Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 8 September 2000
Accepted 4 January 2001

© 2001 International Union of Crystallography
Printed in Great Britain – all rights reserved

An evolutionary computational approach to the phase problem in macromolecular X-ray crystallography

Gordon Webster*† and Rolf Hilgenfeld

Department of Structural Biology and Crystallography, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 Jena, Germany. Correspondence e-mail: gwebster@caregroup.harvard.edu

The *ab initio* computation of the molecular envelopes of two proteins exclusively from their corresponding diffraction amplitudes demonstrates that an efficient and inherently parallel evolutionary search algorithm can assist in the direct phasing of macromolecules for which almost no *a priori* structural information is available. The applicability of this evolutionary computational approach is general and should not be limited to the examples described nor to extremes of data resolution, symmetry or structural size.

1. Introduction

In X-ray crystallography, the Fourier transform of the three-dimensional distribution of electron density in the crystal is sampled by measuring the Fourier amplitudes scattered from the crystal during a diffraction experiment. Since no phase information is recorded with these amplitudes, the direct Fourier reconstruction of the crystal structure from its corresponding diffraction pattern is not trivial and other means of deriving this phase information must generally be used (Karle, 1989). This 'phase problem' is one of the major and rate-limiting steps in macromolecular X-ray crystallography and generally requires recourse to further experimental methods such as multiple isomorphous replacement (MIR) (Watenpaugh, 1985) or multiwavelength anomalous dispersion (MAD) (Hendrickson *et al.*, 1988) in cases where no suitable structurally homologous model can be found for more direct phasing by molecular replacement (Rossmann & Blow, 1962). In theory, the Fourier reconstruction of electron density from amplitudes alone is an overdetermined problem if a sufficiently large number of these amplitudes are combined with constraints derived from some general prior knowledge of the structure, such as the condition that the electron density is everywhere positive (Hauptman & Karle, 1953). The development and application of this theory to the X-ray crystallography of small molecules has for some years now made the direct phasing of structures containing up to 100 atoms essentially routine (Karle & Karle, 1966). With the more recent addition to this field of maximum-entropy concepts (Bricogne, 1993), the extension of direct methods to successively larger structures continues to progress. The latest developments in direct methods, combining searches in both real (scattering model) and reciprocal (phase) space have extended the limit of *ab initio* phasing to structures as large as a few hundred atoms (Miller *et al.*, 1994; Schäfer *et al.*, 1996), however, such approaches are currently only applicable to macromolecular structures for which the ratio of observed to modelled parameters is atypically large [*i.e.* those structures for which a significant proportion of the very high ($\leq$1.2 Å) resolution structural amplitudes can be observed in the experimentally measured diffraction data (Hauptman, 1996)]. Most recently, however, the effective range of this approach has been extended to protein structures using data to 1.55 Å (Mukherjee *et al.*, 2000).

Some success with macromolecule structures at very low resolution (~80 Å) has been obtained by the use of approaches combining low-resolution scattering models with direct methods and entropy maximization (Podjarny & Urzhumtsev, 1997). Monte Carlo search methods, in which the configurational space of a scattering model is searched under the constraint of the solvent fraction and the restraint of the observed diffraction amplitudes, have also been shown to be effective at somewhat higher resolution (~15 Å) (Subbiah, 1991). The complexity of the search model in such methods is necessarily limited by the ability of the search algorithm to sample efficiently the configurational space of the model itself. This problem can be partially addressed by the use of successively finer search steps, in which a lower-resolution model from an earlier stage of the search can be used as a starting point for a subsequent stage of the search on a finer higher-resolution grid (Subbiah, 1991). A coarser initial search is also necessary to avoid stagnation of the algorithm in local minima as a result of the insensitivity of the crystallographic target function to small local differences between different

† Current address: Division of Experimental Medicine, Beth Israel Deaconess Medical Center, Harvard Institutes of Medicine, 4 Blackfan Circle, Boston, MA 02115, USA.

configurations of a partially accurate model (Kleywegt & Jones, 1995). Although such methods are not theoretically limited to very low resolution studies, at higher resolution with the accompanying exponential increase in the complexity of the model that is necessary to maintain adequate sampling of the diffraction Fourier transform, a different approach is required if the algorithm is to converge upon a solution in any reasonable time span or if it is to avoid the increasing likelihood of falling into local minima which may cause it to stagnate and prevent it from ever reaching any acceptable solution.

