Acta Cryst. (1998). D**54**, 750–756

Modified Phased Translation Functions and their Application to Molecular-Fragment Location

KEVIN COWTAN

Department of Chemistry, University of York, Heslington, York YO1 5DD, England.
E-mail: cowtan@yorvik.york.ac.uk

(*Received 13 August 1997; accepted 10 November 1997*)

Abstract

Direct methods at high resolution have depended on the resolution of atomic like features in the map. At data resolutions more typical for protein structures (2–3 Å) individual atoms may not be resolved, so larger features must be identified. At one extreme the whole molecule may be located using the diffraction magnitudes alone by the molecular-replacement method. At the other extreme it is possible to locate individual residues in a well phased map. In this paper an intermediate problem is addressed: the location of multi-residue fragments on the basis of weak phase information. An agreement function based on the mean-squared difference between model and map over a masked region is shown to be more effective than a simple overlap integral, and may be efficiently calculated by Fourier methods. The techniques are compared using poorly phased electron-density maps at ~3 Å for the proteins RNAse and O6-methylguanine-DNA-methyltransferase.

1. Introduction

Kleywegt & Jones (1997) have shown that an exhaustive real-space search with a molecular fragment can give strong indications of the presence and location of helix-like or strand-like features in an electron-density map. In the *ESSENS* procedure a search is performed in six dimensions over all possible positions and orientations of a fragment. The fragment coordinates are mapped into the electron-density-map space using each translation and orientation, and the map densities near atomic centres are compared to obtain a positional score. The best fits are stored and can be interpreted as a map showing how likely a fragment is to be present at any point in the map.

For the procedure to be effective in obtaining the location of short helices in an uninterpreted density map, the choice of function for measuring the agreement between the maps is critical. Kleywegt & Jones recommend a function which is based on the worst agreements between the fragment and the map at the atomic centres in the fragment.

One drawback of this approach is the computational requirement. The time taken for the calculation increases with the number of search positions and orientations explored and with the size of the fragment. A possible approach to reducing the computational requirement is to implement the translation search through Fourier methods.

The simplest function to implement by Fourier methods is the overlap integral or phased translation function, given by the product of the fragment density and the map density summed over the volume of the fragment. Colman *et al.* (1976) suggest the use of this function for the location of an oriented molecular-replacement model using low-resolution phase information. The phased translation function is extended by Read & Schierbeek (1988) to form a simple correlation function.

The overlap integral may be calculated for every position of the fragment by convolution of the fragment density with the map density; this function may be calculated in reciprocal space from the product of the Fourier coefficients of the two density functions. However, as Kleywegt & Jones suggest, this function is a weak discriminator of the correct fragment location.

The construction of more complex agreement functions using Fourier transforms can provide a more powerful discriminator, which may be used to search for arbitrarily large fragments at no additional computational cost. It should be clear that the fragment-matching problem is a special case of the phased translation function, and that the approaches described here are equally applicable to the solution of phased translation search problems.

2. Methods

A simple function which may be expressed in terms of Fourier transforms is the mean-squared difference between the fragment and the map over the volume of the fragment. This shares some properties with Kleywegt & Jones' 'worst-agreement' function, since the greatest differences will contribute most to this function. (Indeed higher order differences, such as the mean fourth power of the difference, behave similarly to a simple maximum function of the differences.) This function has the additional property of testing the agreement between areas of low density (as well as high density) within the fragment.

© 1998 International Union of Crystallography
Printed in Great Britain – all rights reserved

ISSN 0907-4449 © 1998

For the mean-squared difference to be meaningful, both the fragment density and the map density must be placed on the same scale; in these calculations an absolute scale was adopted and the $F(000)$ term was set to the number of electrons in the unit cell. In these tests the overall temperature factor was removed from the data and the fragment density was calculated with the atomic temperature factors set to zero.

Note that the mean-squared difference must only be calculated over the volume of the fragment within which the density from the model fragment and an associated map feature might be expected to match. Therefore, the mean-squared-difference function also includes a masking function which is unity over the volume of the fragment and zero elsewhere (although fractional values of the masking function could also be used in order to weight parts of the model which are uncertain).

