REPORT

PROTEIN STRUCTURE

Protein structure determination using metagenome sequence data

Sergey Ovchinnikov,[1,2,3] Hahnbeom Park,[1,2] Neha Varghese,[4] Po-Ssu Huang,[1,2] Georgios A. Pavlopoulos,[4] David E. Kim,[1,5] Hetunandan Kamisetty,[6] Nikos C. Kyrpides,[4,7] David Baker[1,2,5]*

Despite decades of work by structural biologists, there are still ~5200 protein families with unknown structure outside the range of comparative modeling. We show that Rosetta structure prediction guided by residue-residue contacts inferred from evolutionary information can accurately model proteins that belong to large families and that metagenome sequence data more than triple the number of protein families with sufficient sequences for accurate modeling. We then integrate metagenome data, contact-based structure matching, and Rosetta structure calculations to generate models for 614 protein families with currently unknown structures; 206 are membrane proteins and 137 have folds not represented in the Protein Data Bank. This approach provides the representative models for large protein families originally envisioned as the goal of the Protein Structure Initiative at a fraction of the cost.

There are 14,849 protein families in the Pfam (*1*) database with 50 or more residues, of which 4752 have at least one member with experimentally determined x-ray crystal or nuclear magnetic resonance (NMR) structure, and an additional 3984, for which reliable comparative models can be built on the basis of homologs of known structure detected using the powerful HHsearch fold-recognition program (*2*). There are an additional 902 for which less-confident comparative models can be built, but no structural information available for 5211 of the remaining 6113 families (HHsearch E-value ≥ 1). Until recently, computational methods could not generate accurate models for these 5211 families, as they lack homologs of known structure for comparative modeling, and the very large number of conformations accessible to a polypeptide chain made the sampling problem in de novo protein structure prediction intractable for all but the smallest proteins. The original goal of the Protein Structure Initiative was to determine structures for at least one representative of such families, but this proved to be extremely challenging, and the focus of the initiative shifted to targets of immediate biological interest (*3*).

The increase in the number of known amino acid sequences has enabled the accurate prediction of residue-residue contacts by using evolutionary data (*4*–*10*)—substitutions at positions close in space in the three-dimensional structure covary. Such contact predictions have been used for a wide range of protein modeling efforts (*11*–*22*). Accurate contact prediction requires large numbers of aligned sequences so that residue-residue covariance is clearly distinguished from lineage effects. Although coevolution-based structure modeling has been used to generate models for individual proteins with fold-level accuracy [template modeling (TM) score (*23*) is >0.5 (*5*, *7*, *8*, *10*, *11*, *14*–*18*, *21*, *22*)], it has not been clear whether such data, combined with structure-prediction methodology, can generate accurate models on a larger scale.

Rosetta de novo structure-prediction calculations guided by evolutionary information were recently used to generate models for 58 large protein families (*21*). The structures of proteins in six of these families have since been published, which provides an opportunity to assess this medium-scale prediction effort. Recently solved structures of the lipoprotein signal peptidase II (*24*), prolipoprotein diacylglyceryl transferase (*25*), fluoride ion transporter (*26*), cytochrome bd oxidase (*27*), DMT superfamily transporter YddG (*28*), and fumarate hydratase (*29*) are all very close to computational models published and publicly released well before the structures were solved (Fig. 1). In the case of the three-subunit cytochrome bd oxidase, the computational model of the 788-residue complex generated using both inter- and intra-subunit contact information was used together with experimental phase information obtained from the three heme irons and a single methionine to solve the structure. Because the phase information was weak, it was only possible to place the transmembrane helices and a subset of the side chains on the basis of the density, but the loops, connectivity, location of the CydX subunit, and registration of the amino acid sequence on many of the helices were unclear. Our *Escherichia coli* protein model closely overlapped with the traced helices, and Phenix-Rosetta refinement (*30*) of a model built for the *Geobacillus thermodenitrificans* protein resolved the above ambiguities, enabling rapid completion of structure determination. The final deposited structure is very similar to our previously published model of the *E. coli* protein (Fig. 1A) [TM-align score (*23*) of 0.8]. The power of Rosetta structure-prediction calculations coupled with coevolution data for soluble proteins is illustrated by an extremely accurate blind de novo prediction for a complex protein structure in the CASP11 structure-prediction experiment (*31*) (Fig. 1E). In all of the cases shown in Fig. 1, standard threading or fold-recognition methods fail to identify the correct fold. Taken together, these data show that Rosetta modeling guided by coevolutionary constraints generates accurate models (in all six cases, the TM-align score is >0.7; the models also illustrate some of the limitations of the approach, including the lack of explicit modeling of ligands, cofactors, and lipids) (see supplementary text).

