Protein structure determination by exhaustive search of Protein Data Bank derived databases

Ian Stokes-Rees* and Piotr Sliz**

*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115; and **Laboratory of Molecular Medicine, Children’s Hospital, Boston, MA 02115

Edited by Axel T. Brunger, Stanford University, Stanford, CA, and approved October 20, 2010 (received for review August 19, 2010)

Parallel sequence and structure alignment tools have become ubiquitous and invaluable at all levels in the study of biological systems. We demonstrate the application and utility of this same parallel search paradigm to the process of protein structure determination, benefitting from the large and growing corpus of known structures. Such searches were previously computationally intractable. Through the method of Wide Search Molecular Replacement, developed here, they can be completed in a few hours with the aid of national-scale federated cyberinfrastructure. By dramatically expanding the range of models considered for structure determination, we show that small (<12% structural coverage) and low sequence identity (<20% identity) template structures can be identified through multidimensional template scoring metrics and used for structure determination. Many new macromolecular complexes can benefit significantly from such a technique due to the lack of known homologous protein folds or sequences. We demonstrate the effectiveness of the method by determining the structure of a full-length p97 homologue from Trichoplusia ni. Example cases with the MHC/T-cell receptor complex and the EmoB protein provide systematic estimates of minimum sequence identity, structure coverage, and structural similarity required for this method to succeed. We describe how this structure-search approach and other novel computationally intensive workflows are made tractable through integration with the US national computational cyberinfrastructure, allowing, for example, rapid processing of the entire Structural Classification of Proteins protein fragment database.

p97 ATPase | likelihood functions | scoring methods | grid computing

Can access to vast quantities of computational power be leveraged to advance the study of biological systems in previously unexplored ways? Whereas many domains have driven demand for computational power and novel computational techniques in the process of scientific investigation, there remain areas where the opportunities provided by the most advanced computational infrastructures and tools have not been fully exploited. The last decade has quietly seen the development of significant national and international federated cyberinfrastructures, established primarily to support the half dozen globally distributed particle physics collaborations. In the same way this community established the World Wide Web as a simple, standards-based system for information sharing, the particle physics community has also facilitated sharing of data and computing through development of what is known as “grid computing.”

An area within the field of macromolecular structural biology that can leverage grid computing is harnessing the large and growing set of known protein structures to accelerate protein structure determination. The question of how to benefit from known structures was posed even as the earliest protein structures emerged, following observation of the similarity of the hemoglobin subunits to each other and to the structure of myoglobin.

The method now known as molecular replacement (MR) was first proposed for macromolecular crystallography by Rossmann and Blow (1), based on ideas developed by Hoppe in the context of small molecule crystallography (2). This was in response to the observation of evident family resemblances among different proteins and to the realization that it would be necessary to determine the structure of a particular protein in multiple states and with multiple ligands. The MR approach bootstraps the process of X-ray crystallographic phase determination by placing a known protein structure template in an orientation and position that aligns with that of the unknown protein. MR has now become the most commonly used method in protein structure determination by X-ray crystallography. It accounts for roughly half of all structures recorded in the Protein Data Bank (PDB) (3), which currently contains almost 70,000 depositions. In traditional MR, a suitable template model is selected based on sequence similarity. Other similar methods in structural biology rely on small databases of short protein fragments [e.g., the “lego” feature in O (4), and molecular fragment replacement in NMR (5)], or homologous structures [e.g., low-resolution refinement in crystallography (6)]. The selection of a suitable candidate template model remains a primary limiting factor in all of these methods. Although several approaches have been proposed for automating the selection of MR template models, either based on sequence information (7–9), or adapting MR algorithms to run in parallel on a specialized cluster (10), none have attempted molecular replacement searches using a complete, PDB-derived database of all available macromolecular domains, or considered the new insights provided by examining the aggregated results from large template model sets. Improved template selection would be expected to accelerate the structure determination process, minimize bias, and extend the range of suitable template models to proteins with negligible template identity.

In this paper, we ask three questions. First, can we compare results from independent molecular replacement runs and use these results to discriminate and rank solutions, thereby justifying the use of large template model databases? Second, can we develop improved criteria for recognizing correct solutions, in order ultimately to improve the convergence and speed of MR and further automatic structure determination? Third, can existing applications be scaled, deployed, and executed in a grid computing environment to enable new avenues of investigation, rather than merely faster computation? To answer the first question, we evaluated three diverse structure determination scenarios: (i) optimal selection in cases with several template model candidates; e.g., an MHC–TCR complex with 5,000 potential peptide binding or Ig domains that could be used as a template models in the MR search; (ii) structural homolog searches in cases for which sequence-based searches fail to identify usable MR template models; and (iii) “blind” cases in which the sequence of the crystalized sample is unknown. By adapting the widely used Phaser (11) MR application to the format of grid computing, we demonstrate the

Author contributions: I.S.-R. and P.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

To whom correspondence should be addressed. E-mail: piotr_sliz@hms.harvard.edu.
This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1012095107/-/DCSupplemental.

