Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The JCSG MR pipeline: optimized alignments, multiple models and parallel searches

Robert Schwarzenbacher,[a] Adam Godzik[b] and Lukasz Jaroszewski[b]*

[a]University of Salzburg, Structural Biology, Billrothstrasse 11, 5020 Salzburg, Austria, and [b]Joint Center for Structural Genomics, Bioinformatics Core, The Burnham Institute for Medical Research, 10901 North Torrey Pines Road, La Jolla, CA 92093, USA

Correspondence e-mail: lukasz@burnham.org

Received 26 February 2007
Accepted 12 October 2007

The success rate of molecular replacement (MR) falls considerably when search models share less than 35% sequence identity with their templates, but can be improved significantly by using fold-recognition methods combined with exhaustive MR searches. Models based on alignments calculated with fold-recognition algorithms are more accurate than models based on conventional alignment methods such as *FASTA* or *BLAST*, which are still widely used for MR. In addition, by designing MR pipelines that integrate phasing and automated refinement and allow parallel processing of such calculations, one can effectively increase the success rate of MR. Here, updated results from the JCSG MR pipeline are presented, which to date has solved 33 MR structures with less than 35% sequence identity to the closest homologue of known structure. By using difficult MR problems as examples, it is demonstrated that successful MR phasing is possible even in cases where the similarity between the model and the template can only be detected with fold-recognition algorithms. In the first step, several search models are built based on all homologues found in the PDB by fold-recognition algorithms. The models resulting from this process are used in parallel MR searches with different combinations of input parameters of the MR phasing algorithm. The putative solutions are subjected to rigid-body and restrained crystallographic refinement and ranked based on the final values of free *R* factor, figure of merit and deviations from ideal geometry. Finally, crystal packing and electron-density maps are checked to identify the correct solution. If this procedure does not yield a solution with interpretable electron-density maps, then even more alternative models are prepared. The structurally variable regions of a protein family are identified based on alignments of sequences and known structures from that family and appropriate trimmings of the models are proposed. All combinations of these trimmings are applied to the search models and the resulting set of models is used in the MR pipeline. It is estimated that with the improvements in model building and exhaustive parallel searches with existing phasing algorithms, MR can be successful for more than 50% of recognizable homologues of known structures below the threshold of 35% sequence identity. This implies that about one-third of the proteins in a typical bacterial proteome are potential MR targets.

1. Introduction

Molecular replacement (MR; Rossmann, 2001) has an advantage over experimental phasing techniques because it requires only one data set of reflections obtained from a native protein crystal, which is considerably less resource-intensive

than multiple-wavelength experiments with substituted protein crystals.

Because of advances in structural biology, more and more structures are available through the Protein Data Bank (PDB; Berman *et al.*, 2000). As the number of known protein structures grows rapidly, the main interest shifts from studying individual structures to studying protein complexes, which are fundamental to our understanding of protein interactions in biological mechanisms such as metabolism, the cell cycle or apoptosis. MR is the method of choice for solving the structures of protein complexes because the structures of individual proteins are often known. As a result, the number of protein structures determined by MR increases every year, so any improvements in the method can save considerable time and resources.

The MR phasing algorithms 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 *Phaser* (Storoni *et al.*, 2004), *AMoRe* (Navaza, 2001), *X-PLOR*/*CNS* (Brünger *et al.*, 1998), *MOLREP* (Vagin & Teplyakov, 2000), *EPMR* (Kissinger *et al.*, 1999) and *Queen of Spades* (Glykos & Kokkinidis, 2000). The success of these MR methods depends critically on the quality of the model used and different ways of preparing models are still being explored. MR has been accomplished with models that cover 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 required and the differences between the coordinates of the model and the molecule must be small [usually with a root-mean-square distance of C^{α} atoms (CαRMSD) below 2.5 Å]. The requirements for optimal search models for MR are still being explored. Several interesting ideas regarding search models have been proposed or tested on individual cases or on small sets of structures (Kleywegt, 1998). These ideas include removing or cutting back residues or regions with high temperature factors, the omission of regions where sequence conservation is low, using composite search models (Chen, 2001) and building alternative models based on suboptimal alignments (Jones, 2001). Recently, the analysis of several difficult MR problems from our center has demonstrated that the alignment accuracy and side-chain modeling have a significant impact on MR success rates (Schwarzenbacher *et al.*, 2004). Some of the methods of model preparation have been implemented in the *CHAINSAW* program, written by Norman Stein and included in the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). *CHAINSAW* prepares different variants of pruned (mixed) search models for MR.

