LETTER

doi:10.1038/nature09964

Improved molecular replacement by density- and energy-guided protein structure optimization

Frank DiMaio[1], Thomas C. Terwilliger[2], Randy J. Read[3], Alexander Wlodawer[4], Gustav Oberdorfer[5], Ulrike Wagner[5], Eugene Valkov[6], Assaf Alon[7], Deborah Fass[7], Herbert L. Axelrod[8], Debanu Das[8], Sergey M. Vorobiev[9], Hideo Iwaï[10], P. Raj Pokkuluri[11] & David Baker[1]

Molecular replacement[1–4] procedures, which search for placements of a starting model within the crystallographic unit cell that best account for the measured diffraction amplitudes, followed by automatic chain tracing methods[5–8], have allowed the rapid solution of large numbers of protein crystal structures. Despite extensive work[9–14], molecular replacement or the subsequent rebuilding usually fail with more divergent starting models based on remote homologues with less than 30% sequence identity. Here we show that this limitation can be substantially reduced by combining algorithms for protein structure modelling with those developed for crystallographic structure determination. An approach integrating Rosetta structure modelling with Autobuild chain tracing yielded high-resolution structures for 8 of 13 X-ray diffraction data sets that could not be solved in the laboratories of expert crystallographers and that remained unsolved after application of an extensive array of alternative approaches. We estimate that the new method should allow rapid structure determination without experimental phase information for over half the cases where current methods fail, given diffraction data sets of better than 3.2 Å resolution, four or fewer copies in the asymmetric unit, and the availability of structures of homologous proteins with >20% sequence identity.

The limiting steps in molecular replacement are finding the correct location of the starting model in the unit cell and the interpretation of electron density maps produced using the imperfect phase information from candidate model placements. The left column of Fig. 1 illustrates the problem of initial model-building starting with distant comparative models (20–30% sequence identity) that have been correctly placed in the crystallographic unit cell. Automatic chain tracing methods fail on such maps because they often follow the incorrect comparative model (red) more closely than the actual structure (yellow); breaks in the density make it difficult to recover the correct backbone trace. Nevertheless, the maps contain considerable information about the native structure; for example, portions of the starting model that are not within density are generally incorrect.

Structure prediction methods such as Rosetta search for the lowest energy conformation of the polypeptide chain using physically realistic force fields. Based on previous work showing that accurate structures could be obtained from very sparse NMR data sets[15] by using the data to guide structure prediction searches, we reasoned that structure prediction methods guided by even very noisy density maps might be able to improve a poor molecular replacement model before applying crystallographic model-building techniques. We developed an approach in which electron density maps generated from molecular replacement solutions for each of a series of starting models are used to guide energy optimization by structure rebuilding, combinatorial side chain packing, and torsion space minimization[16]. New maps are generated using phase information from the energy-optimized models most consistent with the diffraction data, subjected to automatic chain tracing, and success is monitored through the free R factor[17].

To investigate the performance of the new method, we obtained 18 crystallographic data sets that had resisted previous attempts at structure determination. We first tested whether a comprehensive set of state-of-the-art molecular replacement approaches using a range of full-length and trimmed templates and homology models could solve any of these structures (Supplementary Information). We were able to solve five of the structures with both the new method and the existing methods (Table 1), leaving 13 challenging data sets highly resistant (Supplementary Information section 1) to structure determination (Table 1). For each of these, we identified homologous proteins of known structure[18] and constructed sequence alignments and starting models[9] from the five closest homologues. Starting models were used to search for up to five

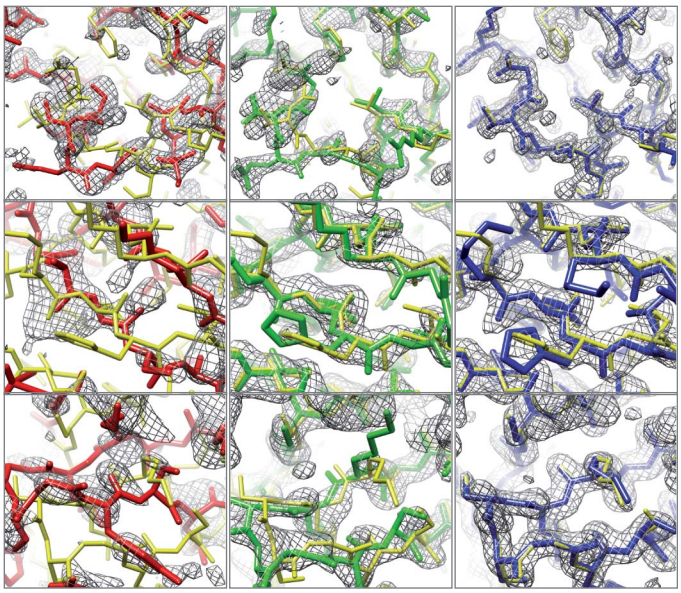

Figure 1 | Examples of improvement in electron density and model quality. Each row corresponds to one of the entries in Table 1. First row: 6 (2.0 Å resolution); second row: 7 (2.1 Å resolution); third row: 12 (1.7 Å resolution). Left column: correct initial molecular replacement solution (not necessarily identifiable at this stage) using starting model and corresponding density. Middle column: energy-optimized model and corresponding density. Right column: model and density following automatic building using the energy-optimized model as the source of phase information. The final deposited structure is shown in yellow in each panel; the initial model, energy-optimized model, and model after chain rebuilding are in red, green and blue, respectively. The sigma-A-weighted $2mF_o - DF_c$ density contoured at 1.5σ is shown in grey.