2. The genetic algorithm approach

We have developed a prototype evolutionary search method for the *ab initio* phasing of macromolecular structures from diffraction intensities based upon the paradigm of natural selection in biological systems. Evolutionary searches as a computational tool were first suggested as a means of searching the state space of systems too complex for traditional Monte Carlo or gradient search methods or for which there exists no satisfactory analytical description of the functional landscape (Holland, 1975). One of the most widely studied of the family of evolutionary methods is genetic algorithms, in which the set of parameters representing one possible configuration of the system is encoded in binary form as a single string of *ones* and *zeros* constituting a binary 'genome' that can be used as a blueprint to generate its corresponding trial configuration or 'phenotype'. In a computational analogy of natural selection, an initially random set of these binary genomes (genotypes) is used to generate a population of configurations (phenotypes) which are tested for their proximity to some given target function with each individual phenotype being assigned a corresponding score. A fitness function is then applied to these scores to weight a subsequent 'reproductive' cycle to favour the recombination of the genetic material from the more highly ranked phenotypes. (A clear distinction must be made between the target function that provides a measure of the divergence of each individual phenotype from an ideal model and the fitness function that determines the corresponding degree of selection advantage that is granted to the phenotype according to its score.) The recombination process consists of a weighted random exchange of one or more corresponding parallel sections of the aligned binary genomes of each pair of individuals selected from the population, analogous to the processes of genetic recombination that are observed to occur in biological systems (Holland, 1975). This weighted genetic recombination can be repeatedly applied to the subset of selected individual genotypes until a sufficient number of new genotypes has been created from which is generated a new population of the corresponding phenotypes that now bear the traits of the better-performing phenotypes from the previous generation. Provided (i) that a true isomorphism exists between the information contained in the genotypes and that of their corresponding phenotypes, and (ii) that a suitable target function for selection can be defined, iterative cycles of such testing and recombination will yield subsequent generations of phenotypes of increasing quality with respect to the target function by favouring the propagation of the genes from the 'fitter' phenotypes until some point of population convergence is reached (Holland, 1975).

The application of these ideas to the phase problem in macromolecular crystallography can now be considered. Since in nature proteins exist in a single enantiomeric form, they give rise predominantly to diffraction amplitudes with noncentrosymmetric phases. A typical protein diffraction data set at medium resolution consists of several thousands of such amplitudes and the possibility of assigning even partially correct phases to these by trial-and-error searching is inconceivable given the size and number of degrees of freedom of the phase space that must be searched. The use of a suitable real-space scattering model from which phases can be calculated is in itself a constraint on the bounds of the phase space to be searched since it limits the search to phase sets whose Fourier transforms are everywhere positive. The degree to which additional prior information about the model is incorporated will correspondingly further limit the phase space until, in the case where a structurally homologous molecular fragment is available as a search model, the search itself may consist of only a handful of parameters describing the orientation of the model within the unit cell (Rossmann & Blow, 1962). In fact, a genetic algorithm approach has already proved to be very successful in such cases and has been shown to be capable of solving macromolecular structures by molecular replacement (Kissinger *et al.*, 1999) as well as some small-molecule structures from their powder diffraction patterns (Kariuki *et al.*, 1997). In certain other special situations, where the phase space can be limited by considerations of symmetry, as in the case of icosahedral viruses (Miller *et al.*, 1996) or by searching for a much smaller subset of phase-determining heavy atoms (Chang & Lewis, 1994), evolutionary algorithms that also model a limited set of state parameters have been tried. Unfortunately, however, the phase problem is most acute in cases where relatively little prior information about the model is available as is typical in macromolecular structure determination and it is to this more general class of problems that the method described here is addressed.

3. Proposed scheme

A general prototype scheme for the derivation of *ab initio* phases directly from structure amplitudes can be outlined as follows. Based solely upon considerations of the dimensions and symmetry of the crystal unit cell, a random population of scattering configurations (phenotypes) and their corresponding linear binary representations (genotypes) is generated. Each individual configuration consists of a regular three-dimensional grid that occupies the unique volume of the unit cell, upon which a number of uniform scatterers is arranged at random. The binary genotype has the same number of locations as the grid and can thus be considered to be a 'linearized' representation of it. Conversely, the phenotype grid can be considered to be a 'folded' three-dimensional representation

of the binary genotype. [It is worth noting that there exists a set of possible mappings between the genotype and its corresponding phenotype under which the isomorphism between these two entities is preserved and that the choice of a particular mapping (or way of 'folding' the phenotype) may influence the performance of the evolutionary algorithm. While this issue has been considered in the design of the algorithm presented here, a full discussion of it is beyond the scope of this article.] As an example, we can consider one way in which this regular three-dimensional grid of scatterers can be represented in the binary genome, choosing a value of *one* to correspond to a gridpoint occupied by a scatterer, with a *zero* representing an empty gridpoint. Clearly, more complex genomes that have a greater number of binary bits per grid location could also be used to represent more elaborate scattering models, although it was decided that a simpler model scheme would suffice for these preliminary experiments to test the applicability of evolutionary algorithms to the phase problem.