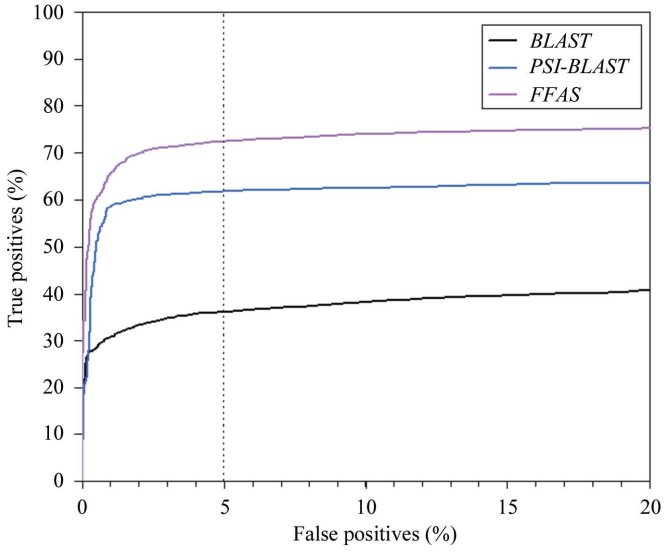


Figure 1
The percentages of correct and incorrect structural predictions derivable by *BLAST*, *PSI-BLAST* and *FFAS* for the representative benchmark set of homologous protein pairs with less than 25% sequence identity based on the SCOP database. With 5% of false positives, *BLAST* correctly detects 35% of such pairs and *PSI-BLAST* finds 60%, while *FFAS* can predict up to 72%.

The most effective methods of protein structure prediction are based on establishing a homology between a protein of interest and an already characterized protein. However, the standard sequence-comparison methods rapidly lose sensitivity in the 'twilight zone' where there is below 30% sequence identity between the protein of interest and the closest known structure (Holm *et al.*, 1992). The sensitivity of fold recognition can be improved by using evolutionary information, which can be extracted from large families of protein sequences. Instead of comparing two sequences, one compares a protein sequence with sequences from an entire protein family represented by a sequence profile as implemented in *PSI-BLAST* (Altschul *et al.*, 1997) or by hidden Markov model (HMM; Eddy, 1998). A logical next step in this strategy is to compare two sequence profiles as introduced in *FFAS* (Rychlewski *et al.*, 2000) or two hidden Markov models as implemented in *HHSEARCH* (Soding, 2005).

The application of sequence profiles has a significant impact on the number of fold predictions one can make from a given set of known structures. A widely accepted way of testing homology-prediction methods is to apply them to representative sets of known structures and to calculate the number of correct predictions and false positives for different score thresholds corresponding to different error levels. Using this procedure, we re-evaluated the sensitivity of remote homology detection using three different methods. We used the *ASTRAL* resource (Chandonia *et al.*, 2004) based on the SCOP database (Murzin *et al.*, 1995) to construct a benchmark set of 5868 protein domain structures with less than 25% sequence identity to each other. The predictions obtained with *BLAST*, *PSI-BLAST* and *FFAS* for this benchmark clearly illustrate the advantage of using sequence profiles for the detection of distant homologues (see Fig. 1). At the 5% error level the profile–sequence comparison method *PSI-BLAST* (Altschul *et al.*, 1997) gives almost twice as many correct predictions as the sequence–sequence comparison algorithm *BLAST* (Altschul *et al.*, 1990). The profile–profile comparison method *FFAS* improves the sensitivity by another 20%.

Other advanced fold-recognition methods based on sequences profiles or similar methods of using evolutionary information include 3*D-PSSM* (Kelley *et al.*, 2000), *FUGUE* (Shi *et al.*, 2001), *BIOINBGU* (Fischer, 2000), *PROSPECT* (Xu & Xu, 2000) and *SAMT*98 (Karplus *et al.*, 1998). These methods are more sensitive than sequence–sequence alignment methods such as *BLAST* and are usually more sensitive than profile–sequence alignment methods such as *PSI-BLAST*.

Besides the accuracy of the model, for more difficult MR problems, the success may critically depend on certain settings of the phasing algorithm, such as the low- and high-resolution limit applied to the crystallographic data. The strong dependence on the resolution limit and cutoff is especially evident for MR phasing algorithms, which are not based on the maximum-likelihood principle. It is rather difficult to propose any useful rules of thumb for selecting optimal low- and high-resolution cutoffs and, as suggested by the authors of MR programs, it is beneficial to test several combinations of these cutoffs. Usually, in difficult MR cases multiple phasing trials with different models and input parameters are performed manually, which imposes practical limits on the number of tested combinations.

We demonstrated that it is possible to extend the limits of the MR method by using several specifically designed protein models based on profile–profile fold recognition and exhaustive MR searches in a parallelized and automated MR pipeline (Schwarzenbacher *et al.*, 2004) built at the Joint Center for Structural Genomics (Lesley *et al.*, 2002).