[1]University of Washington, Department of Biochemistry and HHMI, Seattle, Washington 98195, USA. [2]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [3]University of Cambridge, Department of Haematology, Cambridge Institute for Medical Research, Cambridge CB2 0XY, UK. [4]Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, Maryland 21702, USA. [5]Institute of Molecular Biosciences, University of Graz, Humboldtstrasse 50/3, 8010-Graz, Austria. [6]University of Cambridge, Department of Biochemistry, Cambridge CB2 1GA, UK. [7]Weizmann Institute of Science, Department of Structural Biology, Rehovot 76100, Israel. [8]Joint Center for Structural Genomics and SSRL, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA. [9]Northeast Structural Genomics Consortium, Columbia University, New York, New York 10027, USA. [10]University of Helsinki, Institute of Biotechnology, FI-00014 Helsinki, Finland. [11]Argonne National Laboratory, Biosciences Division, Argonne, Illinois 60439, USA.

Table 1 | Determination of previously unsolved structures using the new approach

ID number	Source*	Resolution (Å)	Seqid (%)	R_{free} after Phaser MR and model-building protocol							
				Autobuild	Arp/Warp	Simulated annealing (SA) + Autobuild	Torsion-space SA + Autobuild	Extreme SA + Autobuild	DEN + Autobuild	Rosetta + Autobuild	R_{free} (current best)
Solved by multiple methods											
1	JCSG	2.1	22	**0.31**	0.50	0.30	**0.30**	**0.30**†	**0.35**	**0.31**	0.22
2	NSGC	2.2	19	**0.29**	0.57	0.29	**0.29**	**0.29**†	**0.30**	**0.29**	0.22
3	UG	2.5	27	**0.34**	0.59	0.29	**0.29**	**0.29**†	**0.35**	**0.27**	0.19
4	JCSG	2.7	21	**0.31**	0.59	0.30	**0.30**	**0.30**†	**0.31**	**0.30**	0.24
5	ANL	1.9	31	0.51	0.59	0.54	0.54	**0.24**	**0.39**	**0.31**	0.24
Only solved by Rosetta											
Rosetta modelling with density required for successful model-building											
6	NCI	2.0	30	0.56	0.59	0.60	0.55	0.55	0.50	**0.34**	0.20
7	WI	2.1	22/15	0.56	0.60	0.54	0.54	0.54	0.56	**0.28**	0.26
8	JCSG	2.8	29	0.52	0.55	0.50	0.50	0.51	0.45	**0.36**	0.36‡
9	UC	3.0	22	0.54	0.56	0.50	0.50	0.47	0.46	**0.32**	0.25§
10	JCSG	3.2	20	0.54	0.57	0.51	0.51	0.53	0.46	**0.39**	0.33‡
11	UG	2.5	18	0.52	0.57	0.54	0.52	0.54	0.55	**0.27**	0.22
	MEAN			0.54	0.57	0.53	0.52	0.52	0.50	**0.33**	
Rosetta homology modelling required for successful molecular replacement											
12	BI,HY	1.7	– (100)	–	–	–	–	–	–	**0.29**	0.22
13‖	JCSG	2.9	29	–	–	–	–	–	–	**0.39**	0.23

The Seqid column gives the sequence identity to the closest homologue identified by HHpred[18], and is shown in parentheses if this is an NMR structure. The next seven columns give the R_{free} of the model produced by different combinations of refinement and autobuilding approaches. The final column gives the R_{free} after further refinement by the crystallographer who provided the data. For structures solved by multiple methods, the new method as well as one or more alternative approaches was sufficient ($R_{free} < 0.4$). In the first subset of structures that could only be solved by the new method (only solved by Rosetta), molecular replacement succeeds (in some cases ambiguously) using the template alone but model-building fails; in the second subset, refinement in Rosetta is required for molecular replacement to succeed. Targets that could not be solved by our approach are listed in Supplementary Table 1.

* JCSG, Joint Center for Structural Genomics; NSCG, Northeast Center for Structural Genomics; UG, University of Graz; ANL, Argonne National Lab; NCI, National Cancer Institute; WI, Weizmann Institute of Science; UC, University of Cambridge; BI,HY, Institute Of Biotechnology, University of Helsinki.

† Because a single SA trajectory was sufficient to solve these cases, Extreme SA was not run. Values from the single SA run are shown for completeness.

‡ Solutions for both are essentially correct based on the selenium positions in the anomalous difference Fourier maps calculated from the experimental data. However, structures are difficult to complete to deposition due to some MR solution model bias, poor or disordered density in numerous regions and low resolution.

§ Refinement ongoing.

‖ This structure was solved and all tests on this template were carried out using the intact template as a starting point. With this template both the molecular replacement step and subsequent rebuilding required Rosetta modelling for success. After determining the structure and completing the tests we found that it was also possible to solve the structure by molecular replacement if the template were split into two rigid subunits and the two domains were correctly chosen.

candidate molecular replacement solutions based on the likelihood of the experimental diffraction data[2]. Electron density maps were computed for each of these solutions, and used to guide energy minimization by first remodelling the unaligned regions and regions which poorly fit the density and then optimizing all backbone and side chain torsion angles. The likelihood of the experimental diffraction data was computed for each optimized model[2]; if top ranked models were similar (see Methods), a map generated from the highest likelihood model was subjected to automatic chain rebuilding, density modification and refinement[5]. If this succeeded in building the majority of the protein and produced a model with free *R* factor[17] significantly better than random ($R_{free} < 0.4$), the structure was considered solved; rebuilt models were further analysed by the crystallographers who supplied the original data. Using this approach, we were able to solve eight of the thirteen challenging cases (Table 1). In some of these eight cases, recognition of the correct placement of the model in the unit cell was only possible after Rosetta refinement (Supplementary Fig. 2); in others the correct placement was clear but the density was too poor for chain rebuilding. In two of the cases (12 and 13), even finding the correct molecular replacement solution first required energy-based refinement[12].