Structure models with the accuracy of those in Fig. 1 would have broad utility for framing biological hypotheses about function and interpreting mutational data, as well as for guiding experimental structure determination. To determine the number of aligned sequences required for contact prediction accuracy sufficient to guide generation of accurate 3D models, we carried out Rosetta structure-prediction calculations for a benchmark set of 27 large protein families (table S1) with known structure. We used both the full sequence alignments and alignments of subsets of the sequences for contact prediction. We also performed structure-prediction calculations using Rosetta to hybridize and refine (*32*) partial structural matches identified by matching predicted contacts with the contact patterns of known protein structures. To do this, we developed an algorithm (map_align) [see the supplementary materials (SM)] that uses iterative double-dynamic programming (*33*). The two approaches are complementary: De novo structure prediction (using only sequence information) (*34*) can succeed where there are no related structures in the Protein Data Bank (PDB), whereas making use of matches to known structures can help for large complex proteins that otherwise present a convergence challenge for de novo structure prediction (structural matches can occur in the absence of detectable sequence similarity because structural similarity is retained over larger evolutionary distances). For large sequence families, combining de novo structure-prediction models and map_align structure matches using the Rosetta iterative hybridization protocol improved accuracy in 14 cases and decreased accuracy in only one (solid line in Fig. 2A) (fig. S1; see SM). Contact prediction accuracy, and hence predicted structure accuracy, depends on the number of sequences in the family, the diversity of these sequences, and the length of the protein. A measure that incorporates all three factors [N_f, the

[1]Department of Biochemistry, University of Washington, Seattle, WA 98105, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98105, USA. [3]Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195, USA. [4]Joint Genome Institute, Walnut Creek, CA 94598, USA. [5]Howard Hughes Medical Institute, University of Washington, Box 357370, Seattle, WA 98105, USA. [6]Facebook Inc., Seattle, WA 98109, USA. [7]Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia.
*Corresponding author. Email: dabaker@u.washington.edu

