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 ATOM coordinates per domain (i.e., multicard domains 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 mtzdmp 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 against each of the 95,838 modified SCOP domains in a monomer search mode, with solvent fraction fixed to 50% and template model to structure sequence identity fixed to 30%. The "automatic" mode was used (MR_AUTO). For each Phaser instance the following quality measures were retained for scoring: rotation Z-score (RFZ), translation Z-score (TFZ), packing faults (PAK), log likelihood gain (LLG_INITIAL), and refined log likelihood gain (LLG). In this work all references to LLG refer to the final refined LLG. In cases where this was not available, the entire model result was ignored. 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).

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. As with Phaser, all other output was discarded. Upon evaluation of the results for the full SCOP search of the MHC–TCR complex, we found 270 PBD domains to be correctly placed. Molrep provides three quality measures: R-factor, contrast, and score. None of the scoring functions provided a distinct cluster of solutions. A weak grouping of the MHC–PBD domain can be achieved by combining score and contrast, but many of the correct solutions remain indistinguishable from the bulk of the incorrect ones (Fig. S3*B*). 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 any suitable orientation for the search model given the reflection data. A third acceptable outcome was a timeout—each instance was run with a 30-minute timeout to eliminate cases where there is unlikely to be convergence to a correct solution. If the MR algorithm returned in any state other than "success," "no solution," or "timeout" it was retried. As an example, for the MHC-TCR structure (2VLJ), WS-MR returned 91,730 SCOP domains with successful placement, no solution was found for 1723, and the remaining 3434 domains timed out. To validate the timeout period had not been set too short it needed to be >5 standard deviations above the mean of the top 200 results (by LLG for Phaser, and by MR-score for Molrep). See Fig. S5*A*, illustrating the top 200 results for the MHC–TCR structure, the mean Phaser runtime, the 5 sigma limit ("a" marks mean of top 200, "b" marks 5 sigma limit, "c" marks timeout, and "d" marks cutoff of top 200), and the actual timeout (30 minutes, in this case).

Structural Alignment, Sequence Identity, and RMSD calculations. To validate the technique we have used structures from 2008 onward that have high resolution PDB files deposited with the Protein Data Bank and reflection data available. 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. S5*B* 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 aide 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 were 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 or 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 calcualate 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. S1*A*). 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. S1*A* 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 using the SCOPCLEAN corpus—this is more than 2 years of serial compute time. Opportunistic use of OSG allows researchers to access high performance computing centers that are part of the OSG federation, consisting of more than 50 institutions and aggregating over 60,000 processor cores.

The global searches described here would commonly execute 2000–5000 concurrent processes at more than 20 computing centers, allowing a single iteration of WS-MR on the full SCOP database to complete in less than 24 hours. The key software components for OSG are provided by VDT (8), Condor (9), and Globus (10) and provide the basic services, security infrastructure, data, and job management tools necessary to create, submit, and manage computations on OSG. To balance computation time with grid infrastructure overhead it was necessary to set time limits on individual molecular replacement instances (typically 30 minutes), and also to group single instances of the MR computation into sets to produce grid jobs that required 0.5–12 hours to complete—shorter or longer than this could produce problems with the grid infrastructure. Scheduling of jobs to sites was managed through a combination of Condor DAGMan (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. cpirate was then run for 5 cycles with default input parameters.

Partial models with the MHC–PBD placed by WS-MR (SCOP domains 1IM3a2, 1MHCa2 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 atomic displacement parameters.

WS-MR for the *Trichplusia 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*A*). These domains correspond to the D1 domain from *M. musculus* p97. We have subsequently reprocessed the dataset with an anisoptropy 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.

Figures of molecules and electron density maps were prepared in CCP4MG (15).

1. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540.
2. Andreeva A, et al. (2008) Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Res* 36:D419–425.
3. Collaborative Computational Project N (1994) The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr D* D 50:760–763.
4. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40: 658–674.
5. Vagin A, Teplyakov A (2010) Molecular replacement with MOLREP. *Acta Crystallogr D* 66:22–25.
6. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 33:2302–2309.
7. Pordes R, et al. (2007) The Open Science Grid. *J Phys Conf Ser* 78:012057.
8. Roy A (2009) Building and testing a production quality grid software distribution for the Open Science Grid. *J Phys Conf Ser* 180.
9. Thain D, Tannenbaum T, Livny M (2003) Condor and the grid. *Grid Computing: Making the Global Infrastructure a Reality* (Wiley, London) pp 299–335.
10. Foster I (2005) Globus Toolkit Version 4: Software for service-oriented systems. *Lecture Notes in Computer Science* (Springer, Berlin), Vol 3779, pp 2–13.
11. Couvares P, Kosar T, Roy A, Weber J, Wegner K (2007) Workflow in Condor. *Workflows for e-Science*, eds Taylor I, Deelman E, Gannon D, Shields M (Springer, New York), pp 357–375.
12. Sfiligoi I (2007) Making science in the Grid world: Using glideins to maximize scientific output. *Nuclear Science Symposium*, ed IEEE (IEEE), pp 1107–1109.
13. Zwart PH, et al. (2008) Automated structure solution with the PHENIX suite. *Method Mol Biol* 426:419–435.
14. Read RJ (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr A* 42:140–149.
15. Potterton L, et al. (2004) Developments in the CCP4 molecular-graphics project. *Acta Crystallogr D* 60:2288–2294.

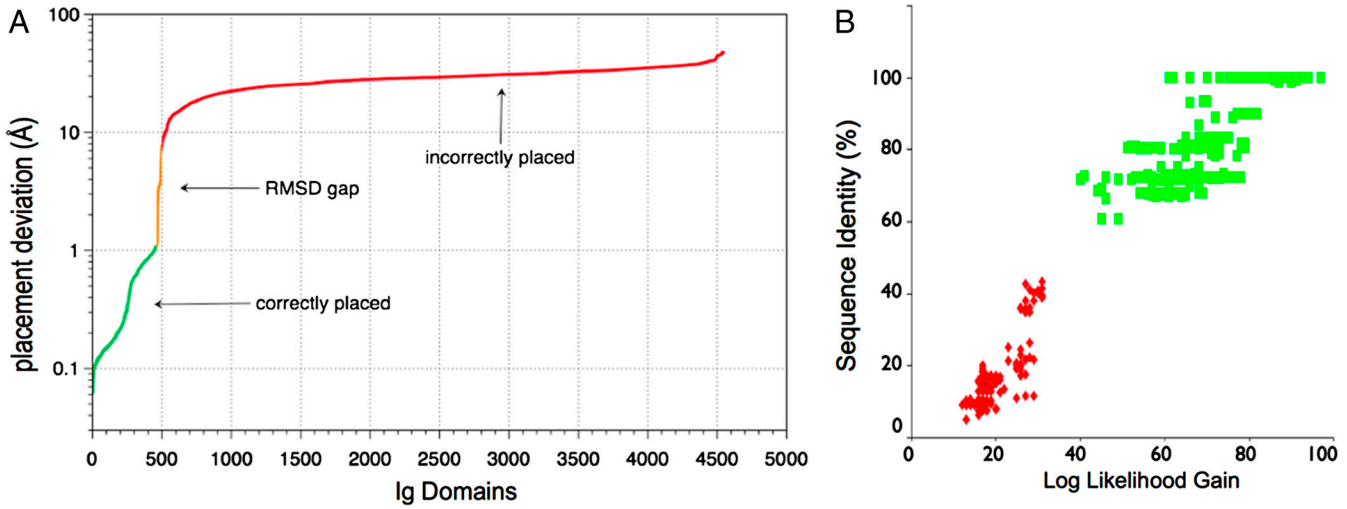

Fig. S1. MHC–TCR case (*A*)—Determination of placement quality. We take a placement quality <1.5 Å as correctly placed and >5 Å as incorrectly placed. The illustration is for the MHC–TCR complex second round WS-MR search for Ig domains, with the MHC–PBD from the first round fixed. The transition in placement 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).