The improvement in electron density produced by density guided energy optimization and autobuilding are illustrated in Fig. 1. The starting molecular replacement models are often quite inaccurate, and the density generated from these models has breaks within the backbone of the actual structure (left panels). After model rebuilding and energy guided structure optimization, backbone breaks are largely closed and both side chains and backbone are more correctly modelled (middle panels). Automatic chain rebuilding into the improved map followed by density modification and reciprocal-space refinement further improve the model and the density (right panels). For all eight cases, the correlation between the final refined density and density from the original molecular replacement solutions is low, increases significantly after energy- and density-based structure optimization, and still further after automatic chain rebuilding (Supplementary Table 2).

For each of the eight challenging cases solved with the new method we also applied a battery of existing methods (Table 1 and Supplementary Information section 1) including simulated annealing in Cartesian and torsion space in PHENIX and CNS[14], deformable elastic network (DEN) refinement[13] in CNS, and PHENIX Autobuild[6] and ARP-WARP[5] for model-building. As noted above, in two cases Rosetta structure modelling was required for the correct placement of starting models in the unit cell, so the alternative methods could not even be applied. In the remaining six cases, final R_{free} values were lower using the new approach than with any of the existing methods (Table 1, Fig. 2a). Whereas conventional simulated annealing in both Cartesian and torsion space had little effect, the recently developed DEN[19] refinement protocol did improve three of the structures slightly, yielding free *R* values of 0.45–0.46 for these targets. Combination of DEN refinement with the method described here could lead to still more powerful approaches.

To benchmark the sequence and structural divergence where the different methods break down, we studied two different protein families for which a total of 59 different template structures covering a broad range of sequence and structural similarity were available (Supplementary Tables 3–5). Each template was correctly placed in the unit cell, and then improved with either Rosetta energy- and density-based optimization, Cartesian- and torsion-space simulated annealing, or DEN refinement. For each resulting model, the correlation with the density of the deposited structure was evaluated. Automatic chain rebuilding beginning with the superimposed starting models was successful for 18 of the 59 cases, consistent with the observation that molecular replacement often fails with templates sharing less than 30% sequence identity with the target sequence. Torsion-space simulated annealing in CNS before autobuilding allowed solution of two additional structures, DEN refinement, three additional structures, and Rosetta energy-based structure optimization, fourteen additional structures (Supplementary Fig. 2 and Supplementary Tables 3–5). We found the radius of convergence of the new method can be further extended by guiding energy based structure optimization by the Patterson correlation[20] rather than electron density (see Supplementary Information). This allowed structure improvement and identification of the correct molecular replacement solution in two additional cases (Supplementary Fig. 2, compare green