21476-21481 | PNAS | December 14, 2010 | vol. 107 | no. 50
www.pnas.org/cgi/doi/10.1073/pnas.1012095107
power of this unique wide search molecular replacement (WS-MR) approach, which can be used to search up to 100,000 domains in a few hours and to provide the range of results necessary to answer the questions posed above. WS-MR successfully identifies the closest structural homologues from a large family of candidates and does so more reliably than traditional, sequence-based approaches. The approach is also successful in identification of domains with marginal sequence identity or coverage. We use the WS-MR method to determine a structure of the full-length insect homologue of p97, a mammalian AAA⁺ ATPase (12–14) that was crystallized as a contaminant and reveals a previously unobserved D1 ADP-free conformation. Based on the extensive collection of results from the completed cases we demonstrate that incorporating multivariate scoring metrics [e.g., Phaser’s log likelihood gain (LLG) and translation function Z-score (TFZ)], or classification and clustering [e.g., Structural Classification of Proteins (SCOP) class and domain size], significantly improves discrimination to identify the best solutions. The computations for WS-MR were performed using the federated computing environment of the Open Science Grid (OSG) (15), illustrating how the national distributed cyberinfrastructure can be effectively used to develop and support unique computational workflows in research areas outside of physics.

Results
1. Comparison of Results from Independent Molecular Replacement Runs. a. Selecting the Best Model from a Large Library of Homologous Structures. We selected the MHC–TCR complex as the first system to validate the WS-MR approach. The structure contains one peptide binding domain (MHC–PBD) and six immunoglobulin (Ig) domains (Fig. 1A). There are over 5,100 candidate domains out of the 95,000 domains found in the Structural Classification of Proteins (SCOP) database (16) (Methods) that could map to parts of this structure, thus providing a useful spectrum of results to correlate the degree of structure coverage, sequence identity, and structural similarity with the quality of the initial phases. WS-MR, using the full SCOP database with the MHC–TCR reflection data [PDB code 2VLJ (17)], was used to determine whether structurally similar models rank best (in terms of various MR scoring metrics) and whether these models can be identified from incorrectly placed domains and from other structures in the database. This case is representative of using WS-MR for an unknown structure with many homologues, where it could be used to select the best model. This would be especially useful in cases where model coverage or sequence identity are low.

The WS-MR search was completed in 12 hours of elapsed time (800 processor-days of computing time) utilizing a small subset of idle computers in the otherwise highly subscribed resources in OSG. This level of performance was typical of all the WS-MR iterations described here. Collected results allowed quick identification of a group of distinct, viable, MR models. Whereas several scoring functions were used to evaluate the quality of Phaser placement results (see section 2), a two-dimensional quality measure based on the LLG and the TFZ provides the best discrimination of results, producing a cluster of approximately 300 candidate domains from the search set of 95,000 (the “top cluster,” Fig. 1B). Domains in the top cluster all belong to SCOP class d.19.1.1, the MHC–PBD domain that represents 20% of the full model, and are all placed correctly by Phaser, in reference to the actual structure (SI Text). Three Ig domains are also identified in the top cluster (12% of search model), and no false positives are observed. The above results provide the boundary for the molecular replacement search to produce correct and identifiable placements for the MHC–TCR example with a model completeness between 12% (in case of the Ig domains) and 20% (for the MHC–PBD). The likelihood of obtaining the correct and identifiable placement with high quality models is very small when searching with 12% of the target (3 in 4,500 Ig domains, Fig. 1B) and dramatically increases for a search with 20% of the target (300 in 550 MHC–PBD domains, Fig. 1B).

WS-MR not only discriminates correctly placed models but, in this case, also orders them by the similarity of the structure and the target molecule (Fig. 1C). For the correctly placed MHC–PBD models, LLG/TFZ is highly correlated to RMSD between the model and the reference structure. For example, the lowest RMSD model also scores the highest on the LLG/TFZ scale. In comparison, selecting models based exclusively on sequence identity results in a wide range of LLG/TFZ values, even for the subset with identities >90% (Fig. S1B). In this test case LLG-based selection provides superior distinction of correct solutions compared to sequence similarity and would therefore provide an advantage for MR model selection.