At least three other groups are also involved in the development of advanced and publicly available MR pipelines, including *CaspR* (Claude *et al.*, 2004), *MrBUMP* (Keegan & Winn, 2008) and *BALBES* (Long *et al.*, 2008). Interesting attempts have also been made to go beyond the 'rigid search model' and generate search models using normal-mode analysis (Suhre & Sanejouand, 2004; Jeong *et al.* 2006).

In this manuscript, we provide a short description of the JCSG MR pipeline, discuss the advantages of using sensitive fold-recognition algorithms and show the benefits of applying parameter-space screening to MR searches. We also give an update on the statistics of the results of the pipeline and further explore methods of generating alternative models for MR.

2. Methods and results

2.1. The JCSG MR pipeline and its results

The parallelized MR pipeline used in the JCSG automatically performs all steps from homology detection through model preparation and MR searches to automated refinement. The pipeline includes the following steps (see Fig. 2).

(i) Firstly, a homology search is carried out in the PDB with the *FFAS* profile–profile fold-recognition method to assure optimal sensitivity in finding homologous templates and the highest accuracy of the alignment. As soon as significant sequence similarity to a known structure can be detected with *FFAS*, the protein is treated as a potential MR target [the sequence identity should exceed 15% and the *FFAS* score should be better (lower) than −15]. In most cases, we also required that at least two-thirds of the structure is included in the search model. However, MR may be feasible with smaller models of high accuracy. For example, individual protein domains with determined structures may be used for the phasing of full multi-domain proteins. The pipeline can be used to attempt MR phasing in such cases.

(ii) PDB files of the top-scoring homologues are obtained, including their biologically relevant oligomers, if available.

(iii) A pool of different types of models is built using the program *WHATIF* (Vriend, 1990): all-atom models with side chains replaced according to the alignment and side-chain conformations optimized, 'mixed' models with side-chain conformations of conserved residues transferred from the template and with the other residues replaced with serine (Schwarzenbacher *et al.*, 2004) and all-atom and 'mixed' models of possible oligomers based on the physiologically relevant oligomers of the templates.

(iv) MR searches are performed with the program *MOLREP*. Exhaustive parameter-space screening is applied to the similarity (SIM) and completeness (COMPL) parameters of *MOLREP*, with other parameters set to default values. For both parameters values of 0.1, 0.3, 0.5, 0.7 and 1.0 are tested, yielding a total of 25 parameter combinations. We found out that finer searches with 100 combinations did not provide solutions which could not be achieved with 25 combinations. In some cases, however, we performed finer grid searches for illustration purposes (see Fig. 3).

(v) All solutions are subjected to rigid-body refinement and restrained

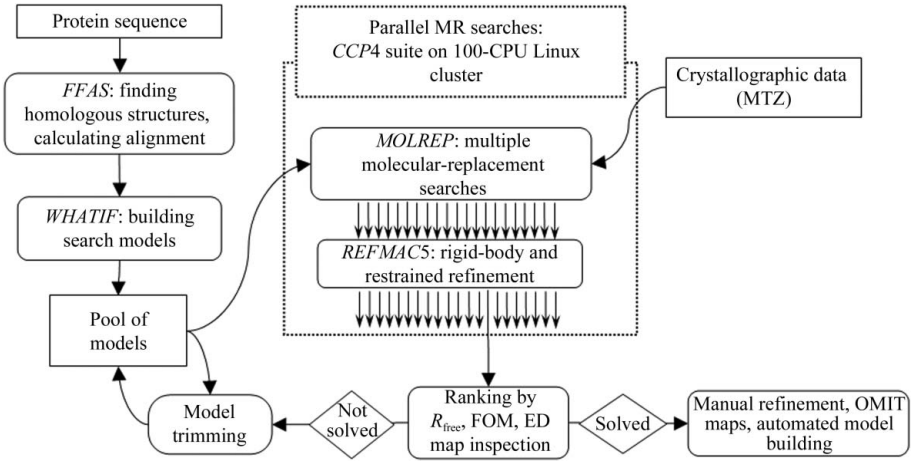


Figure 2
Schema of the JCSG MR pipeline.

refinement with *REFMAC*5 (Murshudov *et al.*, 1997) and the solution with the lowest R_{free} value is selected. In most cases, we performed 5–20 steps of rigid-body refinement and 100–500 steps of restrained refinement. The *REFMAC*5 WEIG parameter controlling the weighting of the X-ray and geometric parts was set to 0.05 and in the most difficult cases additional values in the range 0.02–0.05 were tested.

(vi) If the structure cannot be phased using the procedure described above, large sets of trimmed models may be generated. As suggested by Kleywegt (1998), trimming includes loop regions, regions corresponding to gaps and regions of low sequence conservation in the alignment. The models with all possible combinations of such trimmings are tested in MR searches as described in (iv) and (v) above. The combinatorial trimming step is optional and is not yet fully automated.