After the random population of scattering model phenotypes and their corresponding genotypes is generated, each individual phenotype is assigned a score by being tested against the set of observed diffraction amplitudes using a crystallographic target function. In the subsequent reproductive cycle, pairs of individual genotypes are selected from the population by the use of a weighted random distribution that favours the selection of genotypes whose corresponding phenotypes performed better against the crystallographic target function (the selection weights being computed using the previously mentioned fitness function whose purpose is to 'translate' the score of each phenotype into its corresponding selection probability). The binary genomes of these selected pairs are then recombined under the influence of a set of genetic operators that yield a recombinant offspring potentially bearing genetic material from both parents. The genetic operators are central to the recombination process by which the subsequent generation of individuals is created in each reproductive cycle (Holland, 1975). Three such operators, *crossover*, *inversion* and *mutation* (shown in Fig. 1), are currently implemented in our computational evolutionary search scheme and are also applied according to a weighted random distribution with the weights in this case being supplied by the user at the outset of the computational run. The reproductive step yields a new generation of genotypes from which a fresh population of phenotypes is produced, equal in number to the previous one. This kind of 'steady-state' evolutionary algorithm, in which a constant population size is maintained, has been found to be, in general, computationally more expedient than those in which a freely proliferating population must be managed (Bäck *et al.*, 1997). The phenotypes of this new generation are then in their turn subjected to further cycles of testing, selection and reproduction in an iterative process in which the successive generations evolve towards a configuration that best represents the corresponding Fourier transform of the given set of diffraction amplitudes, provided that a suitable crystallographic target function is used. The rate and extent of this evolution also depends upon the choice of various parameters that govern the computational algorithm such as the fitness function used to weight the selection of the individual genotypes during the reproductive cycle and the weights that control the application of the different genetic operators.

Further consideration of the design of our evolutionary protocol was forced upon us by two initial problems that arose during the first trial runs with the prototype scheme. These were the difficulty and computational expense of calculating a suitable fitness function with which to weight the selection of the individual genotypes during the reproductive stage, and the premature convergence of the evolutionary process as a result of certain well performing phenotypes rapidly proliferating and dominating subsequent populations. Both of these problems were solved by further extending the biological paradigm of natural selection to include the notion of geographical restraints on reproduction (Connor, 1994). Instead of computing a fitness function with which to weight

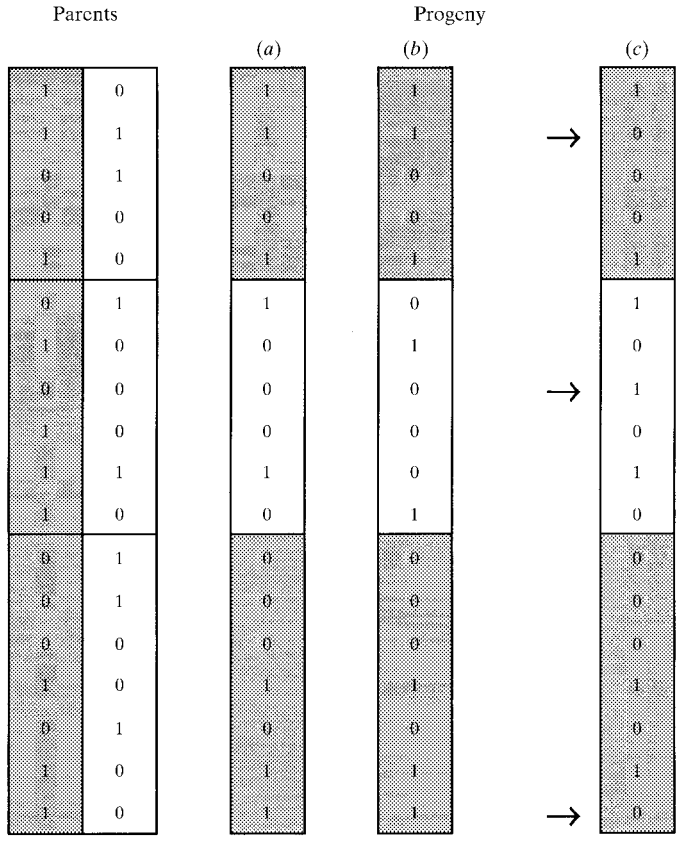

Figure 1
In this demonstration of the genetic operators used in the evolutionary algorithm, three of the many possible progeny genotypes that could result from the genetic recombination of the two parents are shown. (*a*) The effect of the *crossover* operator in which aligned parallel sections of the two genomes are exchanged. (*b*) The effect of the *inversion* operator in which one inherited section from the parental genotypes (in this case, the result of a previous crossover) is inserted as a reversed sequence in the progeny genotype during recombination. (*c*) The effect of the *mutation* operator (combined with a previous crossover step) in which a random fraction of the binary bits in the progeny genotype are toggled to their opposite values during recombination (indicated by adjacent arrows in the diagram).

the selection of individuals for recombination, the population of phenotypes is spatially distributed upon a virtual two-dimensional terrain over which each individual performs a limited random walk, sampling the population in its own locality and undergoing genetic recombination with the best performing phenotype from those it encounters. Under this new scheme, computation of a fitness function is no longer necessary since the choice of each individual's recombination partner from the population subset sampled during the random walk is decided only by its ranking according to the previously calculated crystallographic target function. This avoids both the necessity of making decisions about how best to weight each individual for genetic selection and the accompanying overhead of computing an elaborate fitness function (Connor, 1994). The fitness function itself can be a source of problems. For example, a fitness function that overvalues the differences in the distribution of phenotype quality will tend to lead to the premature convergence of the evolutionary process as previously described, rapidly giving rise to a homogeneous population with limited scope for further evolution. Conversely, a fitness function that undervalues these differences will lead to an unfocused search, hindering the evolution of the phenotype population and possibly leading to stagnation of the algorithm (Blickle & Thiele, 1996). The extended protocol described also has the effect of geographically limiting the convergence of the population. This last feature is particularly important, since in geographically separated regions of the virtual terrain it allows different partial solutions to evolve independently, which are then able to gradually coalesce into superior recombinant solutions in a process that resembles diffusion. In comparison with the former probability-weighted (or 'roulette-wheel') selection method (Bäck *et al.*, 1997) that was implemented in our previous algorithm, this new approach has proven to perform consistently better, allowing high-quality phenotypes to proliferate within the population without dominating it and driving the population to premature convergence.