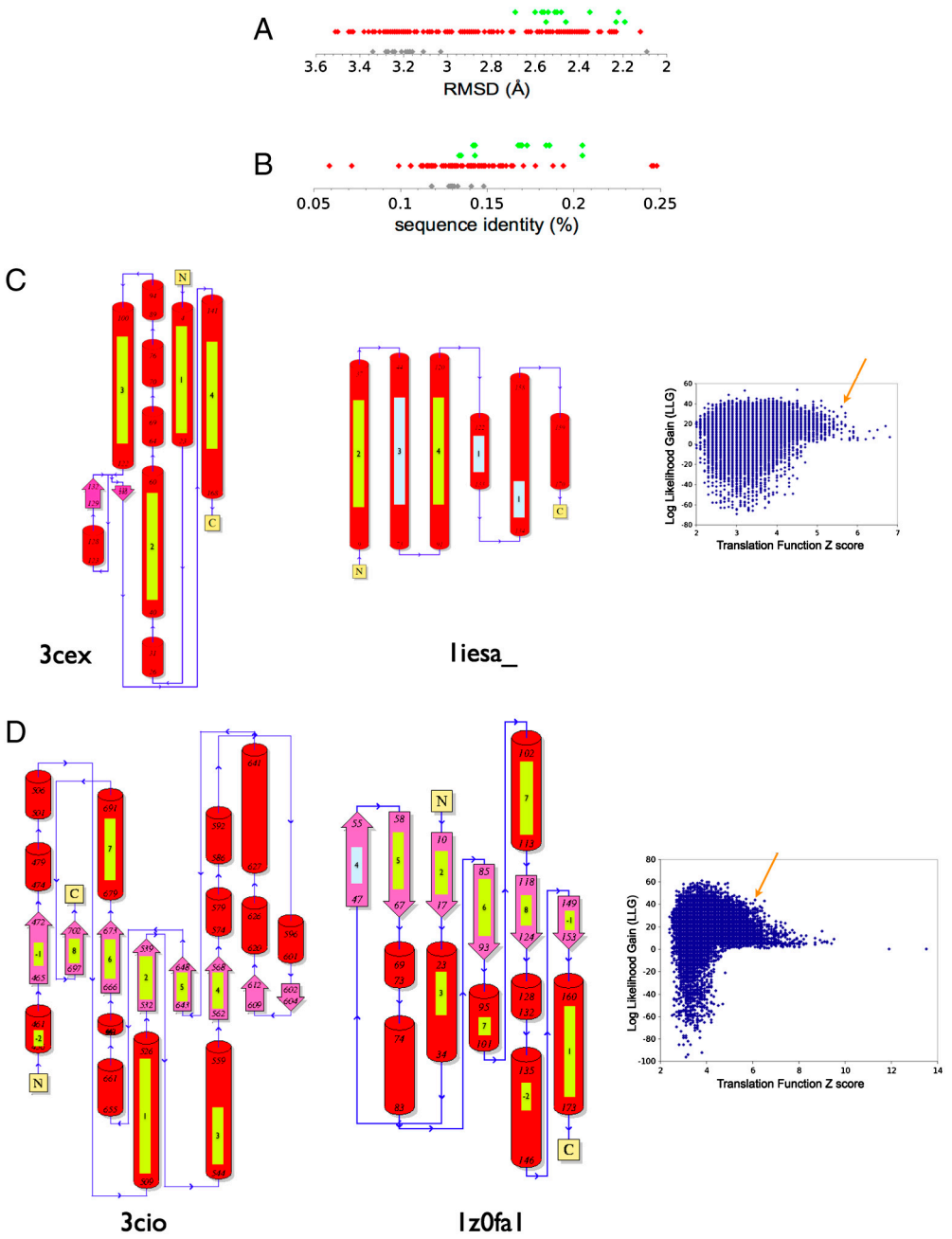

Fig. S2. Screening datasets of distant homologs. (*A*) TM-Align RMSD scores for all c.23.5 domains calculated against actual structure. Top green row represents distinct cluster of 14 correctly placed domains, lower green row represents 4 correctly placed domains with LLG/TFZ scores indistinguishable 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. 4*E*. (*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.