As expected, placement of the best first domain identified by WS-MR (an MHC–PBD) facilitated completion of structure determination. Repeating WS-MR with the MHC–PBD domain fixed placed over 1,000 Ig domains in the top cluster result from the second WS-MR iteration, and further analysis confirmed that all six MHC–TCR Ig domains are found in this set (Fig. 1D). Here the LLG scores correlate strongly with the structural similarity (see linear fit lines in Fig. 2A). Whereas the search for the first MHC–TCR fragment required a minimum 60% sequence identity to obtain identifiable solutions, in the secondary search individual Ig domains with as little as 11.6% sequence identity produced identifiable results (see Section 1b), a noteworthy success of the partially phased MR approach with 20% placed, a 12% search fragment and 68% of the structure still missing.

b. Identifying Good MR Models with Marginal Sequence Identity. Perhaps the most intriguing opportunity for the WS-MR structure determination technique is the application of a blind search in cases where traditional MR techniques fail, and before attempting further experimental phasing methods. Blind WS-MR, where no template filtering is applied and the full template database is
searched, can reveal structures that would not otherwise be identified by sequence alignment algorithms [which generally provide poor results when the best sequence-based homologues have an identity of less than 30% (9)]. Such searches make no a priori assumptions about the target structure and can utilize large databases of PDB-derived models. The infrastructure described in section 3 makes this approach feasible, and the trend of decreasing cost per unit of processing power is such that in the next few years such a workflow could be executed solely by the internal computational resources of a single laboratory.

In a limited number of completed searches we observe that models with borderline sequence identity (between 10–20%) can work well. For example, in the MHC–TCR example described above, in the secondary search with the MHC–PBD placed, the majority of Ig domains with sequence identity below 20% failed to be correctly placed, but the placement of 244 domains was correct (gray vs. colored dots in Fig. 2). All but 17 of the correctly placed domains could be readily identified based on LLG and TFZ scores, indicating a false negative rate for this set (sequence identity below 20%) of 7%, and a clear LLG cut-off of 130, above which 100% of the results were correct, including domains with sequence identity as low as 11.6%.

Further tests of WS-MR were carried out on structures that had previously been determined by experimental phasing methods. A search with data for EmoB (18) (PDB code: 2VZF) was performed with the SCOP database, and returned a clear cluster of 14 solutions (Fig. 3A). All 14 models that belong to SCOP flavoprotein classes c.23.5.4 and c.23.5.8 are positioned properly, while all remaining 182 flavoprotein domains in the bottom cluster, except for 4, are incorrectly placed. The 14 correctly placed and identifiable models have sequence identities of 13% to 21%, and RMSD between 2.2 and 2.7 Å relative to the reference structure (Fig. S2A and B). The top solution can be used to rapidly refine the structure. In contrast, four iterations of a PSI-BLAST search identified 313 candidates, of which two were in the group of 14 identified by the WS-MR approach as suitable MR templates. These two were conformationally similar structures with mutually identical sequence, but a sequence identity to the target structure of only 17%. PSI-BLAST failed to identify any of the other twelve structures, despite having sequence identities in the same range (13–21%). Whereas in this particular case sequence-based approaches should converge on a correct solution, the unpredictability of successful molecular replacement results combined with the difficulty of selecting models by sequence-based searches explain why viable MR models may be missed in other similar cases.

We have also recorded several cases, when performing full SCOP WS-MR searches, where the identified solutions share significant structural characteristics with the target but are too divergent to produce the correct placement. For example WS-MR with experimental data for the kinase domain of *Escherichia coli* tyrosine kinase ETK (PDB code: 3CIO) retrieves no strong results, but after closer inspection of the LLG/TFZ profile, we selected two solutions with relatively high TFZ score (>6), and LLG scores separated from other results. One of those peaks corresponds to SCOP model 1Z0Fa1—a Rab GTPase (Fig. 3B). The two structures have 12.5% pairwise sequence identity, a misleading metric given that the two proteins can only be superposed in a sequence independent manner (Fig. S2C). In another case, a structure of a four helix protein recently deposited by the Midwest Center for Structural Genomics (PDB code: 3CEX) can be superposed on a SCOP domain from ferritin (IIESea) (Fig. 3C). The superposition of the four helical elements is sequence inde-
Lepidoptera, Hi-5 cells) of a mammalian p97, a hexameric AAA does not correspond to the actual sequence identity for the to our solicitation for recalcitrant datasets od on five cases provided for evaluation by colleagues in response database. Four datasets were immediately confirmed as contami- (e.g., migration SDS-PAGE or mass spectrometry profile). For each dataset we performed WS-MR with the full SCOP database. Four datasets were immediately confirmed as contaminants. The most striking was a homolog from Tribolium castaneum (order Lepidoptera, Hi-5 cells) of a mammalian p97, a hexameric AAA+ ATPase, which is characterized by poorly diffracting crystals (6) and multiple nucleotide binding states (19). The Tni protein remains unsequenced, but we expect it to be very similar to the sequenced Bombyx mori transitional endoplasmic reticulum ATPase Terp94 (also order Lepidoptera, accession codes: BAE54254 and NP_001037003), which in turn is 83% identical to the full-length Mus musculus p97. WS-MR clearly identified nine domains in a distinct high scoring cluster (Fig. 4A). The overall architecture of Tni p97 closely resembles the structure of M.musculus p97, and the space group matches the I417 structure. Inspection of fo-fo electron density maps suggests, however, that in contrast to other p97 crystal structures (1E32; 1YQ0; 1YQ1; 1YWP) (Fig. 4B), the Tni p97 is nucleotide-free in the D1 binding pocket (Fig. 4C). Although spectroscopic analysis of the protein sample will be required to confirm that indeed all of the symmetry-related molecules in Tni p97 are ADP-free, the unexpected results of WS-MR in this case reveals another potentially valuable utility of the method. Other contaminants retrieved by WS-MR include carbonic anhydrase (1I6Oa_), inorganic phosphatase (1MJWb_), and pyruvate kinase (1AQFg2). In each case, WS-MR provided a quick, conclusive answer to problems that could not be readily addressed using standard biochemical tools.