Let the discriminator be called $t(x)$. The fragment density is $\rho_f(x)$ and the corresponding fragment mask is $\varepsilon_f(x)$. The discriminator may then be formed from the sum of the mean-squared difference in density between the offset fragment and the map,

$$\begin{aligned} t(x) &= \sum_y \varepsilon_f(y)[\rho_f(y) - \rho(y-x)]^2 \\ &= \sum_y \varepsilon_f(y)\rho_f^2(y) - 2\varepsilon_f(y)\rho_f(y)\rho(y-x) \\ &\quad + \varepsilon_f(y)\rho^2(y-x). \end{aligned} \tag{1}$$

The r.m.s. difference may be formed from the square root of $t(x)/\sum_y \varepsilon_f(y)$. Note that in the expansion the first term is independent of x and so is only calculated once, whereas the second two terms are convolutions and, therefore, may be efficiently calculated in reciprocal space,

$$\begin{aligned} t(x) = \sum_y \varepsilon_f(y)\rho_f^2(y) + (1/V)\mathcal{F}\{&\mathcal{F}^{-1}[\varepsilon_f(x)]\mathcal{F}^{-1}[\rho^2(x)]^* \\ &- 2\mathcal{F}^{-1}[\varepsilon_f(x)\rho_f(x)]\mathcal{F}^{-1}[\rho(x)]^*\} \end{aligned} \tag{2}$$

where $\mathcal{F}$ represents the Fourier transform, $\mathcal{F}^{-1}$ represents the inverse Fourier transform and * represents complex conjugation. If the Fourier coefficients of the density and squared density are pre-calculated, then the translation function for a fragment in multiple orientations may be calculated by three fast Fourier transforms (FFTs) per orientation.

The FFTs must be performed in $P1$, however crystallographic symmetry may be used to reduce either the number of search orientations or the volume of $t(x)$ which must be evaluated.

2.1. *Modified discriminators*

It is possible to refine this function in a number of ways. For example, if the low-resolution reflections are missing, the map will show long-range variations in mean density. If the mean of the map over the volume of the fragment differs from the mean of the fragment, then the mean-squared difference will be increased.

The discriminator may be modified to contain a term which matches the mean of the map over the region covered by the fragment to the mean of the fragment,

$$t(x) = \sum_y \varepsilon_f(y)\{[\rho_f(y) - \overline{\rho_f}] - [\rho(y-x) - \overline{\rho}(x)]\}^2, \tag{3}$$

where

$$\overline{\rho_f} = \sum_y \varepsilon_f(y)\rho_f(y) / \sum_y \varepsilon_f(y),$$

$$\overline{\rho}(x) = \sum_y \varepsilon_f(y)\rho(y-x) / \sum_y \varepsilon_f(y).$$

$\overline{\rho_f}$ and $\overline{\rho}(x)$ are means over the volume covered by the fragment mask. $\overline{\rho}(x)$ is calculated by FFTs.

A further modification involves matching the variance of the map over the region covered by the fragment to the variance of the fragment. The agreement function becomes equivalent to a correlation coefficient,

$$t(x) = \sum_y \varepsilon_f(y)\{[\rho_f(y) - \overline{\rho_f}] - \tfrac{\sigma_{\rho_f}}{\sigma_\rho(x)}[\rho(y-x) - \overline{\rho}(x)]\}^2, \tag{4}$$

where

$$\sigma_{\rho_f}^2 = \sum_y \varepsilon_f(y)\left[\rho_f(y) - \overline{\rho_f}\right]^2 / \sum_y \varepsilon_f(y),$$

$$\sigma_\rho(x) = \sum_y \varepsilon_f(y)[\rho(y-x) - \overline{\rho}(x)]^2 / \sum_y \varepsilon_f(y).$$

σ_{ρ_f} and $\sigma_\rho(x)$ are standard deviations over the volume covered by the fragment mask. $\sigma_\rho(x)^2$ is also calculated by FFTs. This method has the advantage that it is no longer necessary to scale the data, although the temperature factor of the model should still match that of the data. This function differs from the correlation function of Read & Schierbeek (1988) by the use of the fragment mask to limit the correlation calculation to the region of the fragment.

After simplification both the mean-adjusted discriminator and the variance-adjusted discriminator may be calculated using two additional FFTs over the basic discriminator in (2).

2.2. *Reciprocal-space filtering*

An alternative approach to improving the discriminator may be to take into account directly which reflections are available in calculating the initial density map. If all the terms of the discriminator are accumulated in reciprocal space, then only those terms for which map coefficients are available need be used in calculating the final discriminator map. Further restrictions, such as resolution limits, may be placed on the terms which are used.