(vii) Electron-density maps are examined and solved structures are completely refined and deposited in the PDB.

The MR pipeline provided solutions for 33 protein structures with less than 35% sequence identity to their modeling templates (column P in Table 1). These results were compared with results from 'simple' MR runs (column S in Table 1) in which one model based on a *BLAST* alignment was used in an MR search with default parameters. The same model was also used in exhaustive MR searches (column E in Table 1) with a wide range of parameters. By using different types of models based on accurate alignments combined with parallel processing, we can practically double the number of protein structures which can be solved by MR. Our results indicate that MR is usually straightforward if models share more than 30–35% identical residues with their templates (Schwarzenbacher *et al.*, 2004), which is in good agreement with the widely accepted limit of highly accurate homology modeling (Vogt *et al.*, 1995). Almost all MR cases with more than 35% sequence identity between the model and the structure were solved with the 'simple approach' and unsolved problems are most likely to indicate problems with the crystallographic data rather than with model accuracy. Below 35% sequence identity the 'simple approach' was ineffective and successful in only ten out of 33 cases (column S in Table 1). Exhaustive MR searches with standard templates resulted in six additional MR solutions (column E, Table 1). Exhaustive MR searches with different types of models including biologically relevant oligomers, mixed and all-atom homology models based on *FFAS* alignments (column P, Table 1) solved 17 additional structures with less than 35% sequence identity to their templates. Despite exhaustive searches with multiple models, 14 structures with less than 35% sequence identity remained unsolved.

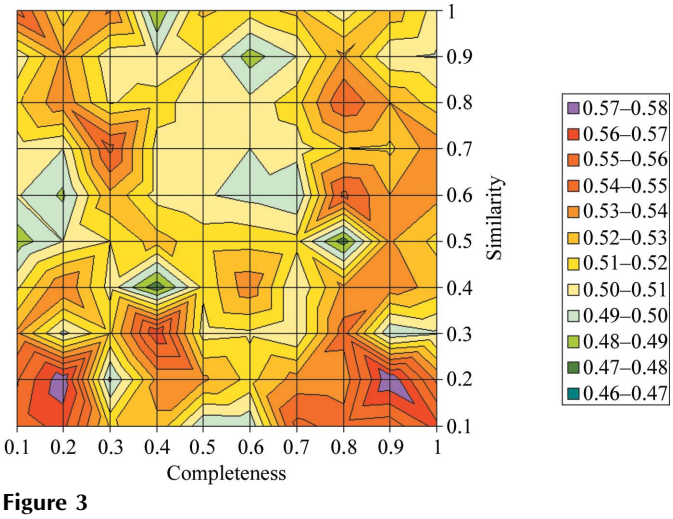


Figure 3
The results of parameter screening applied to MR phasing and automated refinement of JCSG target TM0332. All combinations of similarity (SIM) and completeness (COMPL) parameters of the *MOLREP* program were tested by an exhaustive grid search between 0.1 and 1.0 at intervals of 0.1. All resulting solutions were subject to 20 steps of rigid-body refinement and 500 steps of restrained refinement. The final R_{free} value after restrained refinement is plotted as a contour map.

2.2. Parameter-space screening in MR searches

The procedure of exhaustive testing of different input parameters of crystallographic software has been called parameter-space screening (Liu *et al.*, 2005). In order to complete calculations in a reasonable time, parameter-space screening is usually performed in a parallel way using computer clusters. The results of MR phasing algorithms often depend on several input parameters connected to filters applied to the data and to the anticipated accuracy of the search model. In our pipeline, we relied on the program *MOLREP* (Vagin & Teplyakov, 2000) from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994) because of its robustness, speed and simple usage. Two of the input parameters of the program are related to the expected completeness of the search model and its expected similarity to the structure being solved. The completeness parameter (COMPL) is linked to the soft low-resolution cutoff applied to the crystallographic data and the similarity parameter (SIM) is linked to the high-resolution cutoff. Since we do not have exact information about the accuracy of the model before the actual structure is solved, different combinations of these two parameters are exhaustively tested, as suggested by the authors of the program. In particular, low-resolution reflections and the low-resolution cutoff are known to play important roles in MR phasing. However, instead of examining the low-resolution part of the data and trying to find the optimal low-resolution cutoff, we applied different low-resolution cutoffs by changing the COMPL parameter and tested the correctness of all solutions by refining them. In fact, our tests indicated that in several cases the success of phasing with *MOLREP* was dependent on these input parameters in an unpredictable way, which underscores the importance of exhaustive parameter-space screening. For example, parameter-space screening was used for MR phasing of orotidine 5′-phosphate decarboxylase (TM0332) from *Thermotoga maritima*. *FFAS* detected similarity to the structure of orotidine 5′-phosphate decarboxylase from *Escherichia coli* (PDB code 1eix) with a score of −60, a sequence identity of 24% and the alignment covering 98% of the sequence with six gaps. Fig. 3 shows a contour map of final R_{free} values after restrained refinement obtained for MR solutions calculated with

Table 1
JCSG MR projects for structures with less than 35% sequence identity to the template.

