Acta Cryst. (1997). D53, 756-764

Probing the Limits of the Molecular Replacement Method: the Case of *Trypanosoma brucei* Phosphoglycerate Kinase

BRADLEY E. BERNSTEIN[a,b] AND WIM G. J. HOL[a,b]*

[a]*Departments of Biochemistry and Biological Structure, University of Washington, Box 357742, Seattle, WA 98195, USA, and* [b]*Howard Hughes Medical Institute, University of Washington, Box 357742, Seattle, WA 98195, USA. E-mail: bmsc@gouda.bmsc.washington.edu*

(*Received 19 May 1997; accepted 15 July 1997*)

Abstract

It is of general interest to explore the limits of the molecular replacement method. Described here are the specifics of a successful molecular replacement structure determination in a difficult case: the X-ray crystal structure of a *Trypanosoma brucei* phosphoglycerate kinase (PGK) ternary complex. This ternary complex crystallizes with four 45 kDa subunits in the asymmetric unit, whereas the available search models were all monomers consisting of two distinct domains. Initial molecular replacement attempts using complete subunits were unsuccessful. Attributing this failure to a presumed change in the relative orientations of the N- and C-terminal phosphoglycerate kinase domains, a second attempt was made using each domain as an independent search model. In this way, two N-terminal and then two C-terminal domains could be oriented and positioned in the unit cell. On the basis of this result, a new search model containing both domains in the correct mutual orientation was created and used to identify the two remaining phosphoglycerate kinase subunits. The ability to successfully orient and position an N-terminal domain containing 8.4% of the scattering mass in the asymmetric unit was the key to this structure determination. Further investigations show that a truncated version of this search model containing 5.8% of the scattering mass would have been sufficient for this purpose. A retrospective analysis suggests that the effectiveness of this probe is enhanced by structural conservation, retained temperature factors and a disparity in the degree of order among the various subunits in the *T. brucei* PGK asymmetric unit. Based on these observations, 231 well ordered Cα atoms were selected from a single refined *T. brucei* PGK subunit and it was demonstrated that this collection of atoms, representing just 1.6% of the scattering mass, could be correctly oriented and positioned in the unit cell.

1. Introduction

The molecular replacement method, in which a closely related protein is oriented and fixed in an unknown unit cell to determine phases, is becoming ever more effective with the increasing power and efficiency of hardware and software (see Turkenburg & Dodson, 1996; Rossmann, 1990, for reviews). At the same time the rapidly expanding database of proteins and domains available for use as search models makes this technique amenable to an increasing number of projects. Testament to its power are a series of reports describing the use of remarkably small probes in the molecular replacement functions (Table 1). First, it was reported in 1985 that a very small probe composed of Cα atoms (13% of the total atoms) from the actinidin protein could be oriented and translated into the unit cell of papain, a structurally conserved protein with 48% sequence identity (Schierbeek, Renetseder, Dijkstra & Hol, 1985). The efficiency of the Cα atom search model was further documented by the finding that the Cα trace of the constant region of an Fab containing just 6% of the scattering mass was sufficient for correct orientation in the unit cell of a different Fab (Cygler & Anderson, 1988). More remarkable, is the recent report that a snowdrop lectin search model containing just 65 Cα atoms corresponding to less than 4% of the asymmetric unit could be correctly oriented and translated in the unit cell of amaryllis lectin, a protein 85% identical in sequence to the search model (Chantalat, Wood, Rizkallah & Reynolds, 1996).

At this point it should be stressed that there are three distinct elements of molecular replacement as developed by Rossmann & Blow (1962). These elements include (1) the orientation of a search probe, (2) the translation of a previously oriented probe in the unit cell, and (3) the use of the derived unit-cell contents to obtain crystallographic phases. While a Cα trace might be sufficient to achieve the first two elements, the resulting model would not be adequate for phase determination. Furthermore, as noted by Chantalat *et al.* (1996), the Cα trace search probe may not be a very practical example since in most cases main-chain atoms can easily be added. When these investigators sought the minimal main-chain stretch of snowdrop lectin required for rotation and translation in the amaryllis lectin cell, they found that 47 residues (comprising 21% of protein

Table 1. *A survey of very small search probes used successfully in the molecular replacement functions*

Included in the table are search model and target protein (highly conserved in all cases), scattering mass as percentage of total in the asymmetric unit, and whether the probe is based on main-chain or Cα atoms. Probes composed of Cα atoms, though highly efficient, are not relevant for molecular replacement structure determinations since the rest of the atoms in a search model are typically available and since Cα atoms alone are insufficient to determine phases.

Search model	Target	Scattering mass (%)	Probe	Reference
Actinidin	Papain (48% sequence identity)	12	Cα atoms	Schierbeek *et al.* (1985)
Fab McPC603 C domain	Fab HED10 C domain	6*	Cα atoms	Cygler & Anderson (1989)
Phospholipase A_2	Phospholipase A_2 plus inhibitor (hexamer)	16	Main chain	Oh (1995)
Snowdrop lectin	Amarylis lectin (85% sequence identity)	3.9	Cα atoms	Chantalat *et al.* (1996)
B.s. PGK	*T.b.* PGK (49% sequence identity)	5.8	Main chain	This paper
T.b. PGK subunit	*T.b.* PGK (four in a.u.)	1.6	Cα atoms $B < 25\,\text{Å}^2$	This paper

* Rotation function solution only.

atoms) were required. In fact, the smallest successful *main-chain* search probe reported in the literature contains a significantly greater proportion of the unit cell than the minimized Cα probes: the backbone of a single subunit of phospholipase was oriented and positioned in an asymmetric unit containing six protein copies of the same protein in complex with inhibitor (Oh, 1995). These authors conclude, in the more relevant case of the main-chain search model, that a minimum of 16–17% of the scattering mass is necessary to solve the molecular replacement problem.

