Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The importance of alignment accuracy for molecular replacement

Robert Schwarzenbacher,* Adam Godzik, Slawomir K. Grzechnik and Lukasz Jaroszewski*

Bioinformatics Core, Joint Center for Structural Genomics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Correspondence e-mail: roberts@sdsc.edu, lukasz@sdsc.edu

Received 23 February 2004
Accepted 26 April 2004

Many crystallographic protein structures are being determined using molecular replacement (MR), a model-based phasing method that has become increasingly important with the steady growth of the PDB. While there are several highly automated software packages for MR, the methods for preparing optimal search models for MR are relatively unexplored. Recent advances in sequence-comparison methods allow the detection of more distantly related homologs and more accurate alignment of their sequences. It was investigated whether simple homology models (without modeling of unaligned regions) based on alignments from these improved methods are able to increase the potential of MR. 27 crystal structures were determined using a highly parallelized MR pipeline that facilitates all steps including homology detection, model preparation, MR searches, automated refinement and rebuilding. Several types of search models prepared with standard sequence–sequence alignment (*BLAST*) and more accurate profile–sequence and profile–profile methods (*PSI-BLAST*, *FFAS*) were compared in MR trials. The analysis shows that models based on more accurate alignments have a higher success rate in cases where the unknown structure and the search model share less than 35% sequence identity. It is concluded that by using different types of simple models based on accurate alignments, the success rate of MR can be significantly increased.

1. Introduction

Determining crystallographic structures by molecular replacement (MR; Rossmann, 2001) uses approximate structural models to provide initial estimates of the phases necessary for the determination of the structure. If a sufficiently accurate model is available and can be correctly placed in the unit cell, an initial set of phases can be derived and the final crystallographic structure can be obtained through rebuilding and refinement. MR has an advantage over experimental phasing techniques because it requires only one data set of reflections obtained with a native protein crystal. This is considerably less resource-intensive than multiple-wavelength experiments with substituted protein crystals. As a result, there is considerable interest in methods that can extend the applicability of MR. MR methods pioneered by Hoppe (1957) and Rossmann & Blow (1962) require the identification of the correct orientation and position of the structural model in the asymmetric unit of a new crystal. Currently, several automated computational algorithms for solving this problem are available in popular programs such as *Beast* (Read, 2001), *AMoRe* (Navaza, 2001), *XPLOR*/*CNS* (Brünger *et al.*, 1998), *MOLREP* (Vagin & Teplyakov, 2000), *EPMR* (Kissinger *et al.*,

DOI: 10.1107/S0907444904010145

1999), *Queen of Spades* (Glykos & Kokkinidis, 2000) and *SOMoRe* (Jamrog *et al.*, 2003). The success rate of these MR methods depends critically on the quality of the model used. MR has been accomplished with models covering only a small fraction (<30%) of the molecule (Bernstein *et al.*, 1997), but experience has shown that in order for the procedure to be successful, a significant portion of the molecule (>60%) is usually required and the differences between the coordinates of the model and the molecule must be small [usually within a root-mean-square distance of C^{α} atoms (CαRMSD) below 2.5Å].

To our knowledge, the requirements for optimal search models have not been systematically explored. However, several interesting ideas have been tested on individual cases or on small sets of structures (Kleywegt, 1996). Some of these ideas include (i) using individual domains in MR searches for molecules where flexibly linked multiple domains are present, (ii) removing or cutting back of residues/regions with high temperature factors, (iii) omission of regions where sequence conservation is low and (iv) using composite (multiple) search models (Chen, 2001).

In this contribution, we examine the importance of alignment accuracy and side-chain modeling in MR.

Homology modeling is based on the observation that proteins with similar sequences fold into similar structures. In the first step of the modeling process one has to identify a homologous structure (the template) and to define the alignment: a set of residue-by-residue equivalencies between the target sequence and the template sequence. This is the most crucial step in the process and any errors at this stage lead to significant errors in the models (Sali *et al.*, 1995). Because model quality directly depends upon the accuracy of the underlying sequence alignment, recent progress in the field of distant homology detection, such as fold-recognition methods, was expected to benefit MR (Jones, 2001).

It is well known that simple sequence-alignment methods such as *BLAST* (Altschul *et al.*, 1990) become less accurate when the sequence identity of homologous sequences falls below 30%. Producing the best model in this region requires more sensitive methods. We tested three alignment methods of increasing accuracy: *BLAST*, *PSI-BLAST* (Altschul *et al.*, 1997) and *FFAS* (*Fold and Function Assignment System*; Rychlewski *et al.*, 2000). While *BLAST* aligns sequences one by one, *PSI-BLAST* includes evolutionary information from a set of homologous sequences to improve its sensitivity. It calculates a sequence profile from a multiple-sequence alignment obtained through a sequence-similarity search against a non-redundant sequence database. The profile is iteratively improved by adding more homologous sequences in subsequent ‘profile–sequence’ searches against a non-redundant sequence database. *FFAS* goes one step further by using sequence profiles for the query and all sequences in the database. Thus, it uses evolutionary information for both the query and the database sequences. It has been shown that when comparing representative pairs of remote homologs selected from the Structural Classification of Proteins (SCOP; Murzin *et al.*, 1995) database, *FFAS* surpasses *BLAST* and also *PSI-BLAST* in the number of correct fold predictions (Rychlewski *et al.*, 2000). *FFAS* also yields more accurate alignments than *PSI-BLAST* when alignments of the same lengths are compared (Jaroszewski *et al.*, 2000). Moreover, *FFAS* alignments are usually longer than those from *PSI-BLAST* and therefore closer in lengths to structural alignments.