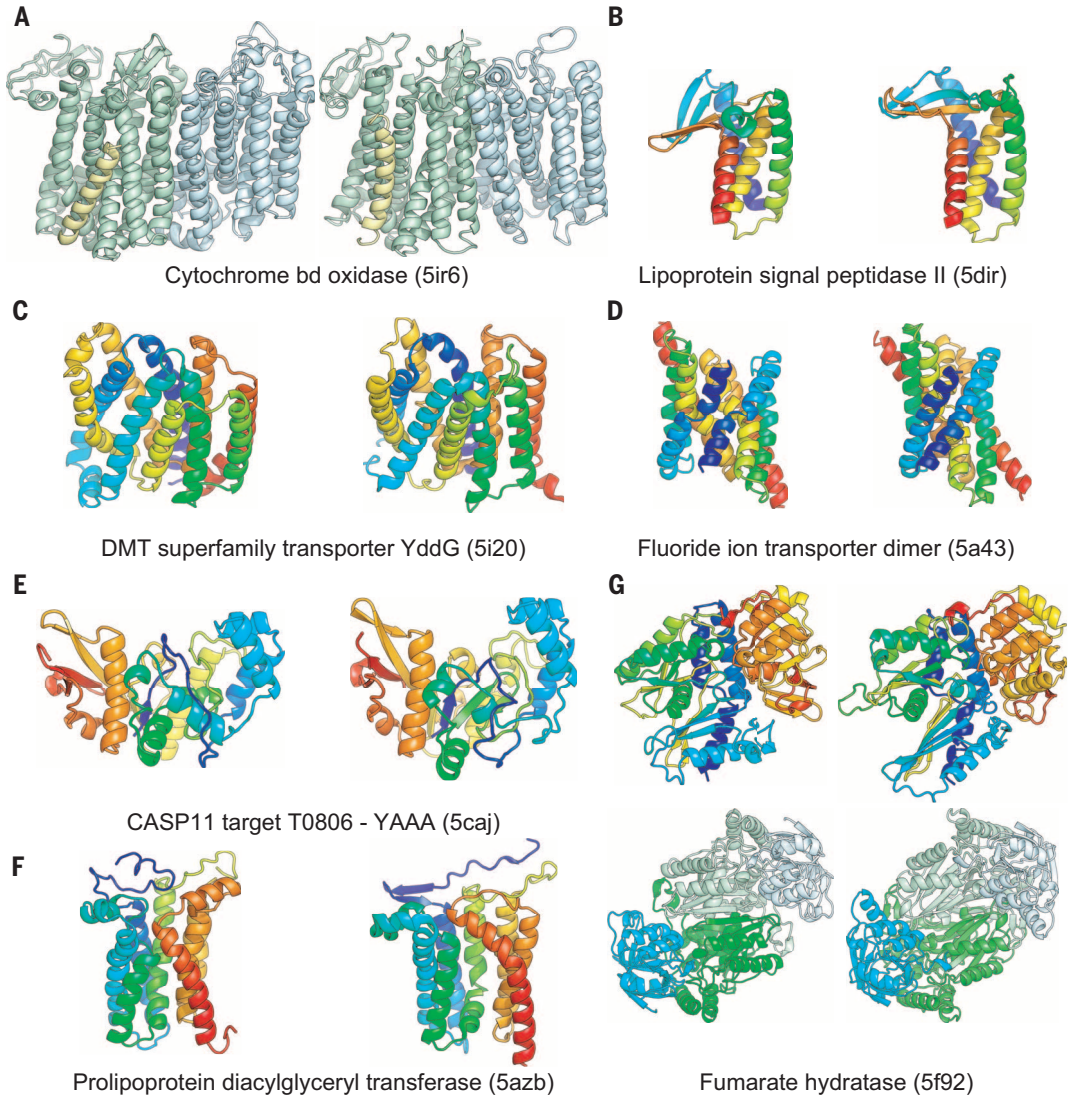

Fig. 1. Comparison of Rosetta models (left) to subsequently published crystal structures (right). The models accurately recapitulate the structural details of the named proteins. The scores are as follows: (**A**) the cytochrome bd oxidase (TM-align score 0.88), (**B**) the lipoprotein signal peptidase II (TM-align score 0.70), (**C**) the DMT superfamily transporter YddG (TM-align score 0.70), (**D**) the fluoride ion transporter dimer (TM-align score 0.69), (**E**) the CASP11 target T0806, (**F**) prolipoprotein diacylglyceryl transferase (TM-align score 0.69), and (**G**) fumarate hydratase [TM-align score 0.80 for monomer (top) and 0.76 for dimer (bottom)].