Another significant advancement in molecular replacement made in recent years concerns the handling of flexible multi-domain proteins which are often not solvable by conventional approaches. This type of conformational variability is found in many proteins, most notably in the immunoglobulins. The effectiveness of the rotation function, which relies on the superposition of Patterson functions, is greatly diminished in cases when domains in the search model and target crystal structure assume different orientations. This obstacle has been addressed in molecular replacement in two fundamental ways. One approach, illustrated by the structure determination of the HED10 Fab fragment (Cygler & Anderson, 1988), involves dividing the search model into individual rigid domains to be used as independent search probes, in this way permitting inter-domain rearrangements. This approach requires that search models representing fractions of the asymmetric unit be successively placed in the unit cell. It is best performed in conjunction with a molecular replacement algorithm that considers information about previously fixed probes. Several programs in widespread use have the capacity to incorporate a previously fixed probe into the translation function (Brünger, 1992; Navaza, 1994; Fujinaga & Read, 1987). Furthermore, a method whereby the theoretical Patterson of a known fragment is subtracted from the observed Patterson has been used to compute an improved rotation function for the still unknown fragment (Zhang & Matthews, 1994).

A second approach for handling inter-domain flexibility is also commonly applied to immunoglobulins. This method uses Patterson correlation refinement (Brünger, 1990) to optimize an approximately oriented search model, in this way allowing for the relative reorientation of domains. This procedure has been applied successfully to a series of problematic multi-domain molecular replacement cases including an anti-dinitrophenyl-spin-label antibody Fab fragment (Brünger, Leahy, Hynes & Fox, 1991) and an Fab 26–10/digoxin complex (Brünger, 1991).

The molecular replacement case we present here is the structure solution of closed form phosphoglycerate kinase (PGK) from *Trypanosoma brucei*. This glycolytic enzyme uses a magnesium cofactor to catalyze the

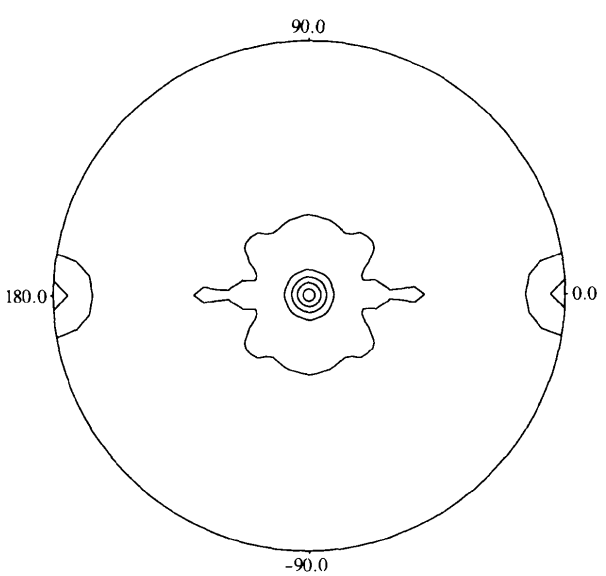

Fig. 1. The self-rotation function $\kappa = 90°$ section generated by *POLARRFN* (Collaborative Computational Project, Number 4, 1994) indicates that at least two of the subunits in the *T. brucei* PGK unit cell are related by 90° about the crystallographic *Z* axis.

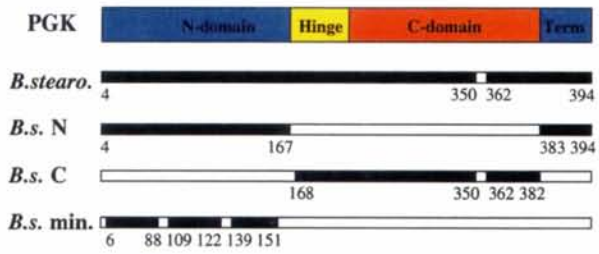

Fig. 2. For the structure determination of open form *T. brucei* PGK, a multi-alanine search model containing residues 4–350 and 362–394 from *B. stearothermophilus* PGK (Davies *et al.*, 1994) was divided into separate N domain (*B.s.* N) and C domain (*B.s.* C) search probes. In a retrospective analysis, a truncated version of the N-domain probe (*B.s.* min.) containing 5.8% of the scattering mass could be correctly oriented and translated in the *T. brucei* PGK unit cell.