2. Improved Criteria for Recognizing Correct Solutions. By collecting a large number of data points in many dimensions for several different target structures, we are able to consider techniques beyond the traditional TFZ score to identify viable MR models. We find that Phaser LLG and TFZ scores, in particular, combine to provide good discrimination of templates when strong MR models exist. When combined with LLG, TFZ scores as low as 3.5 are associated with positive results in the correctly placed top cluster. High TFZ (greater than 7) indicates a good MR solution, but our findings show that a low TFZ can, in some cases, also represent a usable MR solution. It is already well known that the LLG scores for different template models are comparable for the same set of reflection data, and this feature is used by Phaser when presented simultaneously with multiple candidate models. WS-MR greatly expands the number and efficiency of intermodel comparisons by LLG that are possible, and thus, we hypothesized, would improve the process of identifying good MR models. We can further augment the sensitivity of the scoring function by incorporating additional dimensions, such as rotation function Z-score (Fig. S3A), domain length (Fig. S4), or domain class clustering. Other measures such as R-factor improvement or contrast as provided by Molrep (STText and Fig. S3B) (20) are less suitable for cross-model comparison. For example we carried out Phenix refinement protocols for several single domain MR solutions to the MHC–TCR example. Only the best solution has an R-factor that falls below 50.0, and for other cases R-free does not improve, most likely because of the limited convergence of refinement with partial model information.

3. Efficiency and Reliability of Molecular Replacement Computations Executed on Grids. All computations in this project were carried out on “opportunistic” resources of OSG. This required accessing 20–30 computing centers that participate in the OSG federation and have allowed our scientific domain (structural biology) to utilize the otherwise idle computing resources of their clusters. To benefit fully from this national cyberinfrastructure, we established a software and hardware environment that can manage and support both general and specialized types of grid computations. Unlike a desktop or cluster computing environment, where the configuration of the system is fixed and well known, grid computing introduces complexities that require new approaches rather than simple reconfiguration of existing programs. The dynamic nature of grids with a high level of unpredictable faults, federation, geographic distribution, and system heterogeneity present significant challenges. We have therefore developed unique strategies for the synchronization and flow of data and applications at four grid levels: “static” (constantly available), “workflow” (a related set of computations), “grid job” (a single instance of grid resource utilization), and “atomic job” (the smallest computa- tional unit that produces a distinct result as part of the workflow, but may be too small to efficiently run as an independent grid job) (Fig. 5). By tracking application and script versioning, and by considering the permanence and relevance of data, we can reduce the obstacles presented by network latency and multilevel grid faults of caching to maximally localize data and computations while minimizing data movement. We have combined these efforts with caching to maximize and GridSite (24), with an underlying security layer provided by X.509-based public-key cryptography and higher layer workflow.
We have demonstrated that WS-MR is able to discriminate strong molecular replacement template models with marginal sequence identity and coverage, identifying top candidates for subsequent density modification, model building, and refinement steps. In rare cases templates comprising as little as 6% of the scattering matter, or having sequence identity below 20%, have been shown to produce correct MR placement results. Validating or utilizing templates with such characteristics is typically difficult. A routine evaluation of all marginal fragments typically requires several cycles of model building and refinement, can be time consuming, and it is not always clear if the results are correct. Wide search comparison of several domains based on multidimensional scoring metrics greatly accelerates the validation process. Our results suggest that the limit of sequence identity for successful WS-MR search is low enough to allow our method to extend to models that would otherwise be missed by methods that are based on sequence alignment for template selection. Both remote homologues and structural analogs can be detected by WS-MR, with specific examples where models with an identity of 11.6% and an RMSD under 3 Å can be correctly placed and distinguished from negative results. We also show that low completeness with structure coverage of as little as 12% can be sufficient for good WS-MR template models, however in these cases high sequence identity and structural similarity for the covered area are required.