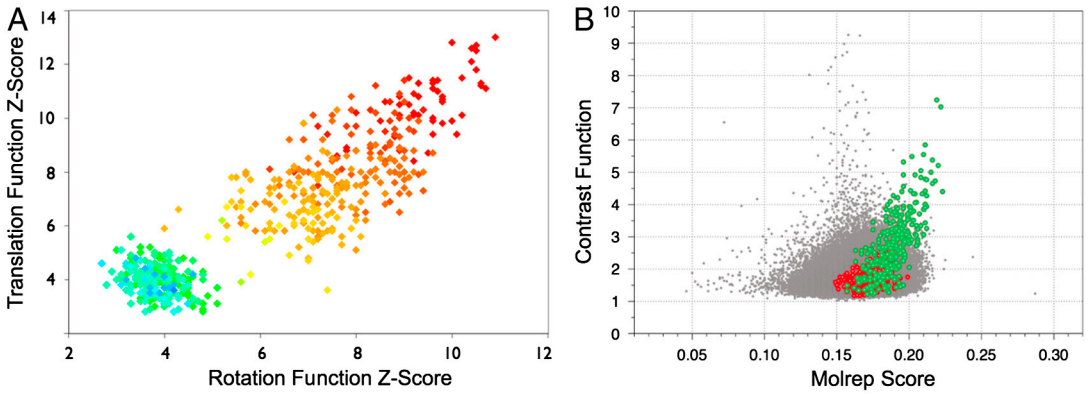

Fig. S3. Evaluation of global search scoring functions for PBD domain of MHC–TCR complex. (*A*) Translation function Z-score vs. rotation function Z-score with LLG heat map for Phaser results of MHC–TCR first round PBD domain search, indicating high correlation of TFZ, RFZ, and LLG in this case. (*B*) 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.

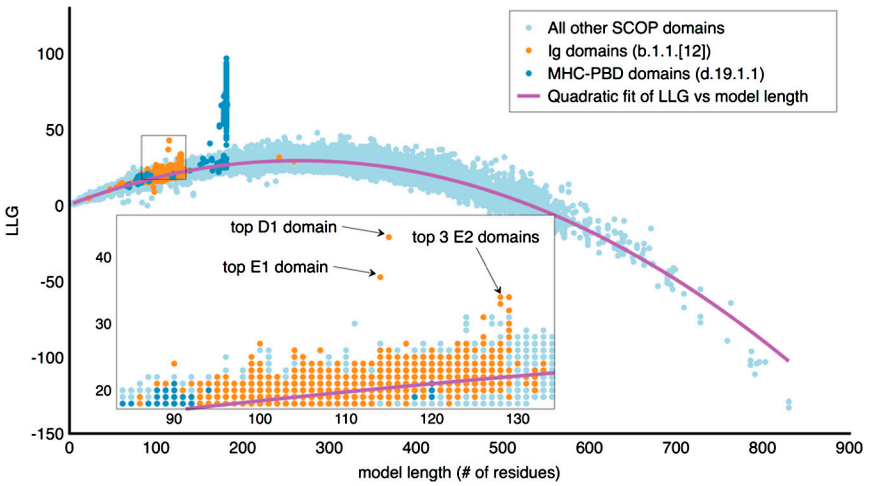

Fig. S4. LLG vs. model length for MHC–TCR complex, first round search. Dark blue points represent MHC–PBD models and cluster predominantly around 181 residues (the length of the actual domain). The correctly placed models are all in the upper vertical cluster. The orange points indicate the Ig domains, 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).

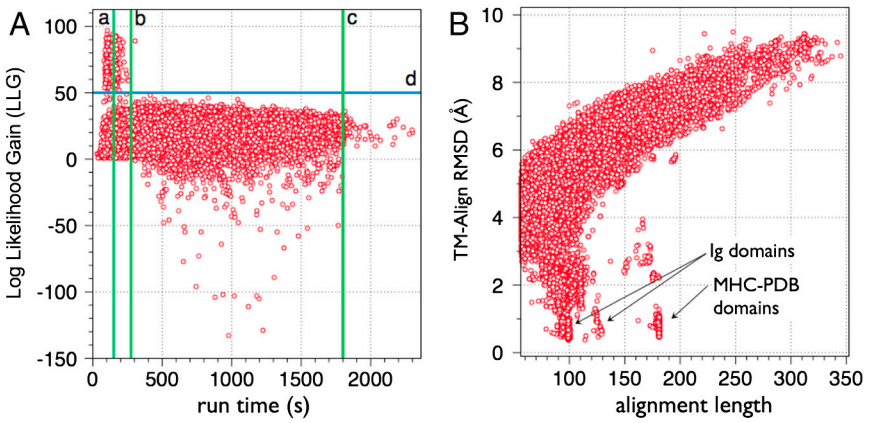

Fig. S5. MHC–TCR complex test case. (*A*) Phaser LLG vs. run-time. This illustrates the clustering of high LLG results all with run times less than 300 s. Timeout was set to 1,800 s (30 minutes). a: top 200 mean run-time (152 s); b: top 200 mean run-time + 5 standard deviations (283 s); c: timeout (1800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