In the light of this broad overview of the method, its individual components can now be discussed in more detail, followed by a demonstration of the application of these ideas to the *ab initio* determination of the molecular envelopes of proteins. The method as described can be conveniently divided into a number of distinct steps for the purposes of discussion.

(i) Based upon the chosen resolution limit, a regular grid of N_g points occupying the unique volume of the unit cell is set up and randomly populated with N_s (initially $N_g/2$) uniform scatterers. A random population of size N_p of these scattering model phenotypes is generated along with their corresponding binary genomes, with each member of the population being assigned to a point on a 'terrain' consisting of a regular two-dimensional grid of N_p unique locations with periodic boundary conditions applying (topographically equivalent to a toroidal surface).

(ii) For each individual model in the population, structure-factor amplitudes are calculated and then tested against the supplied set of diffraction amplitudes by the use of a crystallographic target function, in this case a standard correlation coefficient for observed and calculated structure factors given by

$$\frac{\sum ab - \left(\sum a \sum b\right)/N}{\left[\sum a^2 - \left(\sum a\right)^2/N\right]^{1/2}\left[\sum b^2 - \left(\sum b\right)^2/N\right]^{1/2}},$$

where $a = |F_o|^2$, $b = |F_c|^2$, with F_o and F_c being the observed structure amplitudes and those calculated from the scattering model phenotypes, respectively. The use of a correlation function as opposed to some other quality indicators such as the crystallographic residual is advantageous in this case since it is less sensitive to the relative scaling of the calculated and observed amplitudes, which is particularly important when dealing with structural models of low accuracy (Stout & Jensen, 1989, p. 230).

(iii) After the testing step, for each individual model in the population, a random walk of N_w steps on the two-dimensional population terrain is performed starting from that model's position on the terrain and sampling N_w members of the population in the geographical vicinity (the value of N_w being supplied by the user at the outset). This is followed by genetic recombination of the current model's own binary genome with that of the best-performing phenotype encountered in the random walk (as defined by the crystallographic target function in the previous testing step). For a given population N_p, too large a ratio of N_w to N_p leads to a more rapid initial evolution of the population but also to premature convergence since the population quickly becomes too homogeneous, the lack of diversity caused by the over-sampling limiting its potential for further development.

(iv) The reproductive step consists effectively of aligning the pair of selected genomes and creating a 'daughter' genotype by recombining elements from each in a weighted random fashion by the use of the genetic operators (see Fig. 1) (Holland, 1975). First the genome of the current phenotype is aligned with that of its breeding partner selected during the random walk. Both are then randomly divided into a number of equivalent parallel sections, each pair of which is subjected to a combination of crossover (in which it is exchanged for the equivalent parallel section in the other genome) and/or inversion (in which its sequence order is reversed within the same genome) or is left unchanged, according to the probabilities for crossover P_c and inversion P_i which are supplied by the user. In this way, the recombinant 'daughter' genotype is produced, the phenotype of which is generated and replaces the current parent if it proves to perform better than the parental phenotype in the subsequent testing step. A background of random mutation is also applied, in which some fraction of the binary *ones* and *zeros* of the genome are toggled to their opposite values. The mutation operator is extremely important for ensuring that all of the vast potential search space that must be sampled by the population of N_p binary genomes is accessible to the evolutionary algorithm. The mutation rate M_i for the current generation G_i in a run of G generations is controlled by three parameters which are also supplied by the user: M_a the initial mutation rate; M_z the final

mutation rate and M_s the slope parameter that governs the rate of change of the background mutation level throughout the run, where

$$M_i = M_a + [(G_i/G)^{M_s}(M_z - M_a)].$$

$M_s = 1$ will give a linear rate while $M_s > 1$ or $M_s < 1$ will give an accelerating or decelerating exponential rate, respectively. In this way, the background mutation rate can be gradually reduced from an initially high level that maximizes the population diversity and its evolutionary potential to a much lower final level that introduces little or no noise into the directed evolutionary search carried out by the genetic operators and consolidates the evolutionary gains achieved during the run. Like the mutation operator, the inversion