This approach may avoid the need to correct the mean of the map over the masked region to match the mean of the fragment. Appropriate resolution-dependent scaling (taking into account the size of the fragment) may also replace the correction of variance and temperature factor.

3. Application

3.1. *Translation search*

To compare the discriminators, a translation search was performed on an uninterpretable SIR map of RNAse from *Streptomyces aureofaciens* (Ševčík *et al.*, 1991). The structure consists of 96 amino acids, including one α-helix and a twisted three-strand antiparallel β-sheet. The structure was solved using multiple isomorphous derivatives and refined to 1.8 Å.

This data set was chosen because the derivative data were all available. The phasing was therefore recalculated using a mercury derivative alone, giving a mean figure of merit of 0.26 to 3.2 Å.

The search model was an α-helix of ten polyalanine residues in the correct orientation. A fragment mask was calculated surrounding the atoms of the fragment to a radius of 2.5 Å.

Translation-search functions were calculated using the conventional phased translation function (overlap integral) and the three forms of mean-squared-differ-

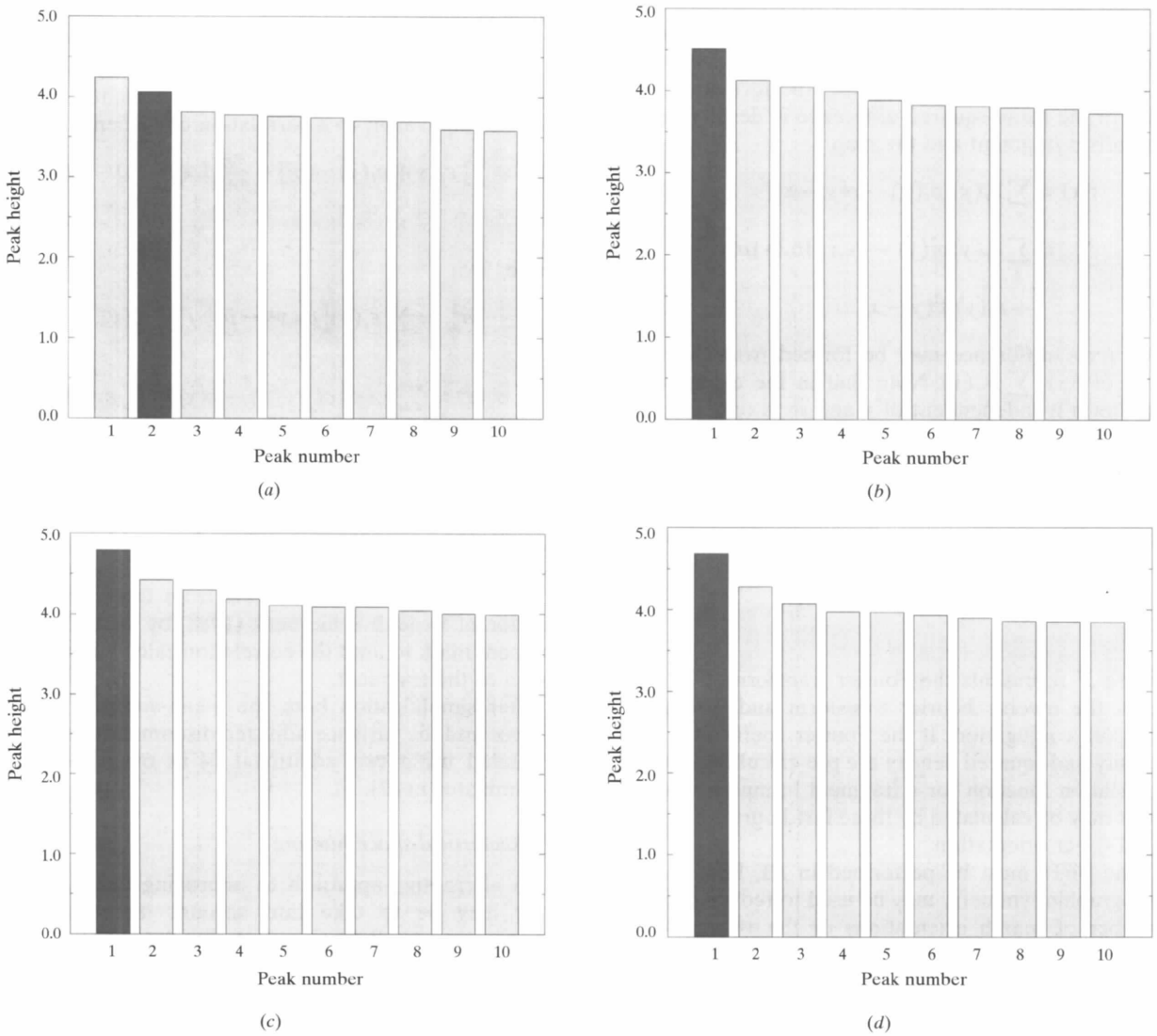


Fig. 1. Heights of the top peaks in the discriminator function using (*a*) convolution/overlap, (*b*) squared-difference, (*c*) mean-adjusted and (*d*) variance-adjusted discriminators. Heights are given in standard deviations above the mean of the discriminator. The data is an SIR electron-density map of RNAse with a mean figure of merit of 0.26 to 3.2 Å resolution.