reversible phosphoryl-transfer reaction between 1,3-bis-phosphoglycerate and ADP to produce ATP. *T. brucei* PGK is 45–50% identical in sequence to PGK from equine, yeast, porcine and bacterial sources, for which crystal structures have been reported (Blake & Evans, 1974; Watson *et al.*, 1982; Harlos, Vas & Blake, 1992; Davies *et al*, 1994). Like these other enzymes, *T. brucei* PGK is monomeric in solution and has similar kinetic parameters (Misset, Bos & Opperdoes, 1986). The crystal structures reveal that PGK consists of an N- and a C-terminal domain (referred to as N and C domains henceforth) connected by a narrow hinge region and separated by a wide cleft. The triose sugar and nucleotide substrates bind to distinct sites on the N and C domains, respectively. Bound in this way the substrates are more than 10 Å apart, a distance

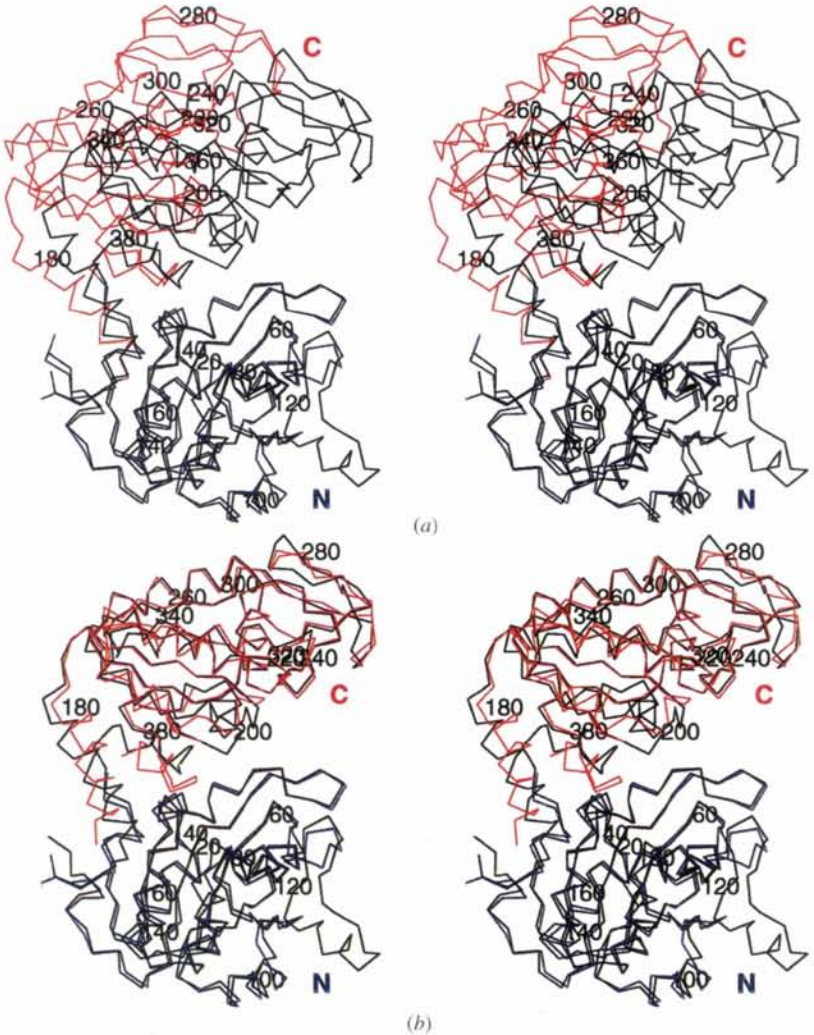

Fig. 3. (*a*) The Cα trace of refined closed form *T. brucei* PGK (black) is compared with that of open form *B. stearothermophilus* PGK (Davies *et al.*, 1994). Regions of *B. stearothermophilus* PGK corresponding to the N- and C-domain search probes are colored blue and red, respectively. (*b*) *T. brucei* PGK Cα trace is shown in black. The N- (blue) and C- (red) domain *B. stearothermophilus* search probes are superimposed separately on the corresponding *T. brucei* PGK domains.

Table 2. *Search models available for T. brucei PGK structure determination*

Sequence identity, resolution and R factor (for refined structures), bound substrates and hinge closure with respect to open form PGK are shown.

PGK crystal structures	Identity (%)	Resolution (Å) (R factor, %)	Substrate	Hinge angle (°)
Horse (Blake & Evans, 1974)	46	3.0	—	0
Yeast (Watson *et al.*, 1982)	48	2.5	ATP	0
Porcine (Harlos *et al.*, 1992)	46	2.0 (21.5)	3-PGA	8
B.stearo (Davies *et al.*, 1994)	49	1.7 (15.6)	ADP	0

clearly incompatible with catalysis. A hinge-bending hypothesis has been invoked to explain this observation, and it proposes that prior to catalysis the enzyme domains rotate towards each other and bring the substrates together (Banks *et al.*, 1979). Low-angle X-ray scattering data for yeast PGK in solution is consistent with this type of motion (Pickover, McKay, Engleman & Steitz, 1979).

We have crystallized *T. brucei* PGK in the presence of the substrate/product combination 3-phosphoglycerate (3-PGA) and MgADP, and solved the structure of the ternary complex (Bernstein, Michels & Hol, 1997). In contrast to all previously reported PGK structures, the *T. brucei* PGK ternary complex resides in a 'closed' conformation. The closed form differs dramatically from previous PGK structures: N and C domains are now rotated inward by approximately 30° and the substrates have correspondingly been brought into close proximity and in-line for catalysis. Conformational change is restricted to the intervening hinge region while the domains move as rigid bodies.