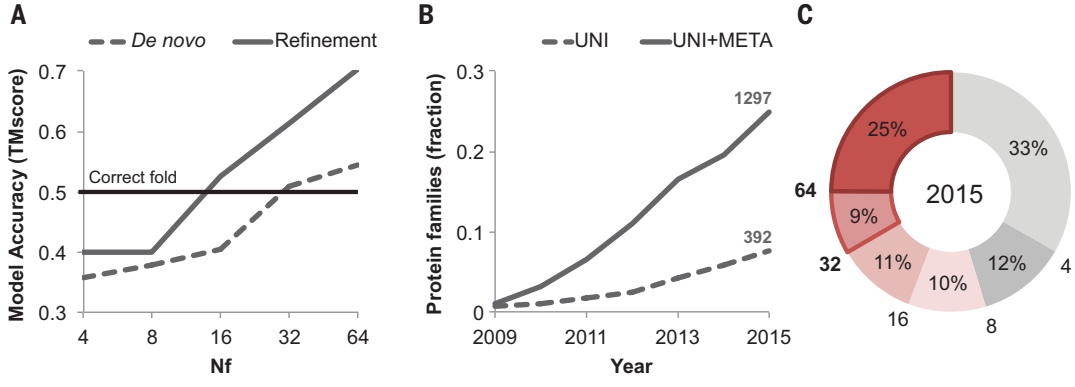

Fig. 2. Metagenome data greatly increased fraction of structures that can be accurately modeled. (**A**) Dependence of coevolution guided Rosetta structure-prediction accuracy on the effective number of sequences N_f (a function of both sequence number and diversity; see methods definition) in the protein family. For each of 27 proteins of known structure, the multiple sequence alignment was subsampled, and residue-residue contacts were predicted by using GREMLIN. Rosetta structure-prediction calculations were then used to generate ~20,000 models, and a single model was selected on the basis of the Rosetta energy and the fit to the coevolution constraints; the average TM score of these selected models over all 27 cases is shown on the *y* axis (dashed line). Hybridization-based refinement of the top 20 models together with the top 10 map_align-based models for each case increases the average accuracy (solid line); models with fold-level accuracy (TM score of >0.5) are obtained for $N_f \geq 16$, and models with accuracy typical of comparative modeling, for N_f of 64. (**B**) Fraction of protein families of unknown structure with at least 64 N_f. Dashed line: including only sequences in UniRef100 database; solid line: including sequences in UniRef100 database together with metagenome sequence data from the Joint Genome Institute (*37*). (**C**) Distribution of N_f values for 5211 Pfam families with currently unknown structure, after the addition of metagenomic sequences; 25% of the protein families have $N_f > 64$, 34% have $N_f > 32$, and 45% have $N_f > 16$.