Therefore, more accurate alignment methods should increase the number of potential MR targets because they provide more accurate models and in some cases may provide models where other methods fail. We tested this hypothesis by comparing models based on alignments of increasing accuracy obtained with *BLAST*, *PSI-BLAST* and *FFAS* on a set of 31 MR data sets.

2. Methods

2.1. Test set

The analysis was performed on MR data sets obtained at the Joint Center for Structural Genomics (JCSG) and included 25 targets from *Thermotoga maritima*, four targets from other bacteria and one target from mouse. Thus, our study focuses on typical bacterial proteins containing one or two tightly interacting domains.

2.2. Sequence-similarity searches

Homology searches against protein sequences from the Protein Data Bank (PDB; Berman *et al.*, 2000) were carried out for each MR target sequence using *BLAST*, *PSI-BLAST* (http://www.ncbi.nlm.nih.gov/BLAST/) and *FFAS* (http://ffas.ljcrf.edu/). *BLAST* searches were performed directly on protein sequences from the PDB. *PSI-BLAST* searches were executed according to a protocol called *PDB-BLAST* (Li *et al.*, 2002), in which an iterative *PSI-BLAST* search is executed against the non-redundant sequence database (NR) and the resulting profile is used to search the PDB. *FFAS* profiles were prepared for all sequences from the PDB and then searched with profiles of target sequences.

An *E* value of 0.001 is a widely accepted threshold for homologs detected with *BLAST* and *PSI-BLAST*. For *FFAS*, a score of −9.5 is the threshold of reliable predictions. All targets included in this study had at least one homolog in the PDB with an *FFAS* score better (lower) than −15. *PSI-BLAST* was able to obtain templates for all targets and *BLAST* was not able to find a template for TM1459 and TM1244.

The structure with the highest sequence identity to the target was used as a template for model building; thus, the same template was used for the *BLAST*, *PSI-BLAST* and *FFAS* alignments. The alignments from the three methods covered at least 70% of each target, including all structural domains.

Characteristics of the target–template sequence alignments obtained with different methods are given in Table 1. Columns CB, CP and CF contain the percentage of the target sequence aligned with the template by *BLAST*, *PSI-BLAST* and *FFAS*,

Table 1
Statistics for JCSG MR projects.

Column captions: Target, TIGR or GeneBank ID (gi No.) of the target protein; L, target-sequence length; SG, crystallographic space group; M, number of molecules in the asymmetric unit; R, resolution (Å) of the crystallographic data set; o/a, number of observations per atom; Template PDB, the closest homolog with known structure (PDB or GeneBank ID); TR, resolution of the template structure; Id, sequence identity between target and template; CB, CP and CF, percentage of the target sequence covered by the alignment from *BLAST*, *PSI-BLAST* and *FFAS*, respectively; GB, GP and GF, number of gaps in the alignment from *BLAST*, *PSI-BLAST* and *FFAS*, respectively; t, standard MR search with standard template; T represents an exhaustive MR search with standard template, A with polyalanine template, Bm and Ba with mixed and all-atom models based on *BLAST* alignment, Pm and Pa with mixed and all-atom models based on *PSI-BLAST* alignment, Fm and Fa with mixed and all-atom models based on *FFAS* alignment, where '+' indicates successful MR phasing and automated refinement; CαRMSD and Lali are the root-mean-square distance between C^{α} atoms of the refined target and the template and the length of the structural alignment (both calculated with *DALI*); Target PDB, PDB code of the solved MR structures (if already deposited in the PDB).