By using an approach in which no a priori knowledge or primary sequence information is required for search model selection, we have expanded the probability of success for difficult molecular replacement problems in X-ray crystal structure determination. Utilizing this system is straightforward, as the only required input is the reflection data. Additionally, initial search constraints (e.g., sequence, predicted secondary structure profile, molecular weight, oligomerization state) can be provided to optimize the search, or previously placed domains in the case of subsequent domain searches for multidomain structures. The output of WS-MR provides both graphical and tabular summary representations of the results, allowing rapid identification of the best candidate MR template models. The user would then attempt to validate a few top scoring solutions using standard approaches, such as packing analysis or interpretation of density modified difference maps. If a particular solution looked plausible, a search for missing components of a given structure, or a manual or automatic rebuilding process could be attempted. To encourage rapid convergence to the best MR models (if they exist), the WS-MR strategy can proceed iteratively, starting with the most promising models based on the specified constraints, for example using the top 100 sequence-similar models, and include a small control set that is widely representative of known domains (for contrasting expected negative results). If no promising models are returned from the initial constrained search, subsequent iterations can relax the selection criteria to associated domain classes, thus expanding the number of search models, eventually considering all known domains. Although it is not possible to predict whether a less-than-exhaustive WS-MR search is necessary (if obvious models existed, conventional MR would suffice), this iterative approach will avoid an exhaustive search if promising models are discovered from the constrained search set. The WS-MR method is accessible and applicable to many crystallographic projects, and it allows the search of arbitrary structure databases, constructed dynamically from selection criteria or from preexisting sets. The WS-MR approach becomes increasingly powerful as more structures are determined and made publicly available.

A benefit of the large result sets produced by WS-MR is the ability to evaluate algorithmic improvements that should result in better scoring and discrimination of search models, in particular a reduction of false negatives. Our work on several WS-MR test cases has provided unique insights leading to improved scoring and model discrimination strategies. By using multiple scoring metrics (such as LLG and TFZ) from the high quality maximum likelihood algorithm in Phaser, it is possible to distinguish correct solutions by cluster identification. In the case of weak (but still valid) MR templates, we have shown that effective model discrimination is significantly aided by these additional metrics. Fig. S4 illustrates how the additional consideration of model size allows for the clear identification of several correctly placed Ig domain models for the MHC–TCR case that were not identifiable from only the LLG and TFZ data. LLG led to the selection of several correctly placed models in the EmoB case (Fig. 3A). Classification (e.g., SCOP class) or MR placement clustering (similar domains placed in the same orientation and location) can also provide a mechanism to identify groups of viable MR models. One important observation for the results of exhaustive WS-MR is that small domains can lead to anomalously high TFZ scores (greater than 10), due either to insufficient statistics or the ability of very small fragments to match accurately to some region of a large unknown structure. Nevertheless, these anomalous results also benefit from the addition of LLG scoring, as they consistently have LLG scores below 20 and can therefore be easily identified and discounted.

Without existing infrastructure, a transition to grid computing requires a significant time investment and presents numerous unexpected hurdles. The challenge in accessing and deploying applications into a grid environment can be simplified for the end user by the development of web-based portals, an approach that has proved successful for many other grid environments [e.g., TeraGrid Science Gateways (31)]. The SBGrid Science Portal that we have developed will make the...
WS-MR technique described here was available to the entire community. Using OSG to perform WS-MR for the cases described here, we typically accessed 2,000–5,000 computing cores concurrently, thus completing what would otherwise have required several years of computing within one day. Access to the national cyberinfrastructure makes it possible for any individual research group to develop novel computational workflows that take advantage of large federated resources, in particular idle cycles that would otherwise be wasted. Computers in a typical scientific computing cluster spend around half their time lightly utilized (less than 10% load), but even then they typically consume more than 80% of the maximum power consumption at full load (32). This presents a tremendous computational opportunity with relatively minor cost overhead. With a transition to a new resource access and scheduling mechanism, using GlideinWMS (26), we have been able to execute up to 7,000 concurrent computations using this pool of otherwise idle computers, well above what is currently available to a typical research group.

Arguably more important than the WS-MR technique itself are the opportunities to reuse the framework that has been developed for large scale data processing and computation. We have started work on problems in NMR, electron microscopy and in other areas of X-ray crystallography that use this foundational infrastructure and the capacity provided by OSG. Any scientific application that can run without active user interaction can be deployed into a grid environment with a suitable workflow management protocol for data staging, results aggregation, and analysis. We have shown that it is not necessary to redesign applications and algorithms to benefit from these advances. Existing applications can be used in new ways with statistical and data visualization techniques applied to aggregate and filter orders of magnitude higher data volumes than the application designers intended, leading to new challenges for interpretation and discovery.

**Methods**

The SCOP domains utilized for WS-MR were taken from the November 2007 (1.73) release (16, 33). Molecular replacement computations were performed with Phaser (version 2.1.4), and Molrep (version 10.2.3). We used a modified version of TM-Align (34) to perform structural alignment and combination of TM-Align and Reforinig (CCPM, version 6.1.2) (35) to calculate placement quality and placement correctness. Scheduling of jobs to OSG sites was managed through a combination of Condor DAGMan (25) and the OSG Match Maker. Density modification and model building of the MHC–TCR and Emo8 models were performed in Phenix Autobuild (36) starting with Phaser Sigma(A)-type weighted fourier maps (FWT/PHWT) (37) and amplitudes with standard deviations from the Protein Data Bank structure factor files. Detailed protocols are described in *SI Text*.