Target, TIGR or GeneBank ID and the name 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; T, the closest homologue with known structure (PDB code); Id, sequence identity between target and template; S, results of a single MR search with a simple template; E, results of exhaustive MR searches with a simple template; P, results of MR pipeline (different types of models based on *FFAS* alignments plus exhaustive MR search); X, successful MR phasing and automated refinement; PDB, PDB code of the solved MR structures (if already deposited in PDB).

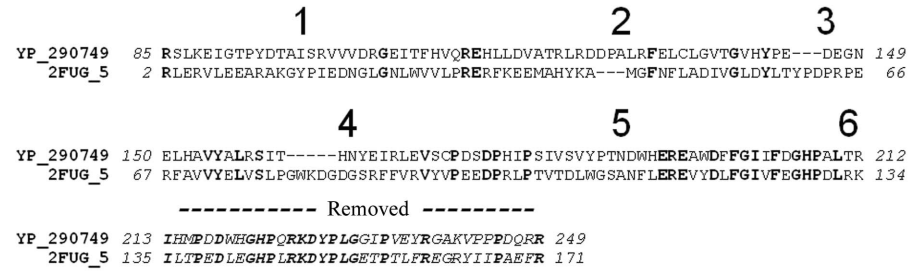
Target	L	SG	M	R	o/a	T	Id	S	E	P	PDB
17134165, hypothetical protein, *Nostoc* sp.	165	$P2_12_12$	2	1.50	18.7	1g76	14			X	1vl7
tm1459, carbohydrate-binding protein, *T. maritima*	114	$P32$	2	1.75	11.8	1lr5	18			X	1o5u
tm1287, oxalate decarboxylase, *T. maritima*	121	$C2$	2	1.70	8.9	1vj2	18		X	X	1o4t
15079298, glia maturation factor-γ, *Mus musculus*	142	$P1$	1	1.35	15.7	1ahq	19	X	X	X	1vkk
tm0603, 30s ribosomal protein s6, *T. maritima*	128	$P4_12_12$	1	1.70	15.0	1lou	19			X	1vmb
17391249, haloacid dehalogenase-like hydrolase, *M. musculus*	248	$P6_122$	1	1.90	12.0	1x42	19			X	2gfh
tm1394, heat-shock protein 33, *T. maritima*	290	$P2_12_12_1$	2	2.00	8.6	1i7f	20			X	1vq0
18044849, bifunctional coenzyme A synthase, *M. musculus*	269	$C2$	1	1.70	15.0	1n3b	22			X	2f6r
tm0820, NADH-dependent butanol dehydrogenase, *T. maritima*	395	$P2_1$	2	1.78	10.0	1o2d	24			X	1vlj
tm0332, orotidine 5′-phosphate decarboxylase, *T. maritima*	201	$C2$	1	1.90	9.2	1eix	24			X	1vqt
10175646, BH3024 protein, *Bacillus halodurans*	126	$P4_12_12$	1	2.40	6.5	1kgs	25	X	X	X	2b4a
NP_394403, GMP synthase, *T. acidophilum*	212	$P2_12_12$	4	2.45	4.4	1gdl	25			X	2a9v
tm0262, DNA polymerase III, β subunit, *T. maritima*	366	$P4_22_12$	1	2.70	4.8	1jqj	26			X	1vpk
tm1419, *myo*-inositol-1-phosphate synthase, *T. maritima*	382	$I222$	1	1.58	22.5	1gr0	26		X	X	1vjp
YP_290749.1, NADH dehydrogenase subunit C, *T. fusca* YX	252	$P4_32_12$	1	2.60	8.6	2fug	27			X	
tm1088A, hypothetical protein, *T. maritima*	143	$P2$	1	1.50	20.3	1lss	27	X	X	X	2g1u
tm0748, SAM-dependent *O*-methyltransferase, *T. maritima*	265	$I222$	1	1.70	16.7	1i9g	28	X	X	X	1o54
tm0544, ABC transporter ATP-binding protein, *T. maritima*	244	$P3_12_1$	1	2.10	10.6	1ji0	29			X	1vpl
tm1128, ferritin, *T. maritima*	182	$H32$	8	2.35	8.1	1eum	30	X	X	X	1vlg
tm0295, transaldolase, *T. maritima*	218	$P2_1$	20	2.40	5.1	1l6w	30			X	1vpx
tm0343, DAHP synthase, *T. maritima*	338	$P2_12_12_1$	3	1.90	8.5	1fwn	31	X	X	X	1vr6
tm1385, glucose-6-phosphate isomerase, *T. maritima*	448	$I2_12_12_1$	3	2.90	6.8	1b0z	31		X	X	
tm1645, quinolinate phosphoribosyltransferase, *T. maritima*	273	$I222$	2	2.80	6.9	1qpn	31			X	1o4u
tm0066, 2-dehydro-3-deoxyphosphogluconate aldolase, *T. maritima*	205	$C222_1$	3	2.30	6.8	1eua	31		X	X	1vlw
tm1393, MEP cytidylyltransferase, *T. maritima*	222	$P6_1$	2	2.60	6.7	1vgz	31			X	1vpa
tm1244, phosphoribosylformylglycinamidine synthase, *T. maritima*	82	$I4_122$	4	2.50	7.0	1t4a	32			X	1vq3
tm0166, dihydrofolate synthase, *T. maritima*	430	$P6_122$	1	2.75	8.9	1fgs	32	X	X	X	1o5z
tm0919, hydroperoxide-resistance protein OsmC, *T. maritima*	138	$P2_1$	4	1.80	12.9	1ml8	33			X	1vla
tm1698, aspartate aminotransferase, *T. maritima*	397	$P2_1$	6	2.50	4.1	1xi9	29	X	X	X	2gb3
tm0604, single-stranded DNA-binding protein, *T. maritima*	141	$F222$	1	2.40	10.0	1qvc	34		X	X	1z9f
tm1169, 3-oxoacyl-(acyl carrier protein) reductase, *T. maritima*	237	$P2_12_12_1$	4	2.50	4.3	1i01	34		X	X	1o5i
17130499, anthranilate phosphoribosyltransferase 2, *Nostoc* sp.	345	$P2_1$	2	2.50	4.8	1kgz	35	X	X	X	1vqu
tm0159, xanthosine triphosphate pyrophosphatase, *T. maritima*	191	$P4_12_12$	2	1.78	18.3	1v7r	35	X	X	X	1vp2