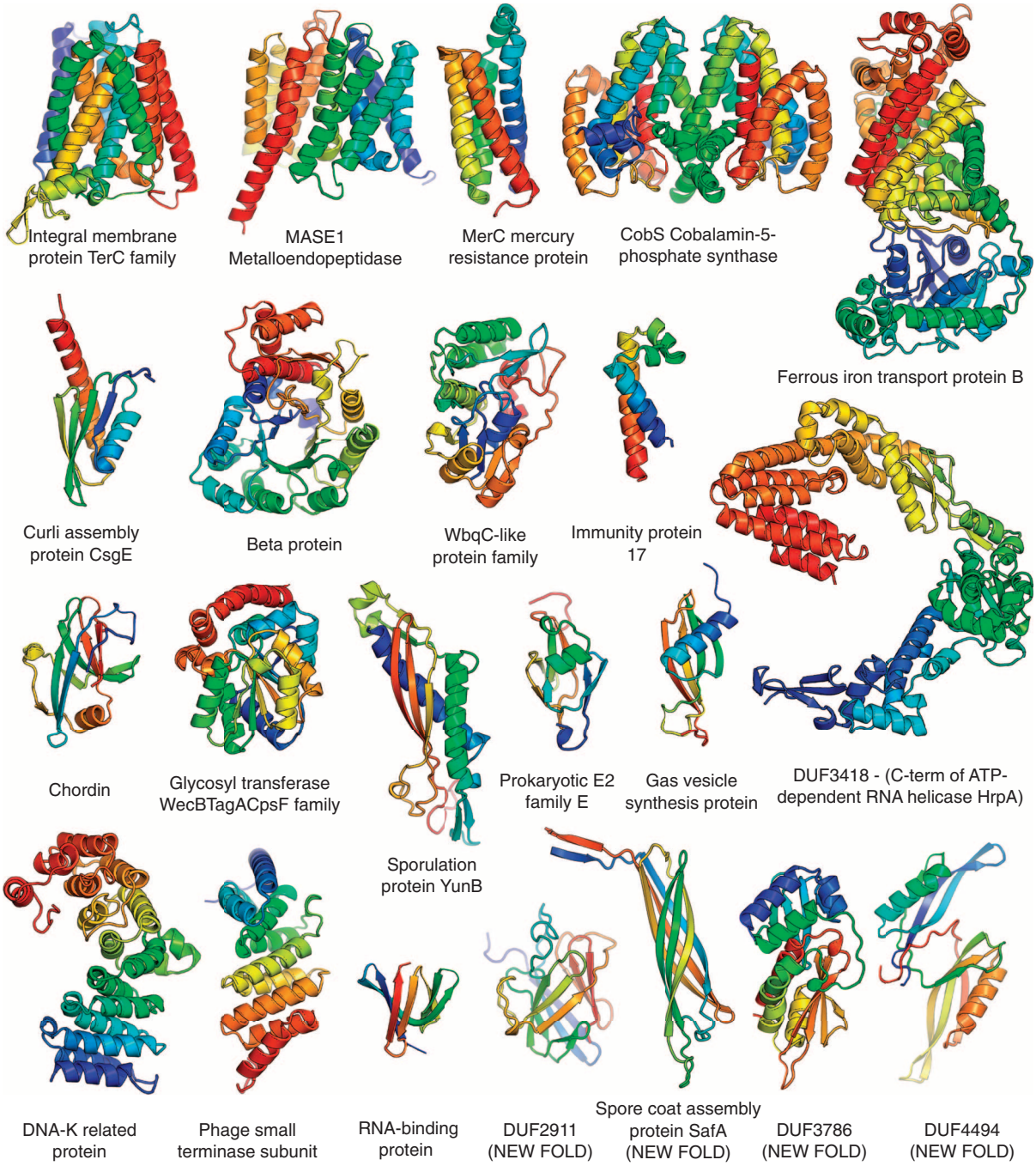

Fig. 3. Representative structure models for selected Pfam families. Membrane proteins are on the top row; new folds on the bottom right. The multidomain models of the iron transporter and RNA helicase and the dimeric model of CobS, an enzyme in vitamin B synthesis, are guided by both intra- and inter-chain coevolution restraints.

number of sequence clusters at an 80% sequence identity–clustering threshold divided by the square root of the protein length (*21*)] correlates well with contact prediction accuracy (*21*) and model accuracy (Fig. 2A and fig. S1) over a broad range of families.

How many protein families with currently unknown structure have N_f values in the range where accurate models can be built? The models in Fig. 1 were all generated for families with $N_f > 64$; accuracy falls off for lower values of N_f (Fig. 2A). As shown in Fig. 2B, fewer than 8% of families have N_f values of 64 or better. Modeling the remaining 92% of families of unknown structure at reasonable accuracy is not currently possible by using the sequence information in the UniRef100 database (*35*).

This limitation in structure modeling can be largely overcome by taking advantage of progress in a completely different research area. Metagenome sequencing projects, in which complex biological samples are shotgun sequenced, have provided insights into biological communities and provide a treasure trove of new sequence data (*36*, *37*). The number of protein sequences determined in metagenome sequence projects is growing considerably faster than the UniRef100 database (solid versus dashed line in Fig. 2B). With the inclusion of metagenome sequence data, the number of sequences increases by as much

as 100-fold for some families (table S2), and the fraction of families with unknown structure that can be accurately modeled using coevolution-guided structure-prediction methods increases dramatically. At $N_f \geq 64$, the fraction increases from 0.08 to 0.25, and at $N_f \geq 32$ [where fold level accuracy can be achieved (Fig. 2A)], the fraction increases from 0.16 to 0.33. To assess structure-prediction and model evaluation accuracy using metagenome data, we carried out a second set of benchmark calculations on 81 Pfam domains with recently solved structures and $N_f \geq 64$ (fig. S1, E and F, and table S5). Structure-prediction accuracy was correlated with the extent of convergence of the lowest energy models and the fraction of predicted contacts present in these models (figs. S1F and S2). For 42 families, the predictions converged with most of the predicted contacts satisfied (see SM for convergence criteria) and of these, 25 had a TM score >0.7 and 13 a TM score >0.6 [in three of the four remaining cases, NMR structures of small transmembrane proteins, our models fit the predicted contacts much better, and in the last case, an intertwined dimer, our monomer model contained all the correct contacts (fig. S13)].