Target	L	SG	M	R	o/a	Template PDB	TR	Id	CB	CP	CF	GB	GP	GF	t	T	A	Bm	Ba	Pm	Pa	Fm	Fa	CαRMSD	Lali	Target PDB
TM0816	142	$C2$	2	2.10	9.7	1lj9	1.60	17	76	92	97	2	0	0										?	?	
TM1459	114	$P32$	2	1.75	11.8	1lr5	1.90	18	—	74	82	—	2	2								+	+	1.8	110	1o5u
TM1287	121	$C2$	2	1.70	8.9	1vj2	1.65	18	81	83	84	4	3	2		+	+	+	+	+	+	+	+	1.7	110	1o4t
17391249	248	$P622$(?)	1	2.23	8.7	1fez	3.00	19	82	95	95	9	7	7										?	?	
15079298	142	$P1$	1	1.35	15.7	1ahq	2.30	19	77	92	93	3	4	4	+	+	+	+	+	+	+	+	+	1.8	130	
TM0820	395	$P2_1$	2	1.78	10.0	1o2d	1.30	24	72	96	97	6	7	6						+		+		2.4	340	
TM1244	82	$P222$(?)	4	2.50	6.8	1gtd	2.56	25	—	93	98	—	1	1										?	?	
TM0262	366	$P4_22_12$	1	2.70	4.8	1jqj	2.90	26	99	99	100	4	3	3									+	2.2	345	
TM1419	382	$I222$	1	1.58	22.5	1gr0	1.95	26	92	97	98	11	8	8			+	+	+	+	+	+	+	2.0	317	1vjp
TM0748	265	$I222$	1	1.70	16.7	1i9g	1.98	28	90	94	99	4	3	3	+	+	+	+	+	+	+	+	+	1.7	258	1o54
TM0222	266	$C2$	2	2.48	5.5	1oxs	1.65	30	81	99	91	8	7	6										?	?	
TM1128	182	$H32$	8	2.35	8.1	1eum	2.05	30	95	95	98	1	0	0	+	+	+	+	+	+	+	+	+	0.9	161	
TM1385	448	$I2_12_12_1$	3	2.90	6.8	1b0z	2.30	31	87	99	99	6	7	7			+	+	+	+	+	+	+	1.4	421	
TM1645	273	$I222$	2	2.80	6.9	1qpn	2.60	31	92	98	99	4	6	5								+	+	2.5	261	1o4u
TM0066	205	$C222_1$	3	2.30	6.8	1eua	1.95	31	95	96	98	5	3	3				+		+		+		1.4	199	
TM0166	430	$P6_122$	1	2.75	8.9	1fgs	2.40	32	98	98	99	9	6	8	+	+	+	+	+	+	+	+	+	2.0	380	1o5z
TM0919	138	$P2_1$	4	1.80	12.9	1ml8 × 2	2.60	33	97	96	100	6	3	3									+	1.4	124	
TM0604	141	$F222$	1	2.40	10.0	1qvc	2.20	34	69	77	91	1	1	1		+				+	+	+	+	0.8	78	
TM1169	237	$P2_12_12_1$	4	2.50	4.3	1i01	2.60	34	96	97	98	7	2	3		+	+	+	+	+	+	+	+	1.6	212	1o5i
17130499	345	$P2_1$	2	2.50	4.8	1kgz	2.40	35	93	96	99	4	4	5	+	+	+	+	+	+	+	+	+	1.9	321	
TM0208	466	$C2$	4	2.30	10.3	1e0t	1.80	41	98	99	100	5	3	4			+	+	+	+	+	+	+	1.5	439	
17130350	381	$P2_12_12_1$	2	2.00	11.8	1h0c	2.50	42	96	99	100	2	2	2	+	+	+	+	+	+	+	+	+	0.6	381	1vjo
TM1521	294	$P2_12_12_1$	2	2.04	10.4	1dhp	2.50	43	99	99	99	3	3	2	+	+	+	+	+	+	+	+	+	1.2	289	1o5k
TM1255	377	$P2_12_12_1$	2	1.85	8.9	1bkg	2.60	43	95	99	100	5	3	3	+	+	+	+	+	+	+	+	+	1.7	370	1o4s
TM1718	220	$P2_12_12_1$	5	3.30	9.7	1rpx	2.05	44	95	97	98	2	2	2	+	+	+	+	+	+	+	+	+	0.9	213	
TM1097	313	$F432$	1	2.50	5.9	1a1s	2.70	48	97	99	99	2	2	2	+	+	+	+	+	+	+	+	+	0.9	308	
15159614	169	$P2_1$	2	1.60	18.0	13421216	2.20	48	96	96	96	1	1	1				+	+	+	+	+	+	1.3	158	
TM0741	161	$P2_1$	4	2.20	11.4	1b6t	1.80	49	94	97	98	0	0	0	+	+	+	+	+	+	+	+	+	1.0	154	
TM0446	171	$I422$	1	1.80	11.7	1qcz	1.50	49	93	93	92	0	1	1	+	+	+	+	+	+	+	+	+	1.5	155	1o4v
TM0721	209	$C2$	4	2.30	6.0	1i5e	3.00	62	99	99	100	0	0	0	+	+	+	+	+	+	+	+	+	1.9	207	1o5o
TM0720	427	$P6_522$	1	2.86	4.6	1kl1	1.93	63	95	96	100	2	3	3	+	+	+	+	+	+	+	+	+	0.7	405	

respectively (the 'coverage' of the target by the alignment). Columns GB, GP and GF show the number of gaps in the *BLAST*, *PSI-BLAST* and *FFAS* alignments, respectively.

2.3. Preparation of search models

All search models were prepared based on alignments derived with the programs mentioned above. All manipulations in the coordinate file, such as the mutation of side-chain residues and optimization of the side-chain packing, were carried out using the 'modeling' option of the program *WHATIF* (Vriend, 1990), which performs exhaustive optimization of side-chain packing but does not build unaligned regions (fragments corresponding to gaps in the alignment).

Four types of models were tested for each alignment.

(i) Standard template. Coordinates and B factors of the structural domains homologous to the target were extracted from the PDB file of the structure (without ligand and water molecules) and directly used for MR searches.

(ii) Polyalanine template. Same as standard template but with all side chains mutated to alanine. It was prepared using the 'cleanup, convert to polyalanine and shift to origin' option of the program *MOLREP* (Vagin & Teplyakov, 2000).