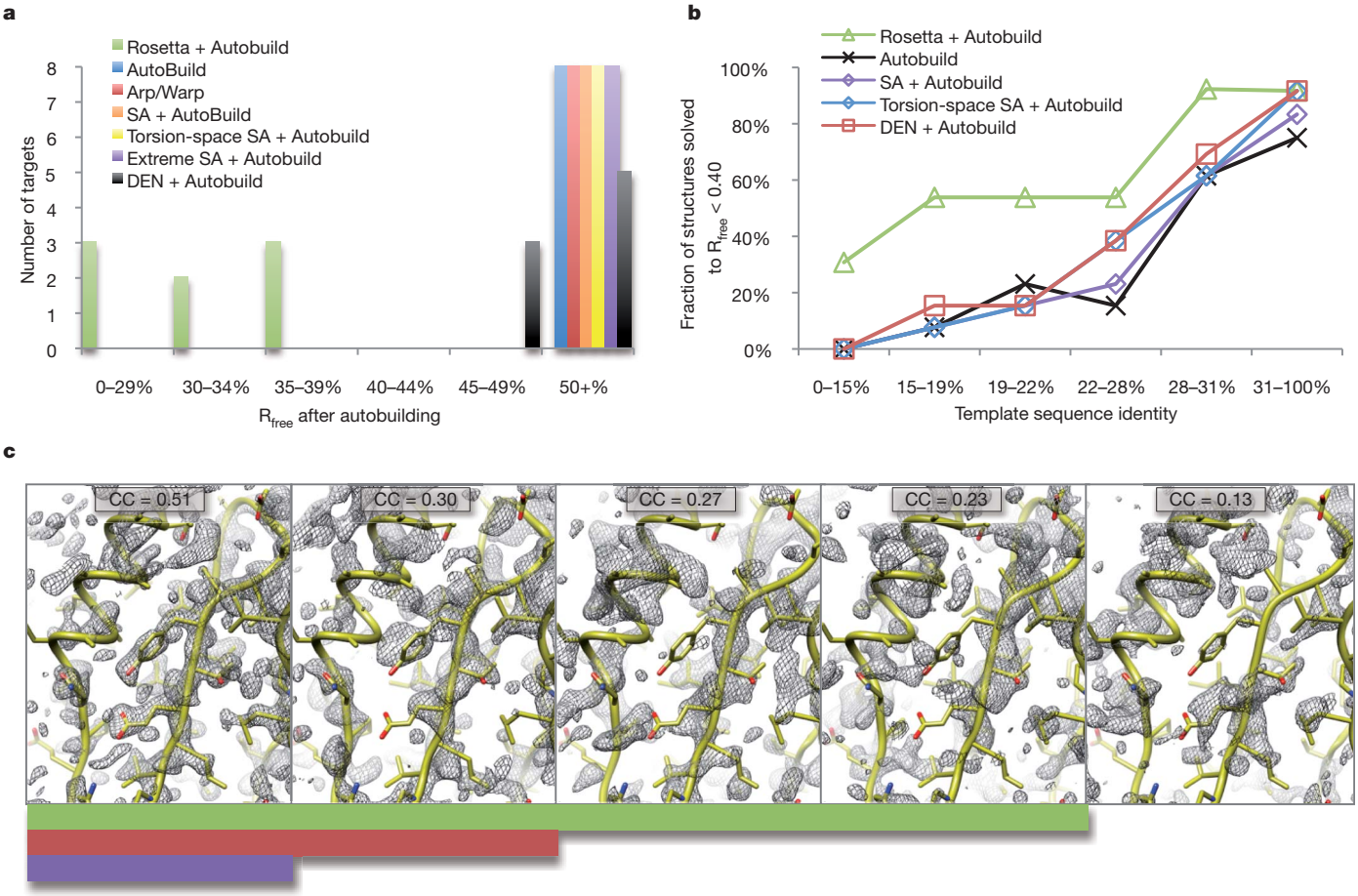

Figure 2 | Method comparison. **a**, Histogram of R_{free} values after autobuilding for the eight difficult blind cases solved using the new approach (Table 1). For most existing approaches, none of the cases yielded R_{free} values under 50%; DEN was able to reduce R_{free} to 45–49% for three of the structures. For all eight cases, Rosetta energy and density guided structure optimization led to R_{free} values under 40%. **b**, Dependence of success on sequence identity. The fraction of cases solved (R_{free} after autobuilding <40%) is shown as a function of template sequence identity over the 18 blind cases and 59 benchmark cases. The new method is a clear improvement below 28% sequence identity. **c**, Dependence of structure determination success on initial map quality. Sigma-A-weighted $2mF_o - DF_c$ density maps (contoured at 1.5σ) computed from benchmark set templates with divergence from the native structure increasing from left to right are shown in grey; the solved crystal structure is shown in yellow. The correlation with the native density is shown above each panel. The solid green bar indicates structures the new approach was able to solve ($R_{free} < 0.4$); the red bar those that torsion-space refinement or DEN refinement is able to solve, and the purple bar those that can be solved directly using the template.

to orange bar); for one of these the improvements were sufficient for autobuilding to effectively solve the structure.

Over the combined set of 18 blind cases and the 59 benchmark cases, Rosetta refinement yielded a model with density correlation as good or better than any of the control methods for all but six structures. The dependence of success on sequence identity over the combined set is illustrated in Fig. 2b. The improvement in performance is particularly striking below 22% sequence identity, where the quality of the starting homology models becomes too low for the control methods in almost all cases. With the new method the success rate in the 15–28% sequence identity range, generally considered very challenging for molecular replacement, is over 50%.

Figure 2c illustrates the dependence of model-building on the quality of initial electron density. Conventional chain rebuilding requires a map in which the connectivity is largely correct (leftmost panel), whereas the new method can tolerate breaks in the chain more than other methods (panels 2–4), as long as there is sufficient information in the electron density map, combined with the Rosetta energy function, to guide structure optimization. The map on the far right contains too little information to guide energy-based structure optimization and hence the new approach fails. In the five blind cases that have not yet been solved the comparative models may have been too low in quality, or there may have been complications in the X-ray diffraction data sets themselves.

Key to the success of the approach described here is the integration of structure prediction and crystallographic chain tracing and refinement methods. Simulated annealing guided by molecular force fields and diffraction data has had an important role in crystallographic refinement[14,21]. Structure prediction methods such as Rosetta can be even more powerful when combined with crystallographic data because the force fields incorporate additional contributions such as solvation energy and hydrogen bonding, and the sampling algorithms can build non-modelled portions of the molecule *de novo* and cover a larger region of conformational space than simulated annealing. The difference between Rosetta sampling and simulated annealing sampling, both using crystallographic data, is illustrated in Fig. 3. Beginning with the homology model placed by molecular replacement in the unit cell for blind case 6, we generated 100 models by simulated annealing at two starting temperatures, and 100 models with Rosetta energy- and density-guided optimization followed by refinement. The $2mF_o - DF_c$ (ref. 22) electron density maps generated using phases from over 50% of the Rosetta models had correlations 0.36 or better to the final refined map, whereas fewer than 5% of models from simulated annealing had correlations this high. Our approach probably outperforms even extreme simulated annealing because the physical chemistry and protein structural information which guide sampling eliminate the vast majority of non-physical conformations.

