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 Sergey Ovchinnikov,1,2,3 Hahnbeom Park,1,2

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 homology 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 6133 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 achievable 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 each family, 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 lipopolysaccharide signal peptide (24), prolipoprotein diacylglyceryl transferase (25), chloride ion transporter (26), cytochrome b d 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 has been established to characterize the structure prediction potential for the Rosetta method (37) (eq. 1).

Downloaded from http://science.sciencemag.org/on January 20, 2017
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 diacylglycerol 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.
number of sequence clusters at an 80% sequence identity–clustering threshold divided by the square root of the protein length \( (\sqrt{L}) \) correlates well with contact prediction accuracy \( (\alpha) \) 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. Meta-genome 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

---

**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.
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 $N_r \geq 64$, the fraction increases from 0.08 to 0.25, and at $N_r \geq 32$ [where fold level accuracy can be achieved (Fig. 2A)], the fraction increases from 0.16 to 0.33. To assess level accuracy can be achieved (Fig. 2A), the by 3Dmol.js (available at http://gremlin.bakerlab.org/meta/, by these models is in table S3; the models are domains. A list of the Pfam families covered calculations) greater than 0.65 for 614 of the 1024 of residues in protein). The structure-prediction (number of long range contacts > half the number of contacts present in these models (figs. S1F and 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 $\geq 0.7$) (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**


**ACKNOWLEDGMENTS**

We thank P. Di Lena, N. Malod-Dognin, and R. Andonov for providing the source code for their software (Al-eigen and a_purva) and for their discussion and advice on contact map alignment. The 3D structures of 487 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.27p4j). We also thank Rosette/Channel 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.

**SUPPLEMENTAL MATERIALS**

www.sciencemag.org/content/355/6322/294/suppl/DC1

Materials and Methods

Supplementary Text

Figs. S1 to S5

Tables S1 to S5

References (41–57)

22 June 2016; accepted 22 November 2016

10.1126/science.aa4043
Editor's Summary

**Filling in the protein fold picture**

Fewer than a third of the 14,849 known protein families have at least one member with an experimentally determined structure. This leaves more than 5000 protein families with no structural information. Protein modeling using residue-residue contacts inferred from evolutionary data has been successful in modeling unknown structures, but it requires large numbers of aligned sequences. Ovchinnikov et al. augmented such sequence alignments with metagenome sequence data (see the Perspective by Söding). They determined the number of sequences required to allow modeling, developed criteria for model quality, and, where possible, improved modeling by matching predicted contacts to known structures. Their method predicted quality structural models for 614 protein families, of which about 140 represent newly discovered protein folds.

*Science*, this issue p. 294; see also p. 248