(iii) Mixed model. This is a simple model where all the non-conserved side chains were mutated to serine and conserved residues were transferred from the template (preserving the side-chain rotamer). The rationale behind this type of model is that rotamers of conserved side chains are expected to be similar to those in the template, while rotamers of non-conserved side chains are more difficult to predict. Mixed models are a compromise between the risk of incorrect side-chain predictions and model completeness. In cases of sequence identities below 35% such models contain about 15–25% less atoms than all-atom models. Mutation to serine instead of alanine is intended to compensate for the loss of side-chain atoms beyond C^{β}. The mixed model depends on the accuracy of the sequence alignment, since the alignment determines model boundaries and conserved residues.

(iv) All-atom model. This is a primitive homology model in which all non-conserved side chains are mutated according to the target sequence. Template main-chain coordinates in the aligned regions were not altered and no rebuilding of gaps and

Table 2
MR results obtained with mixed and all-atom models obtained with *FFAS* alignments.

Column captions: %sm, percentage of scattering matter represented by the search model; #sc, number of side chains corresponding to the target sequence; %χ_1, percentage of χ_1 angles predicted to be within 15° of the real structure; %χ_2, percentage of χ_2 angles predicted to be within 15° of the real structure; $R_{\rm free}$, $R_{\rm free}$ value after 500 steps of automated refinement with *REFMAC*5; NS, no data because target is not solved; NR, no data because target is not finalized.

			Mixed model					All-atom model					
Target	Id	L	%sm	#sc	%χ_1	%χ_2	$R_{\rm free}$	%sm	#sc	%χ_1	%χ_2	$R_{\rm free}$	CαRMSD
TM0816	17	142	NS	21	NS	NS	NS	NS	125	NS	NS	NS	?
TM1459	18	114	28	18	61	70	46	36	69	39	30	44	1.8
TM1287	18	121	34	16	67	50	45	39	91	44	26	41	1.7
17391249	19	248	NS	36	NS	NS	NS	NS	191	NS	NS	NS	?
15079298	19	142	61	22	42	56	43	77	117	38	36	43	1.8
TM0820	24	395	34	57	70	44	46	42	281	47	40	NS	2.4
TM1244	25	82	NS	17	NS	NS	NS	NS	69	NS	NS	NS	?
TM0262	26	366	80	78	21	17	NS	97	314	20	18	49	2.2
TM1419	26	382	80	65	52	28	47	95	267	42	35	48	2.0
TM0748	28	265	88	53	67	50	44	98	211	43	32	45	1.7
TM0222	30	266	NS	51	NS	NS	NS	NS	201	NS	NS	NS	?
TM1128	30	182	10	42	70	68	45	12	140	57	52	36	0.9
TM1385	31	448	39	68	NR	NR	48	47	348	NR	NR	48	1.4
TM1645	31	273	37	65	28	15	49	47	224	43	23	47	2.5
TM0066	31	205	33	36	47	50	39	40	154	37	26	NS	1.4
TM0166	32	430	73	105	59	55	49	90	356	46	35	49	2.0
TM0919	33	138	38	40	55	24	NS	46	115	50	38	49	1.4
TM0604	34	141	81	36	NR	NR	46	89	118	NR	NR	43	0.8
TM1169	34	237	19	53	82	67	46	23	188	55	45	41	1.6
17130499	35	345	41	72	47	42	46	49	248	45	33	44	1.9
TM0208	41	466	21	147	NR	NR	47	25	396	NR	NR	45	1.5
17130350	42	381	42	114	62	47	42	48	294	54	45	42	0.6
TM1521	43	294	42	91	67	56	40	48	236	59	48	38	1.2
TM1255	43	377	41	116	62	58	43	47	311	52	40	38	1.7
TM1718	44	220	16	75	NR	NR	44	19	181	NR	NR	39	0.9
TM1097	48	313	85	110	56	48	44	98	258	45	36	39	0.9
15159614	48	169	44	62	80	59	49	50	138	66	51	37	1.3
TM0741	49	161	22	63	71	41	44	24	135	63	41	43	1.0
TM0446	49	171	74	47	76	68	44	87	118	66	53	42	1.5
TM0721	62	209	23	97	65	51	45	25	168	55	41	37	1.9
TM0720	63	427	89	194	67	45	46	98	326	46	39	47	0.7

insertions was included. The *WHATIF* modeling routine performs exhaustive optimization of side-chain packing. These models should have the best chances in MR in cases where side-chain rotamers can be correctly predicted because they contain most of the atoms present in the target structure. Obviously, the accuracy of this model depends on the accuracy of the alignment.

(v) Oligomer models. In cases where the cell-content analysis indicated more than one molecule in the asymmetric unit, additional mixed and all-atom models of possible oligomers were built based on the biologically relevant oligomer of the template structure.

2.4. Molecular-replacement searches and automated refinement

2.4.1. Standard approach. First, the MR program *MOLREP* (Vagin & Teplyakov, 2000) was used to determine the position of the model in the asymmetric unit. *MOLREP* was run with default parameters, using the option 'cleanup, set *B* values related to accessibility and shift to origin'. The top solutions for each model were subjected to 30 steps of rigid-body refinement, 500 steps of restrained refinement using *REFMAC*5 (Murshudov *et al.*, 1997) and the 'automated model building starting from existing model' routine of *ARP*/*wARP* (Perrakis *et al.*, 2001).