Approaches to molecular replacement combining the power of crystallographic map interpretation and structure prediction methodology

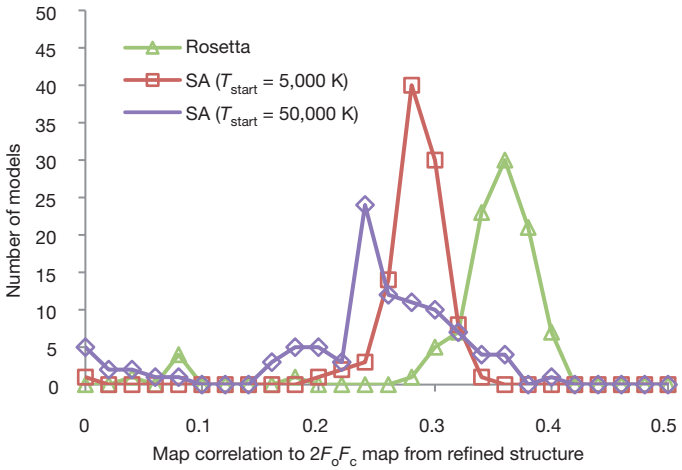

Figure 3 | Comparison of the effectiveness of model diversification using Rosetta and simulated annealing. For blind case 6, 100 models were generated using either simulated annealing with a start temperature of 5,000 K, simulated annealing with a start temperature of 50,000 K, or Rosetta energy- and density-guided optimization. The correlation between $2mF_o - DF_c$ density maps computed from each structure and the final refined density was then computed; the starting model has a correlation of 0.29 and the distributions of the refined models are shown in the figure. Rosetta models have correlations better than the initial model much more often than simulated annealing.

are likely to become increasingly useful in the next few years. First, the number of already-determined structures will continue increasing, making it increasingly likely that there will be a structure with the required >20% sequence identity: the chance there is a structure with a sequence identity of 20% or greater is more than twice that of finding a structure with at least 30% sequence identity[23]. Second, as more work focuses on proteins that cannot be expressed in *Escherichia coli*, the currently preferred methods for experimental phase determination based on selenomethionine replacement may be more difficult to apply. Finally, as protein structure modelling algorithms improve, better initial models should further increase the radius of convergence of the approach.

METHODS SUMMARY

Starting models (templates) for molecular replacement were generated by searching the PDB using HHpred[18] for proteins likely to have structures related to the query. Starting models were constructed from alignments generated by HHpred. Unaligned residues were removed from the template and non-identical side chains were stripped back to the gamma carbon (CG), as suggested in previous work[9]. An initial Phaser search with a low rotation function cutoff (50%) and modest packing threshold (up to 10 clashes) was used to find up to five putative molecular replacement (MR) solutions for each template. Each MR solution for each template was used to obtain an initial estimate of phases and the corresponding sigma-A-weighted $2mF_o - DF_c$ density map was generated[22]. Gaps in the initial alignment, as well as regions around deletions, were rebuilt using the Rosetta loop modelling protocol[12], which alternates insertion of short fragments with similar local sequences and cyclic coordinate descent (CCD) closure[24]. Twenty-four rounds of side chain rotamer optimization and side chain and backbone torsion-space minimization were then used to optimize a linear combination of the Rosetta all-atom energy and a term assessing agreement to the electron density. Following the energy- and density-guided refinement, models were ranked based on the Phaser log-likelihood score. The highest ranked models were then subjected to a second round of modelling using the Rosetta iterative rebuild and refine protocol[12] constrained by density. After this final round of refinement, the model with best agreement to the experimental data (highest likelihood) was used to either find additional models in the asymmetric unit, or as a starting point for Phenix AutoBuild.

The procedures described here require considerable computation as up to several thousand Rosetta models are generated for each structure, typically requiring 0.5–1 h per structure of CPU time. We have developed automated procedures in Phenix (*phenix.mr_rosetta*) that use Rosetta and Phenix modules to carry out and extend many of the methods described here with density modification and density averaging, potentially allowing fewer Rosetta models to be used. All the methods described in this paper are available in release 3.2 of Rosetta.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 August 2010; accepted 22 February 2011.
Published online 1 May 2011.

1. Rossmann, M. G. *The Molecular Replacement Method* (Gordon & Breach, 1972).
2. McCoy, A. J. *et al. Phaser* crystallographic software. *J. Appl. Cryst.* **40,** 658–674 (2007).
3. Brünger, A. T. *et al. Crystallography & NMR system*: a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
4. Vagin, A. & Teplyakov, A. *MOLREP*: an automated program for molecular replacement. *J. Appl. Cryst.* **30,** 1022–1025 (1997).
5. Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols* **3,** 1171–1179 (2008).
6. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density modification with the *PHENIX AutoBuild* wizard. *Acta Crystallogr. D* **64,** 61–69 (2008).
7. DePristo, M. A., de Bakker, P. I. W., Johnson, R. J. K. & Blundell, T. L. Crystallographic refinement by knowledge-based exploration of complex energy landscapes. *Structure* **13,** 1311–1319 (2005).
8. Cowtan, K. The *Buccaneer* software for automated model building. *Acta Crystallogr. D* **62,** 1002–1011 (2006).
9. Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. The importance of alignment accuracy for molecular replacement. *Acta Crystallogr. D* **60,** 1229–1236 (2004).
10. Rodríguez, D. D. *et al.* Crystallographic *ab initio* protein structure solution below atomic resolution. *Nature Methods* **6,** 651–653 (2009).
11. Suhre, K. & Sanejouand, Y. H. On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr. D* **60,** 796–799 (2004).
12. Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450,** 259–264 (2007).
13. Schröder, G., Levitt, M. & Brünger, A. T. Super-resolution biomolecular crystallography with low-resolution data. *Nature* **464,** 1218–1222 (2010).
14. Brünger, A. T., Kuriyan, J. & Karplus, M. Crystallographic *R* factor refinement by molecular dynamics. *Science* **235,** 458–460 (1987).
15. Raman, S. *et al.* NMR structure determination for larger proteins using backbone-only data. *Science* **327,** 1014–1018 (2010).
16. Das, R. & Baker, D. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).
17. Brünger, A. T. Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355,** 472–475 (1992).
18. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21,** 951–960 (2005).
19. Schröder, G. F., Brunger, A. T. & Levitt, M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15,** 1630–1641 (2007).
20. Brünger, A. T. Extension of molecular replacement: a new search strategy based on Patterson correlation refinement. *Acta Crystallogr. A* **46,** 46–57 (1990).
21. Brünger, A. T., Karplus, M. & Petsko, G. A. Crystallographic refinement by simulated annealing: application to crambin. *Acta Crystallogr. A* **45,** 50–61 (1989).
22. Read, R. J. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42,** 140–149 (1986).
23. Vitkup, D., Melamud, E., Moult, J. & Sander, C. Completeness in structural genomics. *Nature Struct. Biol.* **8,** 559–566 (2001).
24. Canutescu, A. & Dunbrack, R. Cyclic coordinate descent: a new algorithm for loop closure in protein modeling. *Protein Sci.* **12,** 963–972 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements R.J.R., T.C.T. and D.B. thank the NIH (5R01GM092802), the Wellcome Trust (R.J.R.), and HHMI (D.B.) for funding this research. F.D. acknowledges the NIH (P41RR002250) and HHMI. D.F. and A.A. acknowledge support from the Israel Science Foundation. G.O. thanks DK Molecular Enzymology (FWF-project W901) and the Austrian Science Fund (FWF-project P19858). The work of A.W. was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. H.I. acknowledges support from the academy of Finland (1131413). S.M.V. was supported by a grant from the Protein Structure Initiative of National Institute of General Medical Sciences (U54 GM074958). The work of P.R.P. at Argonne National Laboratory was supported by the US Department of Energy's Office of Science, Biological and Environmental Research GTL programme under contract DE-AC02-06CH11357. We thank all members of the JCSG for their general contributions to the protein production and structural work. The JCSG is supported by the NIH, National Institutes of General Medical Sciences, Protein Structure Initiative (U54 GM094586 and GM074898).