ence function [equations (2), (3) and (4)]. The heights of the top ten peaks in the translation function are plotted in Fig. 1, with the correct peak shaded. Using the conventional phased translation function (overlap integral), the correct peak is only the second highest in the map, and not very distinct from the others. In the mean-squared-difference map, the correct peak is the highest in the map and is easily identifiable. The mean-adjusted and variance-adjusted discriminators give a slight additional improvement at the cost of some computation.

The absolute peak heights in e $Å^{-3}$ for the difference-based discriminators are listed in Table 1. It is clear that fitting the mean and/or variance of the map improves the fit of the fragment against both genuine and noise features of the map. The actual differences are small, but most pronounced for the variance-corrected form. Similar results are obtained with better maps.

Table 1. *Absolute signal and noise peak heights for difference-based discriminators locating a correctly oriented fragment in an SIR electron-density map of RNAse at 3.2 Å resolution*

Discriminator	Correct peak (e $Å^{-3}$)	Highest noise peak
R.m.s. difference [equation (2)]	1.4332	1.4371
+ mean adjustment [equation (3)]	1.3516	1.3545
+ variance adjustment [equation (4)]	1.1607	1.1849

3.2. *Fragment searching*

A fragment search was performed in an MIR map of O6-methylguanine-DNA methyltransferase (Moore *et al.*, 1994), a DNA repair protein of 178 amino acids. The structure includes six helices and a three-strand β-sheet. The MIR data provided phases with a mean figure of