2.4.2. Exhaustive approach. *MOLREP* runs were repeated, following the author's recommendation, with different values for the *similarity* parameter (values: 0.3, 0.5, 0.7, 1.0) and the *completeness* parameter (values: 0.3, 0.5, 0.7, 1.0). If *MOLREP* did not find the correct solution, the program *EPMR* (Kissinger *et al.*, 1999), which uses a full six-dimensional molecular-replacement search, was used. *EPMR* was run with 500 generations and several combinations of resolution ranges with high-resolution limits of 3, 4 and 5 Å and low-resolution limits of 10, 15 and 20 Å. In order to perform exhaustive MR searches, we have developed an automated MR pipeline. The *FFAS* fold-recognition algorithm has been linked to the homology-modeling routine from the *WHATIF* package, the MR programs *MOLREP* and *EPMR*, rigid-body and restrained refinement from *REFMAC*5 and automated rebuilding from *ARP*/*wARP*. The pipeline has been implemented on an 80 CPU Linux cluster with a Portable Batch System (PBS). This pipeline is designed to allow hundreds of MR searches with different models and parameter sets. In cases where the top MR solutions did not converge, we derived coordinates for up to 100 solutions of each MR run and subjected them to subsequent refinement in *REFMAC*5 and *ARP*/*wARP*.

For all MR trails the convergence of the restrained refinement in *REFMAC*5 and *ARP*/*wARP* ($R_{\rm free}$ < 0.48 and FOM > 0.40) was used as an indicator for a correct MR solution. In all cases the structures were manually rebuilt and refined according to crystallographic standards. Coordinates and experimental structure factors for all the targets are being deposited with the PDB (for accession codes see Table 1, column Target PDB) and are also available from the JCSG. Coordinates for search models and structure factors used in this study are also available from the authors (http://www.jcsg.org/lukasz/mr_models).

2.5. Manual rebuilding and analysis of the results

All model building, structure analysis and superpositions were carried out in the programs *O* (Jones *et al.*, 1991) and *TOP* (Collaborative Computational Project, Number 4, 1994). The *DALI* server (Holm & Sander, 1995) was used to compare the templates and the final target structures. The lengths of *DALI* alignments and values of CαRMSD are shown in Table 1. The *ROTAMER* program (Collaborative Computational Project, Number 4, 1994) was used to calculate side-chain rotamer angles (χ_1 and χ_2) in all MR models and the final structures. Subsequently, predicted rotamer angles from the model were compared with rotamer angles from the final structure. For each side chain, a prediction of the χ_1 or χ_2 angle was counted as correct when the difference between the predicted and real value was smaller than 15°. Rotamer statistics for mixed models were only calculated for conserved side-chain residues. The percentages of correctly predicted χ_1 or χ_2 angles for mixed and all-atom models are shown in Table 2.

3. Results

Results of 31 MR phasing attempts with four different types of models derived from *BLAST*, *PSI-BLAST* and *FFAS* alignments are shown in Table 1. The 'standard approach' (column t in Table 1) was effective for ten out of 11 targets with more than 35% sequence identity to the template, but ineffective below that level (success in only four out of 20 cases). Exhaustive MR searches with standard templates resulted in four additional correct MR solutions in the sequence identity range below 35% (column T, Table 1). Exhaustive MR searches with polyalanine templates (column A, Table 1) gave identical results above 35% sequence identity and solved two more structures below 35%. Exhaustive MR searches with mixed and all-atom homology models based on *BLAST* alignments (columns Bm and Ba, Table 1) found one additional solution above and below 35% sequence identity. Mixed and all-atom homology models based on *PSI-BLAST* alignments (columns Pm and Pa, Table 1) found two additional solutions below 35% sequence identity, namely TM0820 and TM0604. Finally, exhaustive MR searches with mixed and all-atom homology models based on *FFAS* alignments (columns Fm and Fa, Table 1) solved four additional structures below 35% sequence identity. Four out of 31 structures remain unsolved despite extensive modeling and MR trials.

The most interesting cases are those where MR was only possible with mixed and all-atom models. These models are, in contrast to standard templates and polyalanine templates, dependent on the accuracy of the sequence alignment. These examples (TM0919, TM0066, TM1645, TM0262, TM0820 and TM1459) are analyzed and described below in more detail.

TM0066 is a 2-dehydro-3-deoxyphosphogluconate aldolase from *T. maritima*, which belongs to the TIM β/α-barrel fold (resolution 2.30 Å, space group $C222_1$, three molecules in the asymmetric unit, unit-cell parameters a = 79.54, b = 113.19, c = 128.40 Å, α = 90.00, β = 90.00, γ = 90.00°). The closest structural template was identified by *FFAS*, *PSI-BLAST* and *BLAST* as Kdpg aldolase from *Escherichia coli* (PDB code 1eua; 31% sequence identity). The three alignments are very similar, with small differences in coverage (±3%). In this case only the mixed models from each alignment method provided a correct MR solution. The all-atom models failed at the MR step most likely because of too many wrong side-chain predictions (37% of correct χ_1 values in all atom model *versus* 47% of correct χ_1 values in the mixed model).