Author Contributions F.D., T.C.T., R.J.R. and D.B. developed the methods described in the manuscript; F.D., T.C.T., R.J.R., A.W. and D.B wrote the paper. A.W., G.O., U.W., E.V., A.A., D.F., H.L.A., D.D., S.M.V., H.I. and P.R.P. provided the data and refined one or more structures to completion.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to T.C.T. (terwilliger@lanl.gov) or D.B. (dabaker@u.washington.edu).

METHODS

Preparation of templates and identification of initial molecular replacement solutions. For the application of the new method to blind cases, templates were identified using HHpred[18]. For both the blind and benchmark data sets, HHpred was used to generate initial alignments. We prepared templates by removing all unaligned residues and stripping all non-identical side chains to the gamma carbon (CG), as suggested in previous work[9]. An initial Phaser search with a low rotation function cutoff (50%) and modest packing threshold (up to 10 clashes) was used to find up to five putative MR solutions for each template. In two blind cases (12 and 13 in Table 1), Phaser was unable to locate the correct configuration of a molecule using the template alone, but modelling in Rosetta without density-fitting constraints before Phaser search enabled discovery of the correct rigid-body placement of the molecule, with very low Phaser translation function *Z*-scores (TFZ) of 4–6 (after solving 13, it was discovered that breaking the template into two rigid subunits enabled solution of the molecular replacement problem). If point-group symmetry was present in the templates, the initial search (and subsequent steps) were carried out both with monomeric and multimeric models (see subsection on symmetric modelling into density below).

Rebuilding and refinement into density. Each MR solution for each template was used to obtain an initial estimate of phases and the corresponding sigma-A-weighted $2mF_o - DF_c$ density map was generated[22]. Gaps in the initial alignment, as well as regions around deletions, were rebuilt using the Rosetta loop modelling protocol[12], which alternates insertion of short fragments with similar local sequences and CCD closure[24]. Twenty-four rounds of side chain rotamer optimization and side chain and backbone torsion-space minimization were then used to optimize a linear combination of the Rosetta all-atom energy and a term assessing agreement to the electron density. Agreement to density was computed using an extension of a method previously developed for building into cryo-electron microscopy density[25]. Density was calculated from a model using a single-Gaussian approximation to atomic scattering factors. Correlation coefficients between model and map were calculated for each residue: the computed density includes all atoms in the residue and the backbone in the two flanking residues on each side, and the correlation is taken over a mask extending 5 Å from each atom. Scores are proportional to the negative log probability that observed correlations occur by random chance, assuming a normal distribution; parameters are trained matching randomly oriented fragments into synthesized density. In all cases, density was truncated at 3 Å.

Following the energy- and density-guided refinement, models were ranked based on the Phaser log-likelihood score. The highest ranked models were then subjected to a second round of modelling using the Rosetta iterative rebuild and refine protocol[12] constrained by density. Regions that deviated the most from the current estimate of the electron density were rebuilt; clashes between crystallographic (and non-crystallographic) contacts were also always rebuilt. For each template carried over to the second round (typically the top-scoring 3–10 models from the previous round), 2,000 Rosetta models were generated. The likelihood of the diffraction data was again computed using Phaser for the lowest-energy 10% of models, and if the five highest likelihood models were in the same rigid-body configuration (that is if they had density correlations above 0.2 with each other), they were used to re-phase the density and an additional round (24 cycles) of side chain optimization and refinement was carried out in Rosetta. If the top-scoring models differed, then additional templates were considered (if available) or Rosetta homology modelling was used to perturb the initial structures before molecular replacement.