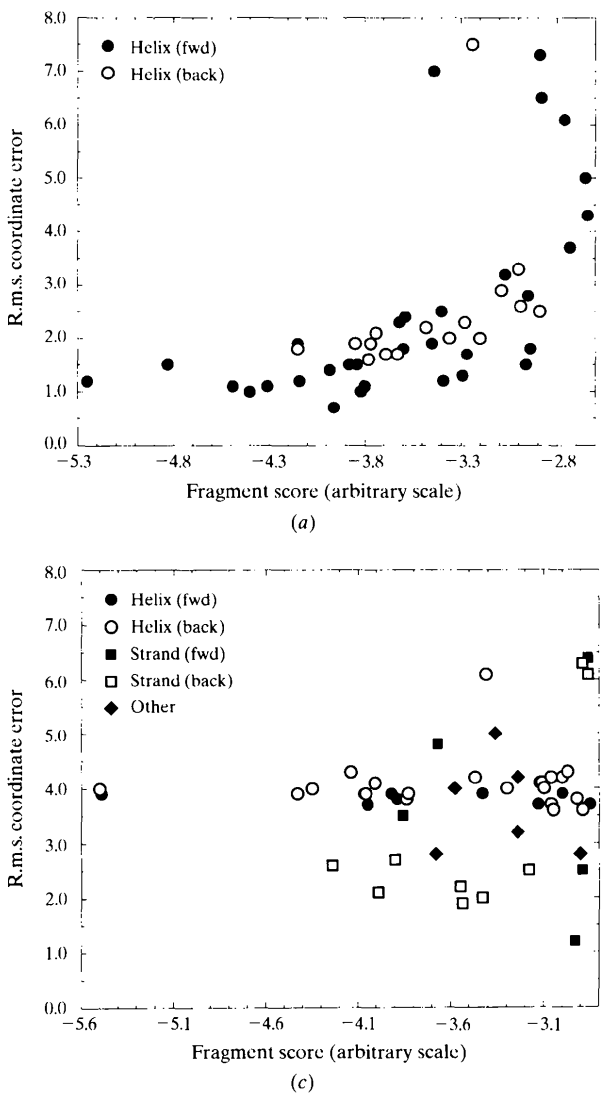


Fig. 2. Fragment-search results for the O6-methylguanine-DNA-methyltransferase data by *ESSENS* in real space using (*a*) ten-residue helix, (*b*) five-residue helix and (*c*) five-residue β-strand. The x axis is the negative of the fragment score, with the best fragment on the left. The y axis is the r.m.s. distance between the Cα atoms in the fragment and the nearest matching section of the true structure. The symbols indicate the chain direction and the secondary-structure type for the matched section.

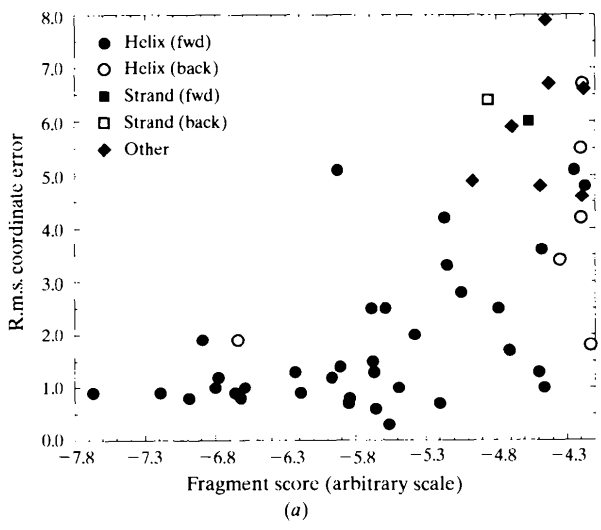
merit of 0.46 to 3.1 Å resolution, which in the original structure solution was sufficient for positioning of about two-thirds of the atoms. The space group is $P2_12_12_1$.

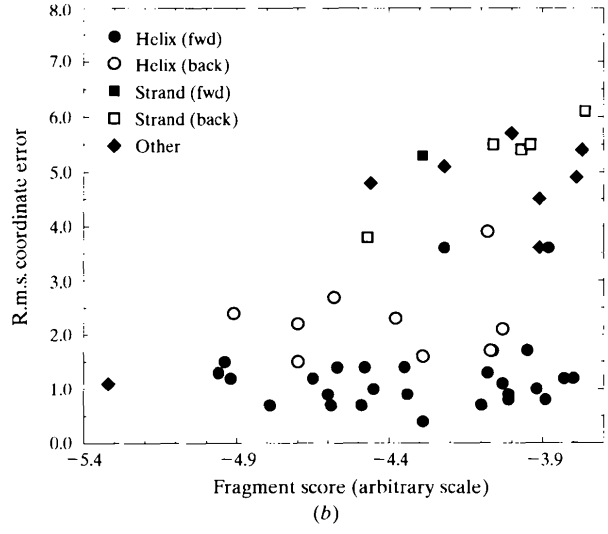
The searches were performed over both translational and rotational spaces with each fragment and the results compared with the *ESSENS* approach of Kleywegt & Jones (1997) in real space and the mean-squared-difference function calculated in reciprocal space. Experiments with the reciprocal-space approach suggested that the translation-search grid should be fairly fine, with best results when the grid is around half the Nyquist spacing. However, the angular search could be quite coarse, with a limit of about 20°.

Three search fragments were tested: a ten-residue α-helix, a five-residue α-helix and a five-residue β-strand (the structure does not contain strands long enough for a ten-residue β-fragment, however, an antiparallel sheet fragment might be usable). The top 50 matches for each fragment were compared with the known structure and the coordinate error between the fragment and the nearest contiguous segment of real molecule was evaluated. The results are shown in Fig. 2 for the *ESSENS* approach and Fig. 3 for the reciprocal-space approach. The CPU timings for the various calculations are shown in Table 2.

Using the ten-residue helical fragment, the first 25 matches from the *ESSENS* approach are correct, of which 17 give the correct chain direction. The first 25 matches from the reciprocal-space approach are also correct, of which 24 give the correct chain direction. The one outlier on the graph (5.1 Å r.m.s. error) is a match to the short seven-residue helix. Beyond the first 25 matches both methods give a high proportion of incorrect matches.