different values of the similarity and completeness parameters. The MR solutions obtained for different input parameters of the program *MOLREP* led to final $R_{\rm free}$ values from *REFMAC*5 ranging from 0.464 to 0.546. The solution with the lowest $R_{\rm free}$ value was manually refined and deposited in the PDB (PDB code 1vqt). The CαRMSD between fully refined TM0332 structure and 1eix is 2.27 Å. A detailed analysis of the solutions with different final $R_{\rm free}$ values showed that most of the solutions with $R_{\rm free}$ values higher than 0.5 were incorrect, underscoring the significance of parameter-space screening for this case.

2.3. Combinatorial trimming of search models

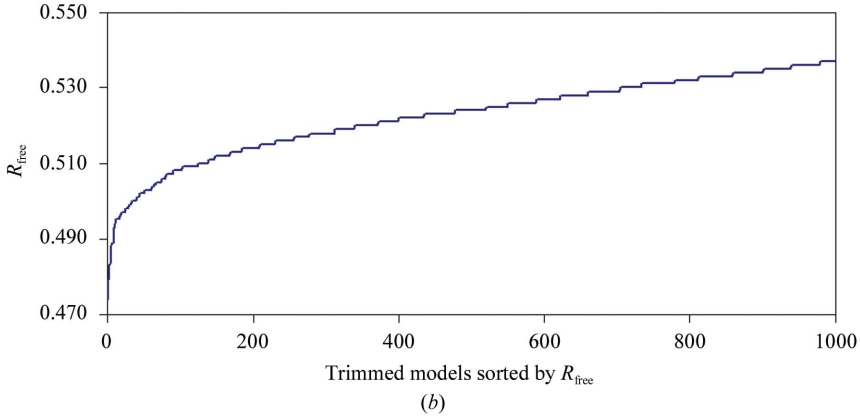
For difficult cases in which the application of exhaustive parameter-space screening combined with multiple models based on different templates does not yield a solution, it is possible to increase the variability of the models used in the pipeline by using models with different combinations of trimmings of possibly unreliable regions.

It is widely accepted that an optimal model for MR phasing should contain all atoms that can be predicted with sufficient accuracy and should not contain any atoms with high co-ordinate errors. Unreliable regions of the model usually include loops, gaps and fragments of low sequence similarity between the model and the template. Such regions are more likely to contain significant errors. Therefore, by removing such regions from the model one can significantly increase its overall accuracy, but some accurately predicted regions can also be removed, since the exact locations of inaccurate regions are not known before the structure is solved. The level of accuracy required for MR models is also not obvious and may vary for different data sets. A brute-force solution to this problem is to use the capabilities of a parallelized MR pipeline and test all combinations of possible trimmings of the model. This procedure allowed MR phasing of the structure of



Region	Tested trimmings
1	None, 85–97, 85–105
2	None, 123–126, 121–126, 130–134, 121–134
3	None, 144–149
4	None, 160–163, 159–164
5	None, 183–191
6	None, 208–212, 203–212

(*a*)



(*b*)

Figure 4
(*a*) The alignment used for modeling of target YP_290749.1 based on PDB structure 2fug. The regions of lower alignment reliability are labeled on the alignment and on the model. The table shows the trimmings applied in these regions. (*b*) Final R_{free} values from restrained refinement obtained for trimmed models tested in the pipeline. All 2000 results were ranked by their final R_{free}. Sorted R_{free} values for the 1000 best ranking models are shown as a graph.