After this final round of refinement, the model with best agreement to the experimental data (highest likelihood) was used to either find additional models in the asymmetric unit, or as a starting point for Phenix AutoBuild. In cases where the R_{free} was better than random but higher than 0.4, and a majority of residues were placed, additional refinement was carried out using models produced by AutoBuild, which allows for recovery from sequence alignment errors. The bond lengths and bond angles were first replaced with ideal values with small compensating changes in the torsion angles to minimize the change in interatomic distances, and the idealized models were then subjected to 48 cycles of side chain rotamer optimization and side chain and backbone torsion minimization. In the first 24 cycles, the Rosetta all-atom energy function was optimized, and in the final 24 cycles a weighted sum of Rosetta all-atom energy and the fit-to-density energy described above was optimized.

Refinement of symmetric complexes into density. Key to solving many of the blind cases was proper treatment of symmetry. In cases where there is point-group symmetry in the asymmetric unit (either from the template or subsequently discovered by molecular replacement search) or there is close contact between crystal partners, the Rosetta symmetric modelling framework[26] was used to reduce the size of the conformational space which must be searched. This occurred in blind cases where either there was point-group symmetry in the template(s) (6 in Table 1), point-group symmetry was found during the Phaser search (13), or tight crystal contacts formed point-group symmetry (8 and 10). In these cases, Rosetta optimizes only the torsion angles in one subunit and the rigid-body degrees of freedom of the corresponding symmetric group. The energy is calculated explicitly over a non-redundant subset of atoms for computational efficiency, but the fit to density is calculated without symmetrization. This is similar to the "strict formulation" of symmetry introduced in ref. 27.

Symmetric modelling in Rosetta requires that the energy of a symmetric complex be expressible in terms of a single subunit or as pairwise interactions between this subunit and other ones. Minimization also only considers gradients from these components. To take advantage of Rosetta's symmetric modelling with asymmetric density data, the gradients of each subunit with respect to the fit-to-density energy must be mapped to a single subunit. The score of a residue *i*'s fit to density is just the sum of the fit-to-density scores over all of *i*'s copies. As a first approximation, the gradient at *i* can be computed as the combined gradients of all of *i*'s copies, rotated by the symmetry operation to rotate the subunit containing *i*'s copy to the one containing *i*. Unfortunately, although this approach correctly handles gradients of internal torsions, the gradients at each symmetric degree-of-freedom are not correctly handled. Proper handling takes advantage of the formulation from ref. 28 to efficiently convert Cartesian gradients to torsion-space gradients. For each atom in the symmetric complex, we compute F_1 and F_2 corresponding to the unrotated gradient with respect to the fit-to-density score. For internal torsional degrees of freedom, the rotation applied to each F_1/F_2 just maps each subunit back to the asymmetric unit. At each symmetric degree of freedom we apply a corresponding symmetry operation; for example, in D3 symmetry (a dimer of trimers) the degree of freedom corresponding to the "spin" of the trimers applies the rotation used to transform between trimers to all the F_1/F_2's in one of the trimers.

Refinement against the Patterson function. In benchmark cases where the Phaser translation search failed to find the correct molecular placement even when many potential solutions were considered, we conducted refinement against the Patterson function. A score function was implemented that assessed the correlation between the computed and experimental Patterson map (next paragraph). The map was truncated to between 3.5 Å and 10 Å resolution (in reciprocal space) and 5 Å to ~75% of the template diameter (in real space). Starting models used the same templates and rebuilding procedure as the density refinement. Because the correct rotation is not known at this stage, the molecule orientation was randomized at the beginning of each refinement trajectory and constraints on backbone atoms were used to prevent the molecule from rotating more than ~5° from this starting orientation.

The scoring function we optimize is the weighted sum of Rosetta's all-atom potential function and the correlation between the calculated Patterson map and the observed Patterson map. To make this tractable in Rosetta refinement, which may require tens of thousands of score-function evaluations per trajectory, simplifications are necessary. Directly computing $\partial p_{calc}/\partial x$ requires three fast Fourier transforms (FFTs) per atom. However, since what is needed is not $\partial p_{calc}/\partial x$ but instead the sum $\partial \sum_{map} p_{calc}p_{obs}/\partial x$, FFTs can be used to compute the change in correlation at every position in the map at once (where p is the Patterson density and ρ is the real-space density):

$$\frac{\partial \sum_{map} p_{calc}p_{obs}/\partial x}{\partial x} = F^{-1}[F[p_{obs}]\cdot F[\rho_{calc}]\cdot[F[\partial \rho_{calc}/\partial x_i]](x) \quad (1)$$

Assuming a fixed B-factor over the molecule, this requires just 3 FFTs per atom type (the correction terms that make this not just the overlap integral but a true correlation can be folded into the same FFT). Then, given a model to refine against the Patterson map, we compute equation (1) once, sum over all the symmetric orientations of the space group, and interpolate the gradient at each atom's position. Given sufficiently fine sampling, this gives a very close approximation to the true derivative in a small fraction of the CPU time.

For side chain optimization, where we must rescore the Patterson correlation for exponentially many combinations of side chain rotamers, exact computation is also intractable. However, first computing the density ρ_{calc} of the backbone only, then computing the correlation scores for each side chain rotamer independently, provides a reasonably good approximation with only several hundred to several thousand function evaluations (one for each rotamer).

Torsion space simulated annealing with DEN restraints. As a control, we ran torsion-space simulated annealing with DEN restraints[13] on the blind tests and on the complete benchmark set of structures related to PDB entries 1XVQ and 1A2B. Using the same template and placement used by Rosetta refinement, initial homology models were built in Modeller[29] (using the same alignment used by Rosetta). DEN refinements were carried out using the refine_lowres.inp script distributed

with CNS version 1.3 as a template. The results of these analyses for the benchmark set of structures are shown in Supplementary Tables 4 and 5, and for the blind tests, as part of Table 1.