TM0262 is the β-subunit of DNA polymerase III (DnaN) from *T. maritima* (resolution 2.7 Å, space group $P4_22_12$, one molecule in the asymmetric unit, unit-cell parameters a = 91.85, b = 91.85, c = 113.72 Å, α = 90.00, β = 90.00, γ = 90.00°). The closest structural template was identified by all alignment methods as the β-chain of DNA polymerase III from *E. coli* (PDB code 1jqj; 26% sequence identity). *BLAST*, *PSI-BLAST* and *FFAS* alignments arrive at the same coverage but differ by three small one-residue shifts. All models based on *BLAST* and *PSI-BLAST* alignments failed at the MR step. This case shows that even small errors in the model resulting from one-residue shifts in the alignment can cause the failure of MR.

All-atom and mixed models based on the *FFAS* alignment gave very similar and correct MR solutions (differences in rotation solution of less than 1°), but only the all-atom model converged in restrained refinement ($R_{\rm free}$ < 50%). The mixed model did not converge ($R_{\rm free} \simeq$ 58%). Only about 20% of side-chain rotamers were correctly predicted both in the mixed and in the all-atom models, which is likely to be connected to the high CαRMSD between the model and the final structure. Better convergence in automated refinement of the all-atom model was probably a consequence of the higher percentage of total scattering matter included in the all-atom model (see Table 2).

TM0820 is a putative NADH-dependent butanol dehydrogenase from *T. maritima* (resolution 1.78 Å, space group $P2_1$, two molecules in the asymmetric unit, unit-cell parameters a = 53.69, b = 129.70, c = 55.23 Å, α = 90.00, β = 103.61, γ = 90.00°). The closest structural template was identified by *FFAS*, *PSI-BLAST* and *BLAST* as iron-containing alcohol dehydrogenase TM0920 from *T. maritima* (PDB code 1o2d; 24% sequence identity). The *BLAST* alignment is considerably shorter than those from *PSI-BLAST* and *FFAS*. In this case exhaustive *MOLREP* searches with different models did