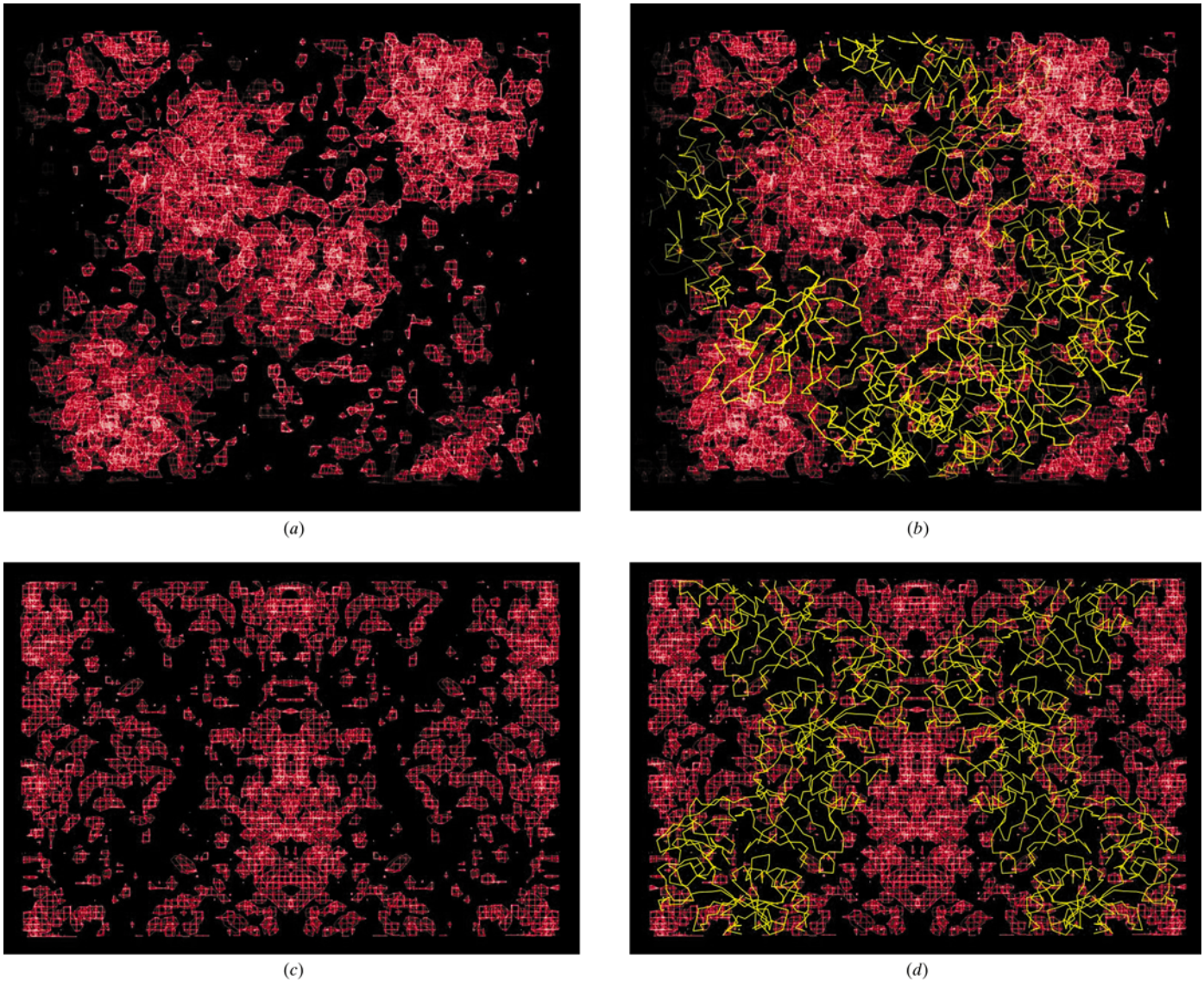

Figure 2
(*a*) The negative-contrast electron-density map obtained with the diffraction amplitudes of the tetragonal form of thermolysin [molecular weight = 34.6 kD, 1 molecule/asymmetric unit, solvent content = 66%, space group $P4_12_12$, $a = b = 98.0$, $c = 108.0$ Å (M. S. Weiss, personal communication)] after 500 iterations of the evolutionary algorithm (correlation before density modification = 0.71) and enhanced by 40 cycles of density modification with phase combination (Podjarny & Rees, 1994). (*b*) The actual thermolysin structure superimposed on the evolved electron-density map. (*c*) The negative-contrast electron-density map obtained with the diffraction amplitudes of elongation factor Tu in complex with a nucleoside triphosphate analogue (Berchtold *et al.*, 1993) (molecular weight = 44.6 kD, 1 molecule/asymmetric unit, solvent content = 63.7%, space group *C*2, $a = 151.38$, $b = 99.95$, $c = 40.45$ Å, $\beta = 94.89^\circ$) after 500 iterations of the evolutionary algorithm (correlation before density modification = 0.66) and enhanced by 40 cycles of density modification with phase combination (Podjarny & Rees, 1994). (*d*) The elongation factor Tu structure (Berchtold *et al.*, 1993) superimposed on the evolved electron-density model. Owing to the origin ambiguities that occur in polar space groups, the original structure was displaced along the *b* axis relative to the evolved electron-density model by ~17 Å, and a corresponding correction was made to allow for this when overlaying the two. Since no phase information from this known structure was used, there is no requirement that the same *b*-axis origin should also apply in the structure yielded by the algorithm. Indeed, the difference in the two origins confirms the absence of any phase bias that would have been introduced had any structural information from the known model been used. It should be noted that all of the ambiguities of origin and (in the absence of anomalous-scattering data) enantiomeric hand that are encountered with more traditional crystallographic phasing methods also apply here and must be accounted for when comparing different evolved models (Drenth, 1994). For both structures, the entire crystal unit cell is shown.