Here we present the structure determination of the closed form *T. brucei* PGK ternary complex starting from an open-form *Bacillus stearothermophilus* PGK search model. Of particular relevance to this structure determination is that there are four copies of the ternary complex in the asymmetric unit of the *T. brucei* PGK crystals, a considerable complication in view of the fact that no multimeric search models were available.

2. Crystallographic data and non-crystallographic symmetry

The glycosomal isozyme of *T. brucei* PGK was expressed and purified, and crystals of the ternary complex were grown in the presence of 10 m*M* 3-PGA and 10 m*M* MgADP by equilibration with 2.5 *M* potassium sodium phosphate (Bernstein *et al.*, 1997). Data were collected to 2.8 Å from a single cryopreserved crystal on the CHESS A1 beamline using an ADSC CCD detector and processed with *DENZO* and *SCALEPACK* (Otwinowski, 1993). The space group is $P2_12_12_1$ with unit-cell dimensions $a = 113.2$, $b = 116.0$ and $c = 171.3$ Å. The Matthews coefficient is 3.05 Å^3 Da^{-1} assuming four *T. brucei* PGK ternary complexes per asymmetric unit (Matthews, 1968). Since no multimeric search model was available, it was unclear whether to attempt to obtain phases by molecular replacement with a model representing only a fraction of the scattering mass, or to proceed with multiple isomorphous replacement. However, a self-rotation function generated a peak, 60% of the origin in height, corresponding to an approximately 90° rotation about the Z axis (Fig. 1). This relationship seemed likely to be useful in the identification of correct solutions from a rotation function list and so we proceeded with molecular replacement.

3. Molecular replacement

3.1. *Search model and trials with open-form PGK subunits*

The binary complex of *B. stearothermophilus* PGK plus MgADP (Davies *et al.*, 1994) was the highest resolution and most well refined PGK structure available and, with the exception of small variations in hinge angle, was highly representative of all known structures of this enzyme (Table 2). Thus, *B. stearothermophilus* PGK (49% identical in sequence to *T. brucei* PGK) was chosen as a starting search model. To optimize this probe, sequence alignments and a structural comparison of each of the known PGK structures was consulted. 161 non-conserved side chains were truncated to alanine and residues 1–3 and 350–362 were removed because of variability among the PGK structures analyzed. The final model contained 380 residues, 210 of which were alanines, and retained original temperature factors. For comparison, the *T. brucei* PGK crystals contain four PGK subunits each with 419 residues, 37 of which are alanines.

A series of molecular replacement attempts were made with this search model using *AMoRe* (Navaza, 1994) and *X-PLOR* (Brünger, 1992). Rotation functions encompassing various resolution ranges between 10 and 3 Å were performed, but none of the potential orientations generated could be positioned in the unit cell by the translation function. Patterson correlation refinement (Brünger, 1990) in *X-PLOR* which allowed for relative movement of the N- and C-terminal domains was performed on each of the potential rotation function solutions. This procedure was intended to produce new search models with optimized hinge angles and orientations. However, none of the refined models could be positioned in the unit cell by the *X-PLOR* translation function.

3.2. *Molecular replacement by domains*

It seemed possible that some large-scale conformational change such as that predicted by the hinge-

bending hypothesis was present in the *T. brucei* PGK ternary complex and precluding molecular replacement. Hence, we created two separate search models representing the N and C domains of the optimized *B. stearothermophilus* PGK search probe described above (Fig. 2). Since the nature of the hinge region was unknown at this time, the *B. stearothermophilus* model was arbitrarily divided after residue 167 (analogous to residue 189 in *T. brucei* PGK). The final 12 residues at the PGK C terminus, numbering 383–394, (408–419 in *T. brucei*) cross over to the N domain and were included with the N-domain search probe (Fig. 3). Since four PGK subunits reside in the asymmetric unit, these N- and C-domain search models represented just 8.4 and 9.6% of the scattering mass, respectively.

Nevertheless, after a series of molecular replacement attempts incorporating various resolution ranges an intriguing list of potential rotation solutions was generated for the N domain using *AMoRe* (Navaza, 1994) (Table 3*a*). This rotation function utilized data between 8 and 3 Å and applied a maximum Patterson vector radius of 20 Å. The top two rotation solutions in this list were related by approximately 90° about the *Z* axis, a value which corresponded to a peak in the self-rotation function. These solutions were 4.8σ and 4.6σ above the mean and 0.9σ above the noise. Furthermore, these potential N-domain orientations also yielded the highest peaks in the translation function (each peak was 5σ above the mean and 0.5σ above the next highest peaks) with correlation coefficients of 22.9 and 23.0%. For comparison, the highest correlation coefficient obtained in the translation function for an incorrectly oriented probe was 21.9%.

