, J. (1951). *Acta Cryst.* **4**, 383.
HAUPTMAN, H. & KARLE, J. (1953). *Am. Crystallogr. Assoc. Monogr.* No. 3. Pittsburgh: Polycrystal Book Service.
HAYES, M. H. (1987). In *Image Recovery: Theory and Application*, edited by H. STARK. New York: Academic Press.
KLUG, A. (1958). *Acta Cryst.* **11**, 515–543.
LUENBERGER, D. G. (1984). *Linear and Nonlinear Programming.* Cambridge, MA: Addison-Wesley.
MILLANE, R. P. (1990). *J. Opt. Soc. Am.* **7**, 394–411.
SJÖLIN, L., PRINCE, E., SVENSSON, L. A. & GILLILAND, G. L. (1991). *Acta Cryst.* A**47**, 216–223.
WANG, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
WILSON, C., WARDELL, M. R., WEISGRABER, K. H., MAHLEY, R. N. & AGARD, D. A. (1991). *Science*, **252**, 1817–22.
WLODAWER, A., WALTER, J., HUBER, R. & SJÖLIN, L. (1984). *J. Mol. Biol.* **180**, 301–310.
WOOLFSON, M. M. & YAO, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.