**ACKNOWLEDGMENTS.** We thank Peter Doherty for grid computing support, Mats Rynge for Matchmaker customizations, Steve Timm for operational assistance, and Stephen C. Harrison and Yusun Nam for discussion and critical review of the manuscript. The work was supported by National Science Foundation Grant 0639193 (P.S.), and National Institutes of Health Grant P01 GM062580 (to Stephen C. Harrison). This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the US Department of Energy's Office of Science.

Supporting Information

Stokes-Rees and Sliz 10.1073/pnas.1012095107

SI Methods

Structural Classification of Proteins Database. The Structural Classification of Proteins (SCOP) domains utilized for wide search molecular replacement (WS-MR) were taken from the November 2007 (1.73) release (1, 2). A later version of the SCOP corpus was released in June 2009 (1.75), however this contains domains from the structures used for validation tests, and was therefore not used. The 97,169 domains in the SCOP-1.73 corpus were filtered to retain only protein residues, with a single set of non-hydrogen C-alpha coordinates per domain (i.e., not against each other for NMR models retained only the first card). This resulted in a modified and reduced set of 95,838 domains. When an occupancy was set to 0.0 it was fixed instead to 1.0 without regard for other atom entries in the residue, including NMR models. Finally a model was only kept if the average ratio of non-carbon-alpha atoms to carbon-alpha atoms per residue was greater than or equal to 4.0. This eliminated models that were insufficiently complete, including all models that contained only the carbon-alpha atom for each residue. The resulting 95,838 modified models are referred to as SCOPCLEAN in the following discussion of methods.

Molecular Replacement Search. Structure factor files for trial structures were retrieved from the Protein Data Bank using a combination of the CIF format data to retrieve unit cell and symmetry parameters and the structure factor data to produce an MTZ format file. A combination of mtzdump and cif2mtz utilities from the CCP4 (3) distribution were used in this conversion (from CCP4 version 6.1.2). Phaser (4) (version 2.1.4, as distributed with CCP4) was then run with the trial reflection data and a single set of non-hydrogen C-alpha coordinates per domain (i.e., not against each other for NMR models retained only the first card). This resulted in a modified and reduced set of 95,838 domains. When an occupancy was set to 0.0 it was fixed instead to 1.0 without regard for other atom entries in the residue, including NMR models. Finally a model was only kept if the average ratio of non-carbon-alpha atoms to carbon-alpha atoms per residue was greater than or equal to 4.0. This eliminated models that were insufficiently complete, including all models that contained only the carbon-alpha atom for each residue. The resulting 95,838 modified models are referred to as SCOPCLEAN in the following discussion of methods.

For algorithmic comparison, the same process was repeated using Molrep (5) (version 10.2.30, as distributed with CCP4). The parameters were set as follows: monomer search, fast mode, 30% structure similarity, 50% completeness, 20 rotation peaks, and 20 translation peaks. For each Molrep instance the following quality measures were retained for scoring: R-factor, MR-score (a Molrep-specific heuristic scoring function that combines the correlation coefficient and the packing function), and contrast (ratio of top score to mean score). In addition, the 3-axis translation and 3-axis rotation values for the placed model, and the execution logging output were retained. All other output was discarded (such as reflection data augmented with phasing and the placed structure PDB format file).

From the bulk of the incorrect ones (Fig. S3B). The ability to discriminate correct solutions from incorrect ones is central to WS-MR due to the size of the search space. Therefore, we selected the Phaser LLG/TFZ quality metrics in all cases discussed here.

Each algorithm, run to completion and with valid input files and parameters, had two possible outcomes: success, or no solution. Success indicated the MR algorithm found some placement for the search model. No solution indicated an inability to find a suitable orientation for the search model. We can then perform a structural comparison of every domain of the validation structure to every domain in SCOPCLEAN to identify the maximum number of molecular replacement candidates the WS-MR technique can be expected to produce for a given validation structure. This assumes that any search model that can be used successfully for molecular replacement will also have a strong structural alignment with one of the domains of the actual structure. The deposited PDB file for each validation structure was decomposed into individual domains. Where possible, these domains were taken from the latest 1.75 version of SCOP, otherwise they were manually prepared. To perform the structural alignment a modified version of TM-Align (6) was used. TM-Align is only able to align a single chain in a given PDB file, therefore in cases where alignment was performed against protein complexes either the chains were arbitrarily merged, resulting in nonphysical representations but still sufficient and accurate for the purposes of structural alignment, or each chain was extracted into an independent PDB file. TM-Align produces the following alignment metrics: residues aligned, sequence identity over aligned region, RMSD of aligned region, TM-Score (a heuristic that combines sequence identity, fraction of residues aligned, and RMSD), fraction of target aligned, and fraction of search model aligned. In addition, it outputs a transformation matrix that can be used to map the search model to the validation target model. Fig. S5B illustrates the TM-Align RMSD error vs. the length of the structurally aligned segment using SCOPCLEAN and the MHC–TCR test case structure.