Table 1
Resolution-dependent breakdown of data completeness for the diffraction data used in the testing of the evolutionary algorithm.

	Data completeness in resolution range (%)	
Resolution range (Å)	Elongation factor Tu	Thermolysin
100.00–8.55	91.2	77.4
8.55–6.79	99.8	99.2
6.79–5.93	99.8	99.4
5.93–5.39	99.6	99.4
5.39–5.00	99.4	99.4
Overall completeness (%)	97.9	94.5
Total unique reflections	2586	2388

operator is designed to allow the algorithm access to all possible states of the model system and is normally applied with very low probability.

Steps (ii), (iii) and (iv) described above are carried out in an iterative fashion over a given number of generations G defined by the user at the outset. The best performing phenotype is periodically output by the program as a three-dimensional coordinate file containing the most evolved distribution of scatterers in the unique volume of the unit cell at the current stage of the evolutionary search, as defined by the crystallographic target function.

4. Feasibility of the method evaluated with test cases

In order to test the feasibility and performance of such an approach, we used only the diffraction amplitudes to a resolution of 5.0 Å from the reflection data sets of two known protein structures and processed them using the iterative computational scheme described above. It was decided that the initial tests of the algorithm's ability to extract structural information from diffraction amplitudes would be made at relatively low resolution, since it allows a simple binary protein/solvent model to be used to demonstrate the feasibility of this approach and the determination of a protein envelope at low resolution is, in any case, often the first step in crystallographic phase determination (Watenpaugh, 1985). It is important to note that at no time was any other structural information from these two proteins used, except in the subsequent assessment of the results. For each of these proteins, initial random populations of 100–500 trial scattering model phenotypes were generated, based solely upon considerations of the symmetry and dimensions of their crystallographic unit cells. The mean ratio of scatterers to empty grid points was 1:1 for this initial randomly generated population, corresponding to an average 50% solvent content. In the test cases described here, the mean solvent content in subsequent phenotype populations was unrestrained, allowing it to fluctuate freely with the weighted random recombination of the scattering model genotypes, although solvent content could potentially serve as a useful restraint in future implementations of this search method. In this preliminary study at relatively low resolution, the simple binary protein/solvent model scheme that was used to represent the molecular structure had the virtue of reducing the computational overhead that would be incurred using a more detailed scattering model while preserving sufficient structural information to model a molecular envelope at that resolution. Each population of models was then subjected to alternate cycles of testing (comparing calculated and observed diffraction amplitudes) and recombination until it was considered that the optimization of the scattering model population with regard to the crystallographic target function was approaching convergence. In the initial computational tests described below, convergence was typically approached after 500–1000 generations as judged by an analysis of the rate of change of the population's mean deviation from ideality (defined by the crystallographic target function). At that point, electron-density maps were calculated based upon phases derived from the highest ranked configuration of the final generation.

Using experimentally measured diffraction data to a resolution of 5.0 Å, a population terrain consisting of 100 to 500 individual configurations which was allowed to evolve over 500 to 1000 generations proved to be sufficient to produce well defined images of the molecular envelopes of the two typical medium-sized proteins that were chosen for testing [elongation factor Tu in complex with a nucleoside triphosphate analog (Berchtold *et al.*, 1993), space group $C2$, $a = 151.38$, $b = 99.95$, $c = 40.45$ Å, $\beta = 94.89^\circ$, data 97.9% complete for all reflections to 5.0 Å (see Table 1); and thermolysin, tetragonal form (M. S. Weiss, personal communication), space group $P4_12_12$, $a = b = 98.0$, $c = 108.0$ Å, data 94.5% complete for all reflections to 5.0 Å (see Table 1)]. For successful test runs like the two described here, typical values for the parameters governing the evolutionary algorithm were: $N_g = 500$, $N_w = 5$, $M_a = 0.001$, $M_z = 0.0$, $M_s = 1.0$. While our initial trials of the algorithm indicated that even larger populations of phenotypes seemed to be preferred (data not shown), population size was limited in these experiments owing to the increased computational overhead incurred when processing very large populations. It is interesting to note that the images that were obtained in the two examples described here turned out to be negative contrast images of the proteins for reasons that will be discussed subsequently. In these negative contrast images, the scatterers on each model grid move preferentially into areas occupied by solvent in the unit cell with the region occupied by the protein appearing as a void. The mean correlation coefficient (*before* any subsequent phase improvement procedures), close to zero for the initial random population, is typically improved to values lying in the range 0.5–0.7, where a value of 1.0 would indicate perfect correlation (Stout & Jensen, 1989, p. 230). For a traditional protein structure determination carried out by multiple isomorphous replacement (MIR) using estimated phases from two or more heavy-atom derivatives of the protein, electron-density maps derived from structure factors corresponding to correlation coefficients in this range are typically of sufficient quality to unambiguously reveal the molecular envelope of the protein (Watenpaugh, 1985).