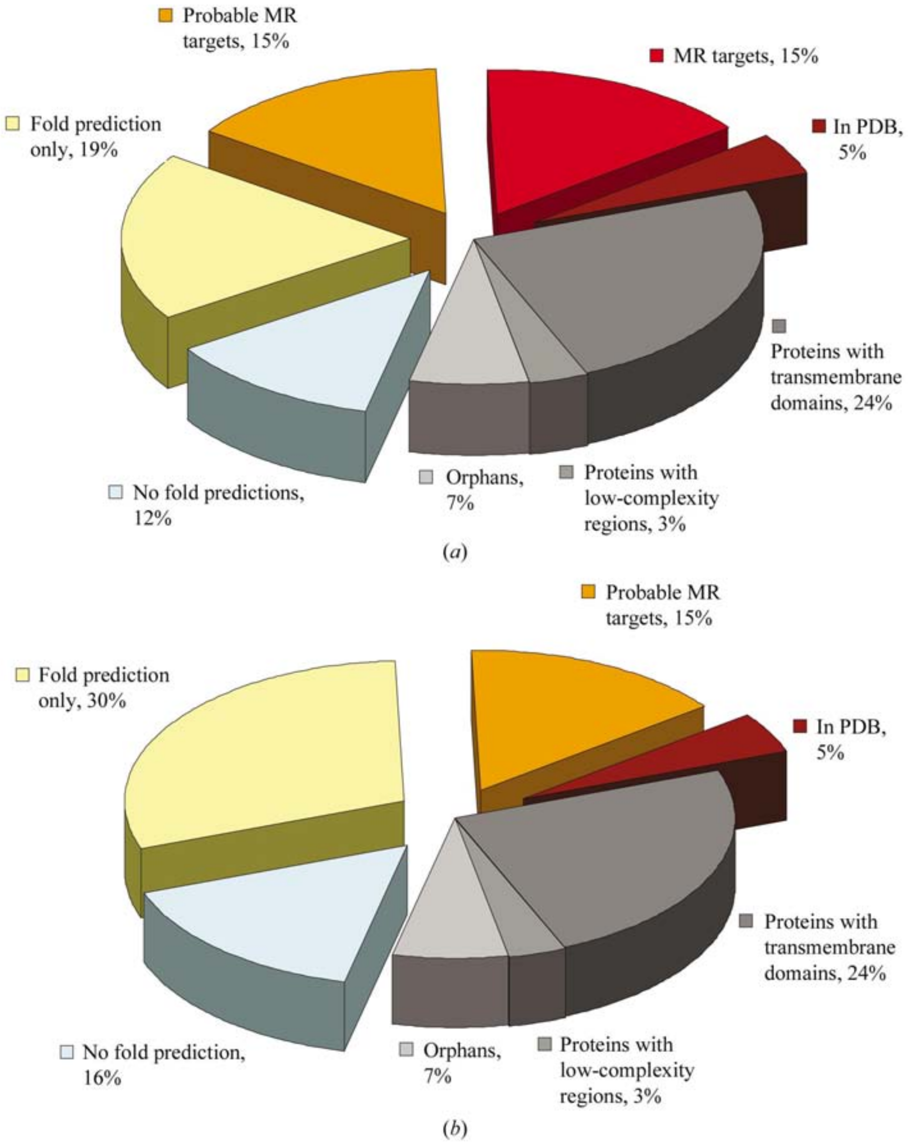

Figure 1
Classification of all *T. maritima* proteins by their suitability for structural prediction and molecular replacement. (*a*) Standard approach. (*b*) Exhaustive MR searches with *FFAS* models. Categories: proteins with transmembrane domains, as predicted with *TMHMM* (Krogh *et al.*, 2001); proteins with low complexity regions, low-complexity regions longer than 40 residues as predicted with *SEG* (Wootton & Federhen, 1993); orphans, targets with no homologs in public sequence databases (as determined with *PSI-BLAST*); no structural prediction, targets without homologs in the PDB; fold prediction only, targets with only remote homologs in the PDB, according to our estimates, not suitable for MR; MR targets, targets with estimated MR success rate higher than 80–90%; probable MR targets, targets suitable for MR in most cases (estimated success rate > 50%); in PDB, targets with structures already deposited in the PDB.

not find the correct solution. TM0820 could only be solved with *EPMR* using mixed models based on the *FFAS* and *PSI-BLAST* alignments. The success of the mixed model is probably owing to the higher accuracy of predicted side-chain rotamers in conserved residues (70% of correct χ_1 values compared with 47% in the all-atom model; see Table 2).

TM0919 is a conserved hypothetical protein from *T. maritima* (resolution 1.8 Å, space group $P2_1$, four molecules in the asymmetric unit, unit-cell parameters a = 64.75, b = 39.79, c = 110.45 Å, α = 90.00, β = 93.79, γ = 90.00°). The closest structural template was identified by *BLAST*, *PSI-BLAST* and *FFAS* as the CRP region of a hypothetical protein from *E. coli* (PDB code 1ml8; 34% sequence identity). Only the *FFAS* alignment covers the full structure with three small one-residue gaps. All MR searches with a monomer model failed. Only a dimer based on the biologically relevant oligomer of CRP constructed from an all-atom model based on the *FFAS* alignment could be correctly placed. The success of the all-atom model is probably owing to higher completeness and because the small backbone difference (CαRMSD of 1.4 Å) between the target and the template allowed a relatively accurate prediction of side-chain rotamers (50% of correct χ_1 values).

TM1459 is a hypothetical protein from *T. maritima* (resolution 1.75 Å, space group $P3_2$, two molecules in the asymmetric unit, unit-cell parameters a = 52.55, b = 52.55, c = 96.27 Å, α = 90.00, β = 90.00, γ = 120.00°). Both *PSI-BLAST* and *FFAS* identified the best structural template as the auxin-binding protein from *Zea mays* (PDB code 1lr5; 18% sequence identity), but the similarity was too low to be detected by *BLAST*. The *FFAS* alignment with 82% coverage is considerably longer than that from *PSI-BLAST* (74% coverage). The correct MR solution could be obtained with an all-atom and a mixed model based on the *FFAS* alignment. Using different types of models based on the *PSI-BLAST* alignment did not give a correct MR solution, which indicates that an ∼10% increase in model completeness was critical for this MR trial.

TM1645 is a nicotinate-nucleotide pyrophosphorylase (NadC) from *T. maritima* (resolution 2.80 Å, space group $I222$, two molecules in the asymmetric unit, unit-cell parameters a = 96.18, b = 126.13, c = 138.20 Å, α = 90.00, β = 90.00, γ = 90.00°). *BLAST*, *PSI-BLAST* and *FFAS* all identified the closest structural template as quinolinate phosphoribosyl transferase from *Mycobacterium tuberculosis* (PDB code 1qpn; 30% sequence identity). The *BLAST* alignment is with 92% coverage, which is significantly shorter than those from *PSI-BLAST* (98% coverage) and *FFAS* (99% coverage). All models based on *BLAST* and *PSI-BLAST* alignments were unsuccessful. Only the all-atom and mixed models based on the *FFAS* alignment yielded a correct MR solution. The biggest difference between the *PSI-BLAST* and *FFAS* alignments was a 13-residue-long shift, which resulted in a wrong residue assignment in the *PSI-BLAST* model.

3.1. Mixed *versus* all-atom models

It is interesting that for two targets (TM0262, TM0919) only all-atom models based on the *FFAS* alignment provided the correct MR solution, while for another two targets (TM0066, TM0820) only the mixed models worked. Thus, the replacement of non-conserved side chains, which involves the prediction of side-chain rotamers, was not always beneficial. Apparently, in the case of TM0820 and TM0066 the prediction of non-conserved side chain rotamers was not accurate enough and thus the all-atom models did not yield an MR solution (see Table 2). At the same time, rotamers of conserved residues included in the mixed model were in significantly better agreement with the real structure, which resulted in a successful MR solution. The failure of the mixed model and success of the all-atom model in the cases of TM0262 and TM0919 corresponds to a situation where there is no significant difference in the accuracy of side-chain rotamers between conserved and non-conserved residues. In this case, the mixed model did not improve the percentage of correctly predicted atoms and all-atom models may have worked better just because they had a higher completeness and total number of correctly placed side chains. In any case, the most important conclusion is that it is definitely beneficial to use both types of models.

4. Discussion

As expected, MR was relatively straightforward when sequences of the target and the template were more than 35% identical. All but one case in this range could be solved using the standard approach. While correct MR solutions could be obtained for all different types of models, it is noteworthy that in our case all-atom models usually converged to lower values of $R_{\rm free}$ in subsequent automated refinement and rebuilding steps (see Table 2).