Determining Model Placement Quality and Placement Correctness. For each instance where a molecular replacement algorithm returns a placed search model it is necessary to ask if the placement is correct. This is a distinct question from “useful,” in that a useful
placement will (i) be correct; and (ii) be sufficient to aid in phasing and full refinement. Due to the nature of molecular replacement as a technique for phase determination, it is clear that a placement algorithm must place a search model sufficiently well for further refinement to have the possibility of converging to the correct solution. Placement quality checking can only be done for validation structures (i.e., ones where the structure is known).

For absolute placement quality checking, we first create a reference placement of the search model with each domain in the validation structure, using the known structure model for that domain. This placement is done using the transformation matrix produced by TM-Align, and does not incorporate any MR placement information. This reference placement is approximately what we would expect from the MR algorithm in the event the given search model was a suitable MR candidate for the given validation structure domain. Next we test whether the actual placement produced by the MR algorithm is equivalent to a symmetry pair of origin-shifted reference placement. We augment both the reference placement and the actual placement with space group and unit cell parameters taken from the validation structure and then utilize the reforigin utility (from CCP4 version 6.1.2) to check for the closest pair between these two placements and calculate placement quality (lowest RMSD between actual and reference copy of the search model). All symmetry equivalents and origin-shifted structures of the reference placement within 100 Å from the actual placement are considered. In all validation tests we observe a rapid transition from low to high placement quality (Fig. S1A). Based on this experience, we define a measure “placement quality” where values less than 1.5 Å are correctly placed, and those greater than 5 Å are incorrectly placed. The placement quality gap between 1.5 and 5 Å typically has less than 2% of the search structures, and therefore 2.0 Å serves as a suitable classification boundary. Fig. S1A illustrates this for the placement of the Ig domains of the MHC–TCR complex, showing 460 domains correctly placed, 4,000 incorrectly placed, and 40 in the placement quality gap. In reference to the “blind” WS-MR results scoring for the first domain search in the MHC–TCR scenario, the top cluster of 300 models consisted entirely of MHC–PBD models, and their placement quality scores were all less than 0.4 Å, which is interpreted as no false positives. In terms of false negatives, there are only half a dozen or so MHC–PBD models that are correctly placed yet not in the top cluster. These are all found on the top fringe of the “bulk” results, as illustrated in Fig. 1 (green dots along bulk results fringe). The 270 MHC–PBD models found mixed in with the bulk results (red dots, Fig. 1 in the main text) all have high values for placement quality (>5 Å) and can therefore be identified as only true negatives.

Computational Infrastructure. To perform numerous iterations of WS-MR on the full SCOP database it was necessary to access the opportunistic compute resources made available by Open Science Grid (OSG) (7). A single Phaser-based global search for a typical crystal requires approximately 20,000 core hours for the Science Grid (OSG) (11) and the OSG Match Maker. More recently GlideinWMS (12), which has allowed over 7000 concurrent jobs, has been deployed. DAGMan provides a mechanism to describe the dependencies between the sets of grid jobs and has facilities that can manage error recovery. The OSG Match Maker is a scheduling system that makes decisions about allocation of grid jobs to available OSG computing centers and maintains status and rank information on computing centers based on the results of previous jobs that have executed there. To reduce network traffic at the job source, the necessary applications and common data (e.g., SCOPCLEAN corpus) were prestaged to each computing center. Maintenance systems ensure these stay up to date. Individual job execution was handled by a wrapper that configures the system environment appropriately and retrieves any job-specific files, such as the reflection data or preplaced structures (for second and subsequent round searches on the same structure). Although both Condor and DAGMan provide mechanisms for error recovery it was still typically the case that 1–5% of results would not be returned from a particular search, due to various forms of failure. Even these failure rates were only achieved after initial experience of >50% job failure rate, and the consequent introduction of system tuning and fault tolerance mechanisms. A semiautomated mechanism was developed to retry any missing results until >99.8% of results were available. All results were then aggregated, filtered, and sorted, then augmented with results from other searches (such as TM-Align comparison, Reforigin placement, or Molrep), and with “static” data related to each individual SCOP domain (such as the SCOP class, the domain size, or the domain description). This process resulted in large tabular datasets that could be processed into reports or analyzed with the assistance of visualization software.

Refinement and Model Building. Density modification was performed in Phenix Autobuild (13) starting with Phaser Sigma(A) (14) type weighted fourier maps (FWT/PHWT) and amplitudes with standard deviations from the Protein Data Bank structure factor files. Sequences of model structures were also included. Version 1.6-289 of Phenix was used, “rebuild_in_place” was forced, and all other default parameters of the Phenix Autobuild Wizard were applied.

For the MHC–TCR complex, density modified maps for the PBD domain with three Ig domains placed were calculated with the CCP4 application Pirate (3). Initial phases were calculated with CCP4 sfall, weighted with CCP4 sigmaa and converted to Hendrickson–Lattman coefficients with chltofom. cpiral was then run for 5 cycles with default input parameters.