5. Discussion

In spite of our electron-density maps being rather noisy and displaying negative contrast images of the proteins, in each case the boundary between the protein and the surrounding solvent could be distinguished (as shown in Fig. 2). Superpositions of the known structures of the protein test samples on their corresponding envelopes generated *ab initio* from the raw structure amplitudes confirm that this approach, even at this nascent stage of its development, displays great potential as a method of extracting structural information from X-ray diffraction data. The excellent three-dimensional correspondence between the generated envelopes and the atomic coordinates is certainly consistent with the high correlation coefficients that were obtained with the evolutionary algorithm and are at least as good as those generally achieved using more traditional crystallographic phasing methods such as MIR (Watenpaugh, 1985). We believe that the generated maps exhibit negative contrast (*i.e.* the protein appears as a void while the solvent regions appear as positive electron density) because the target correlation function compares the magnitudes of squared Fourier amplitudes (Stout & Jensen, 1989, p. 230). Since the information about the signs of these amplitudes is lost, a featureless *positive* image of the structure cannot be distinguished from a featureless *negative* image. Further, since real protein density is not featureless except at extremely low resolution but exhibits marked internal variation, the grid of uniform scatterers of which each configuration consists in this simple model scheme is better able to model the essentially featureless solvent region (Podjarny & Urzhumtsev, 1997). As a consequence, the scattering points are preferentially positioned in the solvent region of the structure during the course of the evolutionary search and a negative image of the protein evolves. It is also worth noting that it proved possible to further enhance the negative contrast of the electron-density maps by a few cycles of iterated real-space density modification and phase combination of the kind normally used to enhance the features of electron-density maps obtained by more traditional crystallographic phasing methods (Podjarny & Rees, 1994). Since in the two experiments described negative-contrast electron-density models are obtained, the normal solvent-flattening operation used in density modification and phase combination is here being applied to the void electron-density region representing the protein volume, further flattening it and increasing the negative contrast of the molecular envelope of the protein.

The resulting electron-density maps (shown in Fig. 2) were obtained in each case after an evolutionary search of about 2 days duration (including the few minutes of CPU time required for the subsequent solvent-flattening procedure) on a DEC Alpha 3000/700 workstation, clearly demonstrating the ability of this algorithm to extract structural information from a set of diffraction amplitudes in a reasonable time span, even using a relatively modest computational platform. The success of this approach undoubtedly rests upon the remarkable ability of evolutionary algorithms to search the vast potential space of possible models in both an efficient and an inherently *parallel* manner, which is of particular importance in crystallographic applications given the general insensitivity of most crystallographic quality indicators to small local changes in a partially accurate scattering model (Kleywegt & Jones, 1995). The efficiency of the algorithm is strikingly apparent when one considers the number of possible states of the scattering model that would have to be searched by trial-and-error methods to yield a reasonable probability of finding a correct solution. A suitable scattering model for a typical protein at 5.0 Å resolution would consist of a grid of several thousand points. Assuming that, for any single trial configuration, approximately half of those points could be occupied by a scatterer with the other half remaining empty (corresponding to the 50% solvent fraction that was used in the cases described here), the number of possible configurations, even for this relatively simple model, is enormous. The actual number of possible model configurations for n grid points and k identical scatterers can be computed using a standard combinatorial formula (Forthofer & Lee, 1995)

$$ {}_nC_k = n!/[k!(n-k)!] $$

and, for the examples described here, yields numbers that are astronomical in comparison with the 50000–100000 configurations that were actually sampled in the evolutionary search protocol (for the elongation factor Tu test case, for example, the scattering model consisted of 6200 grid points any of which could be occupied by one of 3100 scatterers).

The inclusion of a mutation model in our computational tests proved to be essential for their success. The mutation strategy described was evolved in the course of the extensive (and often unsuccessful) testing of our early prototype algorithms (data not shown) and was found to effectively slow the rate at which the population converges. In our experience, we observed that a variable mutation rate generally yields better results than the use of a background mutation rate that is held constant throughout the run. Like the mutation operator, the inversion operator is designed to allow the evolutionary search access to all possible configurational states of the modelled system. We found that setting inversion probability P_i to zero at the outset of the run so that the inversion operator was not used at all made very little difference to the algorithm's performance provided that a suitable mutation model was used (data not shown).

We believe that the negative contrast images produced by the algorithm in these preliminary trials are an artefact of the simple scattering model and the relatively low resolution of the chosen subset of diffraction amplitudes that were used in each case. Indeed, this phenomenon has been previously observed in related *ab initio* phasing methods in which a uniform scattering model is used (Subbiah, 1991) and certainly appears not to be specific to this approach for reasons that we have discussed earlier. By modification of this approach from the simple binary protein/solvent model described here in which the density at each grid point is represented by a single bit (with a value of *one* or *zero*), to a scheme in which each scattering point is described as a binary-coded decimal number of several bits capable of representing a range of

possible electron densities, it should be possible to model scattering ensembles at higher resolutions. Conceptually therefore, the extension of this approach to higher resolutions would be relatively simple, although there would be a significant accompanying increase in the computational overhead of the evolutionary algorithm owing to the increased complexity of the search model and the greater sampling rate of the Fourier transform that is necessary at higher resolutions (Stout & Jensen, 1989, pp. 231–241). Testing of this algorithm at higher resolution with a more sophisticated scattering model may also facilitate a more detailed analysis of its results if it were able to produce positive electron-density images that could be directly compared with the real electron densities for the known structures whose diffraction amplitudes were used for the tests. Certainly, at this stage, there is much further investigation needed to properly characterize the behaviour and performance of the algorithm, but we feel that the results of these initial experiments with the prototype model scheme give grounds for optimism for the future development of this method. In the course of the continuing development of this approach from the prototype system described here, we are currently investigating the behaviour of evolutionary phasing algorithms at higher resolutions, using the more complex scattering models described above. We believe that there also exists ample scope for the further development of the evolutionary protocol itself, not only in its fundamental design (evolutionary protocols using combinatorial methods other than the genetic approach described here, for example) but also in the more systematic optimization of the various parameters controlling the actual evolutionary search.