We were unable to generate a C-domain rotation function list in which high-ranking solutions were related by a 90° rotation about the *Z* axis. However, two C domains could be positioned by the following procedure: first, a rotation function incorporating data between 6 and 3 Å generated a list of 50 potential orientations. Second, each of these 50 potential solutions was tried in a translation search which incorporated the two previously fixed N-domain probes (Table 3*b*). In this way, two C domains (ranked 16 and 25 in the rotation function) could be successfully positioned with peaks more than 10σ above the mean and yielding correlation coefficients of 29.2 and 29.5%. These two C-domain solutions appeared convincing because translation-function correlation coefficients for other potential C-domain orientations clustered about 26%. Furthermore, these C domains assumed orientations that were within 30° of the fixed N domains and were also related by approximately 90° about the *Z* axis. After two N-terminal and two C-terminal domain probes had been positioned in the asymmetric unit and subjected to rigid-body optimization, the correlation coefficient was 34.4% and the *R* factor was 52.3%.

Table 3. *Translation function results for top ranking orientations of the individual N- and C-domain B. stearothermophilus PGK search models in the T. brucei PGK unit cell*

The rotation search interval was 1° about α, β, γ (*CCP*4 Euler angles; model rotated by γ about *Z*, then by β about *Y*, then by α about new *Z*). The top correlation coefficient (c.cf.) and *R* factor obtained in the translation function for each potential orientation are shown.

(*a*) Translation function results for the top ten N-domain search probe orientations

Rotation function solutions 1 and 2, related by an ~90° rotation about the *Z* axis, could be correctly fixed in the unit cell. To fix a common origin for the first two N domains, the second rotation solution was positioned in a unit cell by a translation function calculation that already incorporated the first fixed N domain.

R.f.	Rotation (*CCP*4 Euler)			Translation (fractional)				*R*
solution	α	β	γ	Tx	Ty	Tz	c.cf.	factor
1	53.63	54.51	289.76	0.3103	0.2581	0.4483	22.9	56.4
2	137.01	52.10	290.82	0.0374	0.0670	0.2030	23.0	56.1
3	54.90	27.91	347.59	0.0562	0.2725	0.3888	21.0	56.4
4	7.05	61.45	243.21	0.3148	0.1721	0.3385	20.8	56.8
5	12.98	70.72	94.93	0.3351	0.1180	0.3487	22.0	56.5
6	123.57	88.48	179.96	0.2044	0.0016	0.2951	20.2	56.9
7	175.60	65.65	12.48	0.1954	0.3314	0.1122	21.8	56.4
8	55.40	90.00	−0.37	0.2019	0.0123	0.0400	20.3	57.1
9	162.86	58.08	295.15	0.1553	0.4196	0.3260	21.9	56.6
10	100.14	52.81	155.36	0.4757	0.0437	0.0427	20.6	56.9

(*b*) Translation function results for the top 25 C-domain search probe orientations in a unit cell incorporating two previously fixed N domains.

Rotation solutions 16 and 25 could be correctly fixed in the unit cell. These solutions are related to the previously identified N-domain solutions by a 30° rotation and are also related to each other by a 90° rotation about the *Z* axis.

R.f.	Rotation (*CCP*4 Euler)			Translation (fractional)				*R*
solution	α	β	γ	Tx	Ty	Tz	c.cf.	factor
1	116.20	76.49	37.27	0.3922	0.1725	0.3914	26.1	55.1
2	101.49	74.75	36.17	0.1307	0.0435	0.3853	26.2	55.1
3	30.21	76.61	34.70	0.3548	0.2334	0.4765	25.5	55.5
4	147.30	42.78	203.47	0.8544	0.5087	0.1797	26.1	55.3
5	106.17	64.94	68.91	0.1978	0.7592	0.8796	25.7	55.3
6	24.05	58.86	333.27	0.1282	0.7076	0.4305	26.7	55.0
7	126.29	74.91	274.84	0.3535	0.0613	0.7846	26.5	55.2
8	23.38	63.93	293.49	0.8345	0.1577	0.5198	26.0	55.2
9	115.87	29.89	0.30	0.4519	0.5063	0.7638	26.0	55.3
10	20.22	20.02	72.14	0.3948	0.1918	0.4365	25.9	55.4
11	59.21	78.28	9.10	0.3045	0.2753	0.5509	25.6	55.4
12	22.83	85.93	349.35	0.9712	0.9502	0.3075	26.0	55.3
13	4.79	65.25	128.24	0.5006	0.5680	0.6437	26.0	55.2
14	136.60	7.58	283.40	0.4722	0.5688	0.7508	26.2	55.1
15	2.60	37.61	200.30	0.1389	0.6422	0.9178	26.3	55.0
16	89.61	64.65	262.42	0.4703	0.4040	0.5339	29.2	54.1
17	65.89	56.11	218.73	0.7435	0.2036	0.5346	26.1	55.3
18	94.42	67.67	69.20	0.4404	0.6413	0.7257	25.7	55.2
19	156.56	55.44	221.06	0.2866	0.2272	0.4432	26.1	55.4
20	96.13	82.57	328.97	0.0269	0.0706	0.3833	26.3	55.1
21	15.57	40.48	82.40	0.3681	0.0794	0.7828	25.8	55.1
22	165.54	67.30	131.02	0.9012	0.1652	0.5453	26.1	55.2
23	34.82	42.71	223.80	0.8364	0.2026	0.5880	26.5	55.2
24	33.91	34.51	357.36	0.5967	0.5399	0.7437	26.0	55.2
25	177.19	65.54	265.41	0.3967	0.7327	0.7918	29.5	53.9