Partial models with the MHC–PBD placed by WS-MR (SCOP domains 1HM3a2, 1HM6a2 and 1ZAGb2) were refined in Phenix. Models were first prepared with the “ready-set-go” Phenix utility, and then subjected to three cycles of refinement. Each cycle included rigid body refinement, torsion angle annealing starting at 2,500 K and refinement of atom displacement parameters.

WS-MR for the Trichoplusia ni p97 dataset (with a 3.8 Å resolution limit) identified nine SCOP domains (1R7Ra2, 1S3Sa2/b2/c2/d2/e2/f2, 1E32a2, 1OZ4c2), in a distinct, high scoring cluster (Fig. 4/4). These domains correspond to the D1 domain of M. musculus p97. We have subsequently reprocessed the dataset with an anisotropic correction (15) that extended resolution in two dimensions to 3.2 Å, placed the N-terminal, and D2 domains of the p97 structure, and refined the coordinates with Phenix Autobuild using the silkworm sequence. The R/R-free values
are 25.31/30.29% for the 3.8 Å dataset and 27.6/35.7 for the anisotropically corrected dataset.


Fig. S1. MHC–TCR case (A)—Determination of placement quality. We take a placement quality <1.5 Å as correctly placed and >5 Å as incorrectly placed. The illustration is for the MHC–TCR complex second round WS-MR search for Ig domains, with the MHC–PBD from the first round fixed. The transition in placement quality from correctly to incorrectly placed is not gradual but very rapid. We call the transition region “the placement quality gap” and have found that less than 2% of structures fall into this region, with the remainder clustering strongly into correctly or incorrectly placed groups. Placement quality is calculated taking into account potential origin shift, and symmetry equivalent positions (see SI Methods for full description). (B) Relationship between sequence identity and LLG for MHC–TCR first round search results for MHC–PBD models, illustrating the clear division in both sequence identity and LLG between the correctly placed domains (green) and the incorrectly placed domains (red).
Fig. S2. Screening datasets of distant homologs. (A) TM-Align RMSD scores for all c.23.5 domains calculated against actual structure. Top green row represents distinct cluster of 14 correctly placed domains, lower green row represents 4 correctly placed domains with LLG/TFZ scores indistinguishable from negative cases. Red row indicates remaining c.23.5 domains, which were incorrectly placed, and gray domains represent c.23.5 domains that did not produce MR results. (B) Sequence identity for all c.23.5 domains calculated against actual structure. Coloring as for Fig. 4E. (C) Topology diagram for PDB target (3CEX) and matching model (SCOP code 1IESa_). Corresponding helices in both structures are numbered accordingly. Light blue boxes indicate an antiparallel match. An arrow on the LLG/TFZ graph indicates the 1iesa_ solution. (D) Topology diagram for PDB target (3CIO) and matching model (SCOP code 1Z0Fa1). Corresponding helices in both structures are numbered accordingly. Light blue boxes indicate an antiparallel match (3CEX helices 1-2-3-4 align with 1IESa_ helices 4r-1-2r-3). An arrow on the LLG/TFZ graph indicates the 1Z0Fa1 solution.
Fig. S3. Evaluation of global search scoring functions for PBD domain of MHC–TCR complex. (A) Translation function Z-score vs. rotation function Z-score with LLG heat map for Phaser results of MHC–TCR first round PBD domain search, indicating high correlation of TFZ, RFZ, and LLG in this case. (B) Molrep contrast vs. score for MHC–TCR global siMR search. This illustrates the relatively weak discriminating ability of Molrep to identify a cluster of template candidates. Green indicates 270 correctly placed PBD domains, red indicates 300 incorrectly placed domains, and gray all other SCOP domains with MR results.

Fig. S4. LLG vs. model length for MHC–TCR complex, first round search. Dark blue points represent MHC–PBD models and cluster predominantly around 181 residues (the length of the actual domain). The correctly placed models are all in the upper vertical cluster. The orange points indicate the Ig domains, and the inset window provides a close-up of these results, illustrating several correctly placed Ig domains, some of which are correctly placed but were not identifiable from LLG vs. TFZ scatter plot (the three E2 domain models). This demonstrates the potential for improved model discrimination by introducing additional scoring metrics, for example the model length and an LLG correction (quadratic fit, indicated by purple curve).

Fig. S5. MHC–TCR complex test case. (A) Phaser LLG vs. run-time. This illustrates the clustering of high LLG results all with run times less than 300 s. Timeout was set to 1,800 s (30 minutes). a: top 200 mean run-time (152 s); b: top 200 mean run-time + 5 standard deviations (283 s); c: timeout (1800 s); d: LLG limit of 50 for top 200 results. (B) RMSD vs. alignment length, illustrating clusters of low RMSD solutions for MHC–PDB domains (∼180 residues) and Ig domains (∼100 and ∼125 residues).