The situation changes dramatically with lower sequence identity (<35%), where standard sequence–sequence alignment methods start to be inaccurate and where the values of CαRMSD between target and template tend to be higher (Chothia & Lesk, 1986). This is reflected by the fact that models based on *BLAST* alignments had a significantly lower success rate in MR. Reasons for this are that *BLAST* alignments were significantly shorter and less accurate than those from *PSI-BLAST* and *FFAS*. Furthermore, in two cases *BLAST* could not detect a homolog at all. The differences between alignments from *PSI-BLAST* and *FFAS* are smaller but significant in that *FFAS* alignments often show better coverage and accuracy than those from *PSI-BLAST*. The models based on *FFAS* profile–profile alignments were successful in four cases where those based on *PSI-BLAST* failed. It shows that these subtle errors in the alignment may cause a model to fail in MR or in subsequent automated refinement steps if the overall similarity between template and the model is low. Our results also show that four out of 20 structures below 35% sequence identity still remain unsolved despite extensive modeling and MR trials. This failure is especially intriguing for targets TM1244 and TM0816, which seem to fail for no apparent reason (other than suspected problems with the data set) because their alignments show good coverage with a very low number of gaps. The possible reason of the failure of these targets is the relatively low ratio of observations per atom (see Table 1, column o/a).

Other potential parameters influencing the success of MR are the percentage of scattering matter represented by the search model (Table 2, column %sm) and the resolution of the search model (Table 1, column TR). However, in our set of 31 structures we did not observe any significant correlation between these parameters and the success of MR.

An analysis of all structures deposited in the PDB during the year 2002 shows that out of 1589 unique structures, 769 were reportedly solved by MR. *FFAS* detects 180 additional structures which could have, according to our results, potentially been solved by MR. Thus, using our approach and taking the current status of the PDB into account, one could solve about 10–12% more structures with MR and save a considerable amount of experimental efforts and costs. This improvement is even more pronounced when looking at a bacterial genome such as that from *T. maritima*, where the estimated number of potential MR targets increases from 15% to 30% (see Fig. 1).

Although the results for these 31 data sets do not allow a thorough statistical analysis of MR feasibility, we can sketch the following strategy for molecular replacement below a sequence identity of 35%.

(i) One should use advanced alignment methods to assure the highest accuracy of the alignment. As soon as significant sequence similarity to a protein from the PDB can be detected (*FFAS* score < −15; sequence identity > 15%; coverage > 60%), the protein can be treated as a potential MR target.

(ii) PDB files of the top-scoring homologs should be obtained, including their biologically relevant oligomers, if applicable. Mixed models should be used and if they fail all-atom models should be tried.

(iii) At least two MR programs using different approaches, for example *MOLREP* (rotation and translation search) and *EPMR* (full six-dimensional search), should be tried with different sets of parameters for each type of model. For *MOLREP*, *similarity* and *completeness* are appropriate parameters to be exhaustively explored. In the case of *EPMR*, high- and low-resolution limits play a similar role. The only practical solution for massive MR searches with different parameters is automation and parallelization.

In conclusion, we show that using simple models based on more accurate alignments increases the success rate of MR in cases where the unknown structure and the search model share less than 35% sequence identity. Therefore, a strategy that combines such models with exhaustive MR searches can save a considerable amount of time and resources, especially for structural genomics projects.

The JCSG is supported by the National Institutes of Health Protein Structure Initiative Grant GM62411 from the National Institute of General Medical Sciences.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, B. E., Michels, P. A. & Hol, W. G. (1997). *Nature (London)*, **385**, 275–278.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P, Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Chen, Y. W. (2001). *Acta Cryst.* D**57**, 1457–1461.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **4**, 823–826.

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

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Holm, L. & Sander, C. (1995). *Trends Biochem. Sci.* **20**, 478–480.

Hoppe, W. (1957). *Acta Cryst.* **10**, 750–751.

Jamrog, D. C., Zhang, Y. & Phillips, G. N. Jr (2003). *Acta Cryst.* D**59**, 304–314.

Jaroszewski, L., Rychlewski, L. & Godzik, A. (2000). *Protein Sci.* **9**, 1487–1496.

Jones, D. (2001). *Acta Cryst.* D**57**, 1428–1434.

Jones, T. A., Zou, J.-Y, Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Kleywegt, G. J. (1996). *CCP4 Newsl.* **32**, http://www.ccp4.ac.uk/newsletter/uppsala.html.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. (2001). *J. Mol. Biol.* **305**, 567–580.

Li, W., Jaroszewski, L. & Godzik, A. (2002). *Protein Eng.* **15**, 643–649.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.

Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, V. S. (2001). *Acta Cryst.* D**57**, 1445–1450.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Rossmann, M. G. (2001). *Acta Cryst.* D**57**, 1360–1366.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). *Protein Sci.* **9**, 232–241.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. & Karplus, M. (1995). *Proteins*, **23**, 318–326.

Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.

Vriend, G. J. (1990). *J. Mol. Graph.* **8**, 52–56.

Wootton, J. C. & Federhen, S. (1993). *Comput. Chem.* **17**, 149–163.