3.3. *A closed-form search model*

When the newly positioned N and C domains were visualized in the unit cell they appeared to form two complete subunits. From their relative orientation, it was apparent that the *T. brucei* PGK crystals contained enzyme in a new conformation. Specifically, the domains had rotated approximately 30° inward relative to their positioning in the *B. stearothermophilus* search model (Fig. 3). At this point only half of the unit-cell content had been identified; two other presumed subunits remained unidentified. However, on the basis of the new relative orientation and translation between the fixed N and C domains, a 'closed-state' search model containing both domains could be created. Fixing the two known subunits in the unit cell, we then attempted to orient and position this new PGK search model to find the position of the two still unknown subunits. In the rotation search, this probe yielded two new orientations with peak heights of 5σ and 4.3σ. In a translation search which incorporated the previously fixed probes, these orientations gave solutions that were 18.5σ and 14σ above the mean. With the addition of these two solutions the unit cell appeared complete in a packing analysis and the *R* factor, after rigid-body optimization in *AMoRe*, was 50.1%.

4. NCS averaging and refinement

The molecular replacement solutions were subjected to rigid-body and positional refinement in *X-PLOR* while maintaining strict non-crystallographic symmetry (NCS) constraints. This yielded an *R* factor and *R* free of 45.2 and 48.3%, respectively. Real-space density modification was then performed taking advantage of fourfold averaging, histogram matching and solvent flattening using the program *DM* (Collaborative Computational Project, Number 4, 1994). This treatment yielded a significant improvement in the figure of merit (from 0.512 to 0.801) and a resulting *SIGMAA*- (Collaborative Computational Project, Number 4, 1994) weighted map displayed good density for the N and C domains as well as for the MgADP substrate which had not yet been included in the model. Model building proceeded on the basis of this averaged map. After several cycles of refinement still maintaining strict NCS constraints, the *R* factor and *R* free dropped to 35.1 and 38.3%, respectively. At this point no density had yet been found for the 16-residue trypanosome specific insertion (residues 69–84), the N-terminal side of the first hinge (residues 175–195), or the 25 C-terminal residues. Density for these residues emerged and the *R* factor fell only after the strict non-crystallographic symmetry constraints were replaced with non-crystallographic symmetry restraints imposed separately on, and limited to, the N and C domains (*T. brucei* residues 5–174, and 210–395, respectively). This was presumably due to a 5° variation in hinge angle among the four subunits in the asymmetric unit. Using this protocol, positional refinement yielded an *R* factor and *R* free of 30.1 and 35.6% and the missing 61 residues could be built into density over several more rounds of refinement. The fully built model was subjected to further rounds of positional refinement, simulated annealing, a single overall anisotropic *B*-factor correction, and restrained *B*-factor refinement. The final model contains 12 488 protein atoms and 156 substrate atoms and has an *R* factor of 23.8% and an *R* free of 31.0% for data greater than 2σ between 8 and 2.8 Å.

Table 4. *Search probes positioned in the T. brucei PGK unit cell using AMoRe (Navaza, 1994)*

Relative scattering mass (with respect to the *T.brucei* PGK unit cell) and corresponding peak heights in the molecular replacement functions are shown. Rotation and translation searches used data between 8 and 3 Å, except for the *B. stearothermophilus* (*B.s.*) C-domain rotation function which used data from 6 to 3 A. The maximum Patterson vector radius used in the rotation functions was 20 Å for the *B. stearothermophilus* N- and C-domain probes and 30 Å for other models.

Search probe	Scattering mass (%)	R.f. peak (σ)	T.f. peak (σ)
B.s. PGK (Multi-ala) (Davies, 1990)	18	*	*
B.s. PGK N domain (multi-ala)	8.4	4.8	5.3
B.s. PGK C domain (multi-ala)	9.6	3.2	*
B.s. C domain in cell with two fixed N domains	9.6	—	11
Truncated *B.s.* PGK N domain (multi-ala)	5.8	4.5	5.2
Refined *T. brucei* PGK subunit	25	13	22
Cα atoms *T. brucei* PGK subunit	2.9	5.2	9.2
Cα atoms *T. brucei* PGK subunit ($B < 25$ Å^2)	1.6	4.5	7.9

* Probe could not be oriented/translated.

5. Molecular replacement with very small probes

5.1. *Keys to success of this molecular replacement protocol*

In the structure determination of closed-form PGK it was necessary to divide the open-form PGK search model into domains. This procedure was advantageous for several reasons. First, searching with discreet rigid domains accommodated inter-domain conformational differences between search model and target protein (in this case, a 30° hinge bend). Second, using search probes corresponding to different portions of the target molecule provided a means to validate potential rotation and translation solutions; *i.e.* correct solutions should be consistent between the two domains in terms of approximate orientation, placement in the unit cell, and

crystal packing. By combining this information with expected non-crystallographic symmetry relationships we were able to differentiate weak rotation and translation solutions from noise, and position four very small probes.