NADH dehydrogenase subunit C from *Thermobifida fusca* (GenBank accession code YP_290749). According to *FFAS*, the only structure homologous to this protein is subunit 5 of an oligomeric domain in respiratory complex I from *Thermus thermophilus* (PDB code 2fug). *FFAS* aligned 66% of the sequence of YP_290749 with the sequence of 2fug, with a score of −79 and a sequence identity of 27%. Residues 213–249 of the target sequence were aligned with the region of 2fug subunit 5 which extends from its globular domain and binds to another subunit in the complex. However, since the present crystals only contained the isolated domain, we expected that this particular region may have a different conformation and removed it from the model. This resulted in a decrease in the sequence identity to 22% and in the sequence coverage by the model to 50% (see Fig. 4*a*). Since the asymmetric unit of 2fug contains four slightly different copies of subunit 5 (chains 5, E, N, W), each of them was used to build models of the target. Model trimmings were proposed based on the sequence alignment, in which six potentially unreliable regions of the model were identified. We applied up to four alternative trimmings in each of these regions (see Fig. 4*a*). By applying all combinations of these trimmings, we produced 540 trimmed models from each copy of subunit 5, yielding a total of 2160 models. All search models were submitted to the MR pipeline. MR searches were completed in about 5 h on a 50 CPU Linux cluster. Because of time limitations, parameter-space screening was not used and MR solutions obtained with default *MOLREP* parameters went directly to 30 cycles of restrained refinement in *REFMAC*5.

Interestingly, only a small subset of trimmed models led to successful phasing as indicated by significantly lower R_{free} values from *REFMAC*5 (see Fig. 4*b*).

3. Discussion

The JCSG MR pipeline increases the success rate of MR by using accurate modeling methods, large numbers of alternative models and applying parameter-space screening to phasing algorithms. We observed that MR was relatively straightforward when the sequences of the target and the template were more than 35% identical. Based on our results, we tend to accept 35% as a limit of straightforward MR, since almost all cases in this range could be solved using the standard approach.

This situation changes when the sequence identity drops below 35%: standard alignment methods start to be less accurate and CαRMSD values between structures of related proteins increase significantly (Chothia & Lesk, 1986). Although the relationship between the sequence identity of pairs of protein structures and their CαRMSD values is well established, the character of this relationship varies significantly among protein families, as it becomes apparent when structural alignments of large families are calculated and analyzed (Reeves *et al.*, 2006). Therefore, one can expect that the limit of accurate homology modeling (which is also the limit of feasible MR) may be different for different protein families. In some cases, the chances of successful MR phasing can be estimated based on the structural variability observed among known structures from a protein family of interest. If known structures from a family show only small differences in the protein core, then unknown structures from this family are also likely to have a well

conserved core. Members of such protein families could be suitable for MR, even when the sequence identity to the closest known structure is very low. Therefore, as an element of experiment design one may perform homology searches in the PDB database using sensitive fold-recognition methods such as the *FFAS* server (Jaroszewski *et al.*, 2005; available at http://ffas.burnham.org). Then, if homologous structures are found one can assess the structural similarity between them using a multiple structural alignment method such as *POSA* (Ye & Godzik, 2005; available at http://fatcat.burnham.org/POSA). The *POSA* server provides a quantitative measure of the structural similarities between submitted structures along with a graphical interface, which we found very helpful in determining the extent of the conserved structural core in the family. At this point it is rather difficult to provide general quantitative limits of the applicability of MR based on such analyses, but in many cases it is possible to tell whether MR phasing is worth considering.

Below 35% sequence identity models based on *BLAST* alignments had a lower success rate, since in most cases they are shorter and less accurate than the alignments from *PSI-BLAST* and *FFAS*. Furthermore, in two cases (targets 17134165 and TM0603) *BLAST* could not detect a homologous structure at all, while remote similarity detected using *FFAS* led to successful MR phasing. This observation implies that some difficult MR problems can be solved by using publicly available fold-recognition servers.