6. Conclusions

The use of evolutionary search algorithms may also offer the potential to extend the effective range of other current approaches to direct phasing such as the combined real- and reciprocal-space 'Shake-and-Bake' and 'Half-Baked' methods (Miller *et al.*, 1994; Schäfer *et al.*, 1996). Their efficiency and implicit parallelism could be applied to searches in any functional space of the modelled system for which a suitable target function can be defined. Indeed, the evolutionary computational approach is completely general and in the development of new approaches to phasing, or in its application to the real- and reciprocal-space searches of existing direct-methods approaches, it may facilitate the more rapid discovery of macromolecular structures. While the results presented here may represent only a first step towards the ultimate goal of the full *ab initio* determination of protein structures at high resolution, it is felt that they constitute a very significant step. In extending the effective resolution range of current direct methods for the determination of macromolecules under typical conditions of symmetry, data quality and *a priori* structural knowledge, we believe that the examples described here amply demonstrate the potential power of such combinatorial methods in overcoming the phase problem. With the development of ever more sophisticated parallel computational architectures, the future application of such methods may eventually help crystallographers to determine the three-dimensional structures of proteins at rates closer to those at which the genes encoding them are being discovered. This would be a far cry from the current situation in which the structural determination of proteins can often lag years behind the discovery of their corresponding genes.

We would like to thank Dr M. S. Weiss and Dr J. R. Mesters of the IMB, Jena, for furnishing the diffraction data and coordinates used for this study, and Nancy DesRosiers and Janet Delahanty of Harvard Institutes of Medicine for the production and layout of the graphics. RH thanks the Fonds der Chemischen Industrie for support.

References

Bäck, T., de Graaf, J. M., Kok, J. N. & Kosters, W. A. (1997). *Bull. Eur. Assoc. Theor. Comput. Sci.* **63**, 161–192.
Berchtold, H., Reshetnikova, L., Reiser, C. O. A., Schirmer, N. K., Sprinzl, M. & Hilgenfeld, R. (1993). *Nature (London)*, **365**, 126–132.
Blickle, T. & Thiele, L. (1996). *Evolutionary Comput.* **4**, 361–394.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Chang, D. & Lewis, M. (1994). *Acta Cryst.* D**50**, 667–674.
Connor, R. (1994). *Practical Handbook of Genetic Algorithms: Applications*, edited by L. D. Chambers, pp. 57–74. Boca Raton: CRC Press.
Drenth, J. (1994). *Principles of Protein X-ray Crystallography*, pp. 161. New York: Springer-Verlag.
Forthofer, R. N. & Lee, E. S. (1995). *Introduction to Biostatistics*, pp. 128–129. London: Academic Press.
Hauptman, H. & Karle, J. (1953). *Solution of the Phase Problem. I. The Centrosymmetric Crystal.* American Crystallographic Association Monograph, No. 3.
Hauptman, H. A. (1996). *Acta Cryst.* A**52**, 490–496.
Hendrickson, W. A., Smith, J. L., Phizackerley, R. P. & Merritt, E. A. (1988). *Proteins*, **4**, 77–88.
Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan Press.
Kariuki, B. M., Serrano-Gonzalez, H., Johnson, R. L. & Harris, K. D. M. (1997). *Chem. Phys. Lett.* **280**, 189–195.
Karle, J. (1989). *Acta Cryst.* A**45**, 765–781.
Karle, J. & Karle, I. L. (1966). *Acta Cryst.* **21**, 849–859.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.
Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst.* **27**, 613–621.
Miller, S. T., Hogle, J. M. & Filman, J. D. (1996). *Acta Cryst.* D**52**, 235–251.
Mukherjee, M., Maiti, S. & Woolfson, M. M. (2000). *Acta Cryst.* D**56**, 1132–1136.
Podjarny, A. D. & Rees, B. (1994). *Computational Crystallography 5*, pp. 361–372. Oxford University Press.
Podjarny, A. D. & Urzhumtsev, A. G. (1997). *Methods Enzymol.* **276**, Part A, pp. 641–658.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Schäfer, M., Schneider, T. R. & Sheldrick, G. M. (1996). *Structure*, **15**, 1509–1515.
Stout, G. H. & Jensen, L. H. (1989). *X-ray Structure Determination, a Practical Guide*. New York: Wiley.
Subbiah, S. (1991). *Science*, **252**, 128–133.
Watenpaugh, K. D. (1985). *Methods Enzymol.* **115**, Part B, 3–15.