Perhaps the most surprising aspect of this structure determination is that it was possible to orient and position an N-domain PGK fragment containing just 8.4% of the scattering mass in the asymmetric unit. Two such fragments could be positioned with relative orientations that were consistent with the self-rotation function. The fact that these solutions were the top peaks in both the rotation and translation functions indicates that this probe could have been positioned without knowledge of non-crystallographic symmetry. To the best of our knowledge, this probe contained the smallest proportion of scattering mass of any non-Cα fragment that could be positioned in the unit cell of a protein crystal.

We were interested in identifying aspects of the PGK N domain which make it such an effective search probe. One aspect can be recognized by a comparison with the PGK C domain, a significantly less effective search probe that could only be positioned after two N domains had been fixed in the unit cell. When the *B. stearothermophilus* PGK search model was divided into N and C domains in the initial stages of molecular replacement, little was known about the hinge points for this flexible molecule since all existing crystal structures depicted open-form enzyme. Hence, the model was arbitrarily divided at what we know in retrospect is the N-terminal side hinge point. Consequently, the C-domain probe contains the hinge region of open-form PGK in a different conformation than in closed form *T. brucei* PGK. The r.m.s.d. over Cα atoms between the C-domain probe and the *T. brucei* PGK ternary complex is 1.8 Å. This contrasts with an r.m.s.d. over Cα atoms for the N-terminus of 1.1 Å and explains, at least in part, the relative difficulty in positioning the C-domain probe. Hence, the structural conservation of the *B. stearothermophilus* PGK search model and in particular the N domain was critical for the successful rotation and translation of this very small search probe. This finding stresses the importance of using a search model that is as well conserved as possible, with non-conserved side chains and loops removed to minimize incorrect vectors from the model Patterson. Indeed, we found that an N-domain PGK search probe *which retained* the 78 non-conserved side chains did not yield high-ranking peaks in the rotation and translation functions.

Another partial explanation for the effectiveness of the N-domain search probe is related to the observation that the two N and two C domains initially identified by molecular replacement belong to the same two subunits. An analysis of the temperature factors in the various subunits of *T. brucei* PGK provides a likely explanation for this finding: the two subunits identified first both have average B factors of 29.5 Å^2, whereas the subunits which were only located once a complete 'closed'-state search model had been created have average B factors of 34.6 and 46.7 Å^2. The atoms in the subunits with higher average B factors are each distributed over a larger volume and consequently contribute less scattering intensity with increasing scattering angle. The corrected scattering factor (f) as a function of temperature factor (B) and resolution (d) is given by:

$$f(B, d) = f_o \exp[-B/(4d^2)].$$

At a resolution of 4 Å, it follows that a subunit with an average temperature factor of 46.7 Å^2 would contribute approximately 23% less to the scattering factors than a subunit with an average temperature factor of 29.5 Å^2. Furthermore, at this resolution, the less ordered subunit will theoretically contribute 41% less to the Patterson function used in the rotation and translation functions since the Patterson depends on intensities. Consequently, in the molecular replacement functions calculated with data between 8 and 3 Å (or 6 and 3 Å, as in the case of the C-domain rotation function) the subunits with lower B factors which contribute more to the discreet scattering will bias the molecular replacement functions, enhancing their particular solutions. In this way, the effectiveness of the small search probes in locating the subunits with lower overall temperature factors, is increased. By a similar phenomenon, atoms in the search model with low temperature factors will dominate in the molecular replacement functions. A search model which retains B factors will accentuate the more well ordered atoms of the protein core and the polypeptide backbone. Thus it is generally advantageous to retain B factors, since the protein core and backbone atoms are typically the best conserved parts of a search model. Indeed, we have found that N- and C-domain search probes in which B values have been reset to a constant value (B factors in the range between 20 and 40 Å^2 were tried) do not yield as high peaks in the rotation and translation functions as probes retaining their original crystallographic temperature factors.

5.2. *Testing the limits of molecular replacement*

Since the finding that a very small main-chain search probe can be oriented and fixed in the unit cell is likely to have implications towards other potential molecular replacement protocols, we were interested in determining the minimum size for such a probe, at least in our system. Starting with the *B. stearothermophilus* N domain we removed successive stretches of polypeptide that were (relatively) less well conserved and tried to rotate and translate the result. In this way, a multi-alanine *B. stearothermophilus* PGK fragment (Fig. 2) containing just 5.8% of the scattering mass was identified as the minimal main-chain search probe that

could be correctly oriented and positioned in the *T. brucei* PGK unit cell. In the *AMoRe* rotation function performed with this minimized probe, the first and fourth ranking solutions corresponded to correctly oriented probes (the top peak was 4.8σ above the mean). Furthermore, the top rotation solution could be correctly translated (yielding a 5.2σ peak, 0.5σ greater than the next) with a correlation coefficient of 21.5%. After fixing the first solution in the unit cell, the fourth ranking orientation could be successfully translated (9.2σ peak, 1.7σ greater than next) with a correlation coefficient of 25.1%. Hence, it appears that the *T. brucei* PGK structure determination could have proceeded with an N-domain search probe comprising just 5.8% of the scattering mass.