Using the five-residue helical fragment, the first 15 matches from each method are correct, including in both





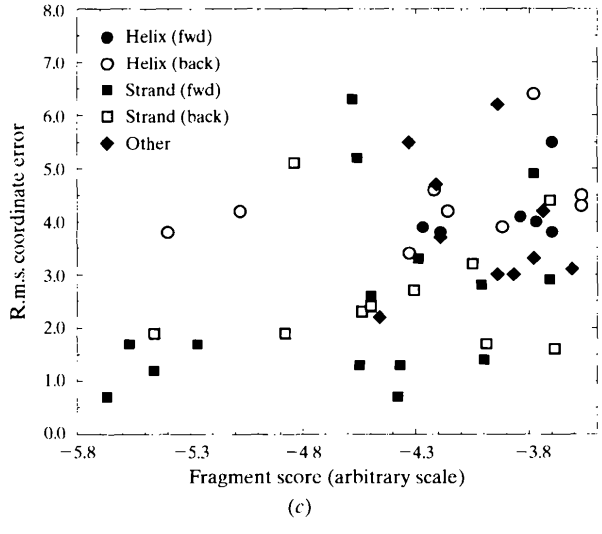


Fig. 3. Fragment-search results as for Fig. 2 but by the mean-squared-difference method in reciprocal space, using (*a*) ten-residue helix, (*b*) five-residue helix and (*c*) five-residue β-strand. The *x* axis is the fragment score, with the best fragment on the left. The *y* axis is the r.m.s. distance between the Cα atoms in the fragment and the nearest matching section of the true structure. The symbols indicate the chain direction and the secondary-structure type for the matched section.

Table 2. *Time comparisons (in hours) for real- and reciprocal-space fragment searches*

Calculations were performed on a 200 MHz MIPS R10000 CPU, searching over 786432 grid points and about 5000 orientations. Notes: For speed a mask restricted the search to a single molecule and map points above the mean (63242 points). Some further speed improvement may be obtained by increasing the density cutoff. The FFT was optimized for the fine map grid, however, neither the crystallographic symmetry nor aggressive compiler optimization were used.

Method	Five-residue fragment	Ten-residue fragment
Real space (*ESSENS*)	5.4	11.5
Reciprocal space (mean-squared difference)	2.0	2.1

cases a turn which shows roughly helical geometry (this is the best match from the reciprocal-space method). Again the correct chain direction is indicated in the majority of cases.

Using the five-residue strand fragment, the reciprocal-space approach gives six correct matches from the top eight, with the remaining matches fitting the strands along the centre of helices. Most of the top matches from the *ESSENS* approach fit the strand fragment along helices, although some correct matches are obtained further down the list.

The molecule is shown in Fig. 4, along with fragment structures assembled from the 25 best ten-residue helix matches and 15 best five-residue strand matches found by each approach. The reciprocal-space approach gives matches to all the helices and three sections of β-strand. The *ESSENS* approach gives matches to three helices and two sections of β-strand. The missing helices are also weak or missing in the *ESSENS* feature maps, and include the short seven-residue helix and two helices which deviate significantly from the α-geometry, however, some of the missing features may be extracted using a five-residue search helix.

The two approaches appear to be broadly comparable in performance, however, the sensitivity of the mean-squared-difference method to regions of low as well as high density in the model gives it greater power to distinguish strand features from helices. The reciprocal-space approach is considerably faster, especially for larger fragments.

4. Conclusions

It has been shown that it is possible to use a modified phased translation function to locate small-density fragments in low-quality maps. The squared-difference function and variants may be implemented in reciprocal space. For very small fragments (two residues or less with the current code) this will be slower than a direct real-space calculation. However, the computational requirement is almost independent of fragment size, thus for the five- to ten-residue fragments typically used when searching for strand and helix motifs the reciprocal-space approach is faster.