We generated coevolution based contact predictions using GREMLIN (*4*, *12*) for the 1297 protein families with $N_f \geq 64$ and built models for the 921 protein families (1024 domains) with many contacts between positions separated by more than five residues along the linear sequence (number of long range contacts > half the number of residues in protein). The structure-prediction calculations converged on models with predicted TM scores (based on the benchmark calculations) greater than 0.65 for 614 of the 1024 domains. A list of the Pfam families covered by these models is in table S3; the models are available at http://gremlin.bakerlab.org/meta/, along with an interactive 3D interface powered by 3Dmol.js (*38*) and D3.js (*39*) for visualization of coevolution contacts on the models. These structures provide close templates for comparative modeling of 487,306 UniRef100 and 3,868,268 Integrated Microbial Genomes metagenomic unique (less than 80% pairwise identity) sequences.

The converged models for the 614 Pfam families (table S3) provide a view of the hitherto unseen protein universe. To determine whether the models belong to known protein folds, we carried out structure-structure comparisons against the Structural Classification of Proteins (SCOP) (*40*) domain database. For 477 of the families, the models matched a protein of known structure over nearly the entire length and, hence, can be assigned to SCOP folds (52 distinct all alpha, 29 alpha/beta, 51 alpha+beta, and 28 all-beta folds). In a number of cases, the SCOP classifications are consistent with previous functional information; for example, the restriction endonuclease Xho I is assigned to the restriction enzyme fold, and a family of prokaryotic putative ubiquitin-like proteins is assigned the beta-grasp fold (to which ubiquitin belongs). For 137 of the domains, there were no significant structure matches of the models to the PDB (TM-align score < 0.5), and hence, these have new folds. Space limitations preclude showing here even a small number of the 614 models; instead, we show a small selection of the 3D structures in Fig. 3. They include the key developmental regulator Chordin; a key enzyme in cobalbumin synthesis; a metalloendopeptidase; and mercury and iron transporters. Six are transmembrane proteins, four have new folds, and several have complex topologies. These and the remaining 590 structure models not shown in Fig. 3 should provide a basis for understanding molecular function and mechanisms and should guide experimental structure determination (such efforts should be informed of the limitations of the modeling approach described in the supplementary text). While this manuscript was in preparation, crystal structures of members of 5 of the 614 families were published and are similar to the corresponding models (TM-align score ≥ 0.7) (see fig. S3 and table S4).

The models presented in this paper fill in about 12% of the structural information missing for known protein families. That this could be accomplished using computational modeling methods was not at all apparent 5 years ago. This progress required integration of advances in disparate research areas: metagenome sequencing, coevolutionary analysis, and de novo protein structure-prediction methodology. This combined approach has a bright future: Extrapolating from the data in Fig. 2B suggests that in several years the majority of families will have sufficient number of sequences for accurate structure modeling. A current limitation is that most sequence data are for prokaryotes, but as fungal and other simple eukaryote genome structure prediction sequencing projects ramp up, the approach should become applicable to eukaryote specific protein families.