Massive-sampling simulated annealing. To test the role that massive sampling around the conformation of the input structure plays in the success of our new methods, we developed an 'extreme simulated annealing protocol', where 1,000 models were produced by simulated annealing refinement, the best of these models is used as the starting point for automated model rebuilding, density modification and refinement with PHENIX, and the resulting model is used as the starting point for a second iteration of the procedure. In this procedure, simulated annealing was carried out in phenix.refine using the flag 'simulated_annealing = True' and the default starting temperature of 5,000 K.

Implementation in Phenix and Rosetta. The procedures described here require considerable computation as up to several thousand Rosetta models are generated for each structure, typically requiring 0.5–1 h per structure of CPU time. We have developed automated procedures in Phenix (phenix.mr_rosetta) that use Rosetta and Phenix modules to carry out and extend many of the methods described here with density modification and density averaging, potentially allowing fewer Rosetta models to be used. Beginning with correctly placed templates (including all copies of each molecule, and placed domains for 13), each of 13 blind test cases in Table 1 can be solved with phenix.mr_rosetta using 20 Rosetta models during each rebuilding cycle, yielding free R values of 0.42 or lower (mean $R_{free} = 0.33$), and requiring from approximately 30 to 130 CPU-hours to complete.

All the methods described in this paper are available in release 3.2 of Rosetta. An application, 'mr_protocols,' is included which was used (together with Phaser and Phenix Autobuild) to generate all the results in this paper. The flags files used for Rosetta are shown below.

Comparative modelling (with target sequence target.fasta, alignment target_template.ali, and template template.pdb) in the context of density:

```
-database $DB
-MR:mode cm
-in:file:extended_pose 1
-in:file:fasta target.fasta
-in:file:alignment target_template.ali
-in:file:template_pdb template.pdb
-loops:frag_sizes 9 3 1
-loops:frag_files aa1xxx_09_05.200_v1_3.gz aa1xxx_03_05.200_v1_3.gz none
-loops:random_order
-loops:random_grow_loops_by 5
-loops:extended
-loops:remodel quick_ccd
-loops:relax relax
-relax:default_repeats 4
-relax:jump_move true
-edensity:mapreso 3.0
-edensity:grid_spacing 1.5
-edensity:mapfile target.map
-edensity:sliding_window_wt 1.0
-edensity:sliding_window 5
-cm:aln_format grishin
-MR:max_gaplength_to_model 10
-nstruct $STRUCTS
```

In cases where Rosetta was used to 'pre-refine' the structure before Phaser, the same command line was used without the -edensity:* flags. Modelling with symmetry used the flags above in addition to the flag '-symmetry_definition symm.def', where symm.def defines the symmetry in the template. Symmetry definition file creation is automated using a script; see the Rosetta documentation for more details.

Additional refinement (both after comparative modelling and after autobuilding in some cases):

```
-database $DB
-MR:mode relax
-in:file:s rosetta_model.pdb
-relax:default_repeats 4
-relax:jump_move true
-edensity:mapreso 3.0
-edensity:grid_spacing 1.5
-edensity:mapfile target.map
-edensity:sliding_window_wt 1.0
-edensity:sliding_window 5
-nstruct 5
```

Comparative modelling against the Patterson function (the experimental Patterson map, target_pat.map, is computed outside Rosetta):

```
-MR:mode cm
-in:file:extended_pose 1
-in:file:fasta target.fasta
-in:file:alignment target_template.ali
-in:file:template_pdb template.pdb
-loops:frag_sizes 9 3 1
-loops:frag_files aa1xxx_09_05.200_v1_3.gz aa1xxx_03_05.200_v1_3.gz none
-loops:random_order
-loops:random_grow_loops_by 5
-loops:extended
-loops:remodel quick_ccd
-loops:relax relax
-relax:default_repeats 2
-relax:jump_move true
-edensity:grid_spacing 1.6
-edensity:mapfile target_pat.map
-edensity:use_spline_interpolation true
-edensity:realign random
-edensity:use_symm_in_pcalc true
-edensity:patterson_lowres_limit 3.5
-edensity:patterson_hires_limit 10.0
-edensity:patterson_minR 5.0
-edensity:patterson_maxR 14.0
-edensity:patterson_B 0.2
-edensity:patterson_cc_wt 0.5
-cm:loop_rebuild_filter 500
-cm:aln_format grishin
-cm:max_loop_rebuild 10
-cm:min_loop_size 4
-MR:max_gaplength_to_model 10
-nstruct $STRUCTS
```

Most of the data used in this paper is available at http://www.phenix-online.org/phenix_data/terwilliger/rosetta_2011/ (additional blind cases will be made available as the structures are deposited).

25. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of protein structures into low-resolution density maps using Rosetta. *J. Mol. Biol.* **392,** 181–190 (2009).
26. André, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proc. Natl Acad. Sci. USA* **104,** 17656–17661 (2007).
27. Weis, W. I., Brünger, A. T., Skehel, J. J. & Wiley, D. D. Refinement of the influenza virus hemagglutinin by simulated annealing. *J. Mol. Biol.* **212,** 737–761 (1990).
28. Abe, H., Braun, W., Noguti, T. & Gō, N. Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins general recurrent equations. *Comput. Chem.* **8,** 239–247 (1984).
29. Eswar, N. *et al.* Comparative protein structure modeling with MODELLER. *Curr. Protoc. Bioinform.* (Suppl.) **15,** 5.6 doi:10.1002/0471250953.bi0506s15 (2006).