We were also interested in identifying the absolute minimum collection of atoms that could be oriented and positioned in the *T. brucei* PGK unit cell. We postulated that such a collection of atoms should (1) be structurally conserved with respect to the target molecule, (2) be well ordered and (3) map out a pattern uniquely representative of the target structure in three-dimensional Patterson space such that unambiguous solutions in the rotation and translation functions could be obtained. This final requirement is particularly well satisfied by the Cα trace which contains sparsely sampled atoms that yield long inter-atomic vectors in the Patterson map. The Cα trace from one subunit of the refined *T. brucei* PGK ternary complex satisfies conditions (1) and (3) and it was, therefore, chosen as a starting search point. To define a collection of well ordered atoms, Cα atoms with temperature factors greater than 25 Å^2 were removed. Although the resulting search probe contained just 231 atoms, representing 1.6% of scattering mass in the *T. brucei* PGK asymmetric unit, these atoms could easily be fixed in the unit cell. A rotation search was performed using data from 8 to 3 Å and the third highest rotation solution could be positioned correctly with a 7.9σ peak in the translation function and a correlation coefficient of 15.3%, 1.3% higher than that obtained for incorrect orientations.

6. Conclusions

The sequential addition of fragments containing extremely small portions of the unit cell makes this structure determination a unique application of the molecular replacement method. The original structure solution relied on the ability to correctly position a main-chain search probe containing just 8.4% of the scattering mass. This surprising finding has relevance for other projects in which partial search models are available. Even in cases where a more significant probe can be used, dividing the model into discreet rigid domains may still be an effective strategy. Individual domains can then be brought into the molecular replacement functions in parallel or in series in an attempt to identify a complementary set of correctly fixed solutions. The knowledge that these subunits must fit into the unit cell in a biologically meaningful way provides an unbiased means for filtering weak solutions from noise.

Through a retrospective analysis, this work also provides insight into aspects of a search model and crystal system which facilitate the rotation and translation functions. Characteristics of an effective search probe include high structural conservation, well ordered atoms, and a sparse, yet spatially extended, representation of the target structure. A collection of 231 Cα atoms from the refined *T. brucei* PGK structure (1.6% of the scattering mass) with these attributes could be correctly oriented and positioned in the unit cell, whereas search models which retained non-conserved features or had constant temperature factors were less effective in the molecular replacement functions.

We would like to thank Drs F. Athappilly, E. Merritt and E. Pohl for helpful discussions and advice. We are indebted to Dr Paul Michels for providing the *T. brucei* P6K overexpression system. The work was supported in part by the Murdock Charitable Trust. BEB was supported by the Poncin Fund and Howard Hughes Medical Institute Graduate Student Support.

References

Banks, R. D., Blake, C. C. F., Evans, P. R., Haser, R., Rice, D. W., Hardy, G. W., Merrett, M. & Phillips, A. W. (1979). *Nature (London)*, **279** 773–777.

Bernstein, B. E., Michels, P. A. M. & Hol, W. G. J. (1997). *Nature (London)*, **385**, 275–279.

Blake, C. C. F. & Evans, P. R. (1974). *J. Mol. Biol.* **84**, 585–601.

Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Brünger, A. T. (1991). *Acta Cryst.* A**47**, 195–204.

Brünger, A. T. (1992). *X-PLOR*, Version 3.1, Yale University, New Haven, CT, USA.

Brünger, A. T., Leahy, D. J., Hynes, T. R. & Fox, R. O. (1991). *J. Mol. Biol.* **221**, 239–256.

Chantalat, L., Wood, S. D., Rizkallah, P. & Reynolds, C. D. (1996). *Acta Cryst.* D**52**, 1146–1152.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cygler, M. & Anderson, W. F. (1988). *Acta Cryst.* A**44**, 38–45.

Davies, G. J., Gamblin, S. J., Littlechild, J. A., Dauter, Z., Wilson, K. S. & Watson, H. C. (1994). *Acta Cryst.* D**50**, 202–209.

Harlos, K., Vas, M. & Blake, C. C. F. (1992). *Proteins*, **12**, 133–144.

Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Misset, O., Bos, O. J. M. & Opperdoes, F. R. (1986). *Eur. J. Biochem.* **157**, 441–453.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Oh, B.-H. (1995). *Acta Cryst.* D**51**, 140–144.
Otwinkowski, Z. (1993). In *Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Burley, pp. 56–62. Warrington: Daresbury Laboratory.
Pickover, C. A., McKay, D. B., Engelman, D. M. & Steitz, T. A. (1979). *J. Biol. Chem.* **254**, 11323–11329.
Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Schierbeek, A. J., Renetseder, R., Dijkstra, B. W. & Hol, W. G. J. (1985). In *Molecular Replacement. Proceedings of the Daresbury Study Weekend, 15–16 February 1985*, edited by P. A. Machin, pp. 16–21. Warrington: Daresbury Laboratory.
Turkenburg, J. P. & Dodson, E. J. (1996). *Curr. Opin. Struct. Biol.* **6**, 604–610.
Watson, H. C., Walker, N. P. C., Shaw, P. J., Bryant, T. N., Wendell, P. L., Fothergill, L. A., Perkins, R. E., Conroy, S. C., Dobson, M. J., Tuite, M. F., Kingsman, A. J. & Kingsman, S. M. (1982). *EMBO J.* **1**, 1635–1640.
Zhang, X. J. & Matthews, B. W. (1994). *Acta Cryst.* D**50**, 675–686.