REFERENCES AND NOTES

1. R. D. Finn *et al.*, *Nucleic Acids Res.* **44** (D1), D279–D285 (2016).
2. J. Söding, *Bioinformatics* **21**, 951–960 (2005).
3. G. T. Montelione, *F1000 Biol. Rep.* **4**, 7 (2012).
4. H. Kamisetty, S. Ovchinnikov, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
5. D. S. Marks *et al.*, *PLOS ONE* **6**, e28766 (2011).
6. F. Morcos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
7. T. A. Hopf *et al.*, *Cell* **149**, 1607–1621 (2012).
8. T. Nugent, D. T. Jones, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E1540–E1547 (2012).
9. D. T. Jones, D. W. Buchan, D. Cozzetto, M. Pontil, *Bioinformatics* **28**, 184–190 (2012).
10. D. S. Marks, T. A. Hopf, C. Sander, *Nat. Biotechnol.* **30**, 1072–1080 (2012).
11. J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, J. N. Onuchic, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 10340–10345 (2012).
12. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S. I. Lee, C. J. Langmead, *Proteins* **79**, 1061–1078 (2011).
13. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **87**, 012707 (2013).
14. S. Wickles *et al.*, *eLife* **3**, e03035 (2014).
15. P. Tian *et al.*, *J. Am. Chem. Soc.* **137**, 22–25 (2015).
16. S. Hayat, C. Sander, D. S. Marks, A. Elofsson, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5413–5418 (2015).
17. T. A. Hopf *et al.*, *Nat. Commun.* **6**, 6077 (2015).
18. L. A. Abriata, *Biorxiv* 10.1101/013581 (2015).
19. S. Ovchinnikov, H. Kamisetty, D. Baker, *eLife* **3**, e02030 (2014).
20. T. A. Hopf *et al.*, *eLife* **3**, (2014).
21. S. Ovchinnikov *et al.*, *eLife* **4**, e09248 (2015).
22. S. Antala, S. Ovchinnikov, H. Kamisetty, D. Baker, R. E. Dempski, *J. Biol. Chem.* **290**, 17796–17805 (2015).
23. Y. Zhang, J. Skolnick, *Proteins* **57**, 702–710 (2004).
24. L. Vogeley *et al.*, *Science* **351**, 876–880 (2016).
25. G. Mao *et al.*, *Nat. Commun.* **7**, 10198 (2016).
26. R. B. Stockbridge *et al.*, *Nature* **525**, 548–551 (2015).
27. S. Safarian *et al.*, *Science* **352**, 583–586 (2016).
28. H. Tsuchiya *et al.*, *Nature* **534**, 417–420 (2016).
29. P. R. Feliciano, C. L. Drennan, M. C. Nonato, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9804–9809 (2016).
30. F. DiMaio *et al.*, *Nat. Methods* **10**, 1102–1104 (2013).
31. S. Ovchinnikov *et al.*, *Proteins* **84** (suppl. 1), 67–75 (2016).
32. Y. Song *et al.*, *Structure* **21**, 1735–1742 (2013).
33. W. R. Taylor, *Protein Sci.* **8**, 654–665 (1999).
34. K. T. Simons *et al.*, *Proteins* **34**, 82–95 (1999).
35. B. E. Suzek *et al.*, *Bioinformatics* **31**, 926–932 (2015).
36. V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, P. Hugenholtz, *Microbiol. Mol. Biol. Rev.* **72**, 557–578 (2008).
37. V. M. Markowitz *et al.*, *Nucleic Acids Res.* **42** (D1), D568–D573 (2014).
38. N. Rego, D. Koes, *Bioinformatics* **31**, 1322–1324 (2015).
39. M. Bostock, V. Ogievetsky, J. Heer, *IEEE Trans. Vis. Comput. Graph.* **17**, 2301–2309 (2011).
40. A. Andreeva *et al.*, *Nucleic Acids Res.* **36** (Database), D419–D425 (2008).

ACKNOWLEDGMENTS

We thank P. Di Lena, N. Malod-Dognin, and R. Andonov for providing the source code for their software (AI-eigen and a_purva) and for their discussion and advice on contact map alignment. The 3D structures of 614 Pfam domains modeled in the study are available at http://gremlin.bakerlab.org/meta/. Other data are archived at the Dryad Digital Repository (doi:10.5061/dryad.27p4s). We also thank Rosetta@home and Charity engine participants for donating their computer time. The work performed by N.V., G.A.P., and N.C.K. was supported by the U.S. Department of Energy (DOE) Joint Genome Institute, a DOE Office of Science User Facility, under contract no. DE-AC02-05CH11231. Research reported here was supported by National Institute of General Medical Sciences, NIH, under award number R01GM092802. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/355/6322/294/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S13
Tables S1 to S5
References (*41*–*57*)

22 June 2016; accepted 22 November 2016
10.1126/science.aah4043