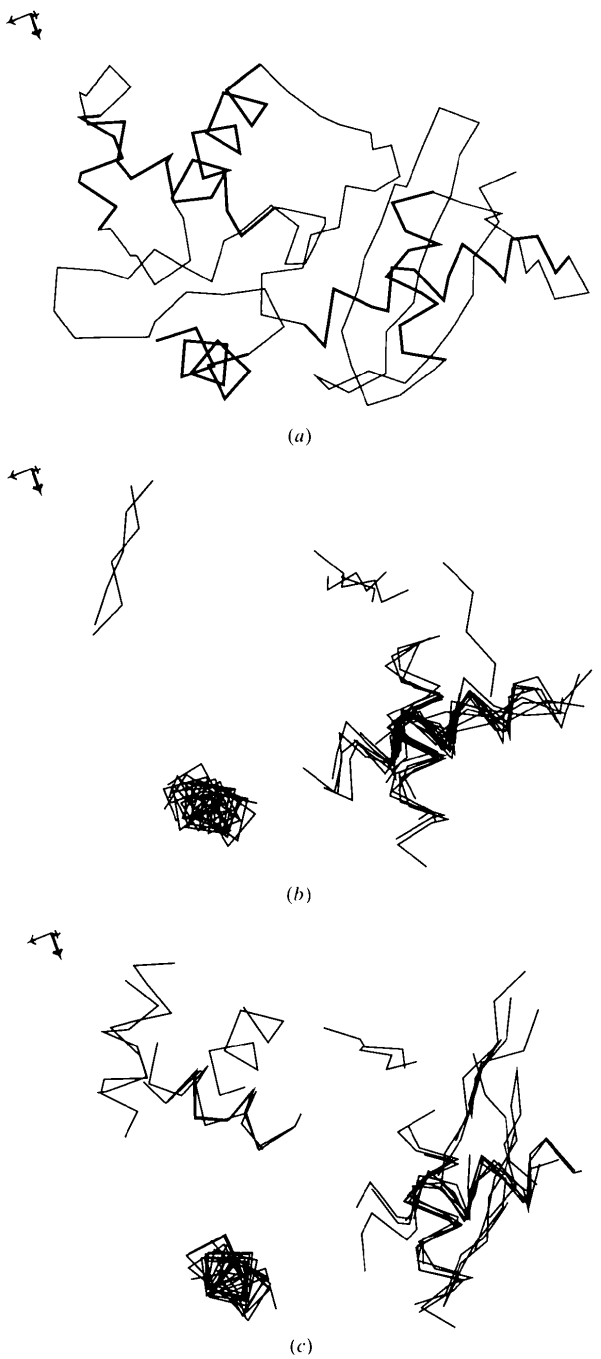


Fig. 4. Cα trace of O6-methylguanine-DNA methyltransferase and the helix and strand fragments found using real-space and reciprocal-space approaches. (Plots generated using *QUANTA*, Molecular Simulations Inc., 1997.) (*a*) Cα trace, with helices highlighted. (*b*) Helix and strand fragments from the *ESSENS* real-space approach. (*c*) Helix and strand fragments from the reciprocal-space approach.

Calculation in reciprocal space places some limitations on the complexity of the discriminator, although higher order polynomials based on density differences could be implemented. In this respect the calculation is less versatile than the real-space approach, but in the case of helical fragments the results can be significantly improved through the use of larger search models. In other cases, such as the fitting of β-strands in the test case, matching of both high and low density within the fragment improves discrimination. It is also possible that optimization of the fragment mask may further improve sensitivity. The construction of optimal masks, as well as alternative search motifs, will require a systematic search of the protein databases for correlated-density fragments.

Using a library of search fragments, including helices, strands, sheets and common loop motifs, the potential exists to model the unit-cell contents in terms of the best-fitting density fragments and solvent on the basis of comparatively poor initial phasing. Overlapping fragments may be combined and inconsistent overlaps removed (this has been partially implemented in Kleywegt & Jones' *SOLEX* program). The resulting structure fragments could then be used to provide new phase information in a density-modification method, or input to a refinement method.

Dr Cowtan is grateful to the United Kingdom BBSRC for funding this work (grant number 87/B03785).

References

Colman, P. M., Fehlhammer, H. & Bartels, K. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, K. Huml & B. Sedlacek, pp. 248–258. Copenhagen: Munksgaard.

Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* **D53**, 179–185.

Molecular Simulations Inc. (1997). *QUANTA*, Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752, USA.

Moore, M. H., Gulbis, J. M., Dodson, E. J., Demple, B. & Moody, P. C. E. (1994). *EMBO J.* **13**, 1495–1501.

Read, R. J. & Schierbeek, A. J. (1988). *J. Appl. Cryst.* **21**, 490–495.

Ševčík, J., Dodson E. & Dodson G. G. (1991). *Acta Cryst.* B**47**, 240–253.