Because of its high computational cost, the method of combinatorial model trimming was only applied to a few unsolved MR problems. The example of the phasing of NADH dehydrogenase subunit C using this method is interesting because the distribution of $R_{\rm free}$ values for trimmed models has a very narrow minimum. It is impossible to make general conclusions based on one example, but this observation suggests that the results of MR and refinement are highly susceptible to the ratio of correctly and incorrectly predicted atoms in the search model. This implies that combinatorial trimming, which allows maximization of this ratio in some models, may provide solutions to problems that are beyond the reach of models based on one optimal alignment. It has to be noted that the method of combinatorial trimming is currently only partly automated and requires manual intervention. For example, the model regions to be trimmed were proposed based on visual inspection of the alignment. In principle, one can imagine full automation of such a procedure by using known methods of assessing the local accuracy of the model. The method needs to be tested on more examples before it can be fully automated.

The results obtained for 47 data sets still do not allow a thorough statistical analysis of the feasibility of MR, which depends on too many features of the data and the model. Nevertheless, we can roughly estimate that the success rate is about 50% for proteins with an *FFAS* score better (lower) than −15, a sequence identity in the range 15–35% and a model which covers at least two-thirds of the sequence.

The main conclusion of our tests is that search models based on alignments from sensitive fold-recognition algorithms together with the latest MR phasing techniques in combination with parameter-space screening do improve the success rate of MR phasing. This improvement will be critical for solving protein complexes and may save a considerable amount of time and resources, especially for structural genomics projects.

It has to be noted that the procedures described above are very CPU demanding and in most cases impractical without a computer cluster. At JCSG we used 25–50 CPUs of a Linux cluster for most calculations. Completion of most searches still took several hours.

The *FFAS* program is available as a web server at http://ffas.burnham.org and is linked to a modeling server which can produce all-atom and mixed models based on *FFAS* alignments. The authors are preparing a distribution version of the JCSG MR pipeline scripts and it will be made available to the academic community on request.

The results presented in this publication were possible thanks to the effort of the entire JCSG team. The authors are especially grateful to their colleagues from the JCSG Structure Determination Core at Stanford Synchrotron Radiation Laboratory, who obtained all data sets used in this work and helped with their crystallographic expertise. The JCSG is supported by the NIH Protein Structure Initiative grant U54 GM074898 from the National Institute of General Medical Sciences (http://www.nigms.nih.gov). RS is supported by EC grant MEXT-CT-2006-033534.

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, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2004). *Nucleic Acids Res.* **32**, D189–D192.
Chen, Y. W. (2001). *Acta Cryst.* D**57**, 1457–1461.
Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **4**, 823–826.
Claude, J. B., Suhre, K., Notredame, C., Claverie, J. M. & Abergel, C. (2004). *Nucleic Acids Res.* **32**, W606–W609.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Eddy, S. R. (1998). *Bioinformatics*, **14**, 755–763.
Fischer, D. (2000). *Pac. Symp. Biocomput.* **5**, 119–130.
Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.
Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). *Protein Sci.* **12**, 1691–1698.
Hoppe, W. (1957). *Acta Cryst.* **10**, 750–751.
Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. (2005). *Nucleic Acids Res.* **33**, W284–W288.

Jeong, J. I., Lattman, E. E. & Chirikjian, G. S. (2006). *Acta Cryst.* D**62**, 398–409.
Jones, D. T. (2001). *Acta Cryst.* D**57**, 1428–1434.
Karplus, K., Barrett, C. & Hughey, R. (1998). *Bioinformatics*, **14**, 846–856.
Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 119–124.
Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. E. (2000). *J. Mol. Biol.* **299**, 501–522.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kleywegt, G. J. (1998). *News From The Uppsala Software Factory*. http://xray.bmc.uu.se/usf/factory_6.html
Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
Liu, Z.-J., Lin, D., Tempel, W., Praissman, J. L., Rose, J. P. & Wang, B.-C. (2005). *Acta Cryst.* D**61**, 520–527.
Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.
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.
Reeves, G. A., Dallman, T. J., Redfern, O. C., Akpor, A. & Orengo, C. A. (2006). *J. Mol. Biol.* **360**, 725–741.
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.
Shi, J., Blundell, T. L. & Mizuguchi, K. (2001). *J. Mol. Biol.* **310**, 243–257.
Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. (2004). *Acta Cryst.* D**60**, 1229–1236.
Soding, J. (2005). *Bioinformatics*, **21**, 951–960.
Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.
Suhre, K. & Sanejouand, Y.-H. (2004). *Acta Cryst.* D**60**, 796–799.
Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.
Vogt, G., Etzold, T. & Argos, P. (1995). *J. Mol. Biol.* **249**, 816–831.
Vriend, G. J. (1990). *J. Mol. Graph.* **8**, 52–56.
Xu, Y. & Xu, D. (2000). *Proteins*, **40**, 343–354.
Ye, Y. & Godzik, A. (2005). *Bioinformatics*, **21**, 2362–2369.