Protein structure determination using evolutionary information

Sergey Ovchinnikov

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

David Baker, Chair

Philip Bradley

Frank DiMaio

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Protein structure determination using evolutionary information

Sergey Ovchinnikov

Chair of the Supervisory Committee:
Professor David Baker
Biochemistry

For billions of years, nature has been conducting the greatest experiment of all time. Imagine one day gaining access to the detailed notes from these experiments. Today, with worldwide expeditions to collect samples from all habitats, single-cellular sequencing of unculturable microbes and rapid drop in sequencing cost, we can finally tap into nature and gain access to these notes. Natural selection acts upon a gene to optimize its sequence to perform a task. For protein-coding genes, the task includes folding, stability, and function. The record of the evolutionary process, which in itself is probabilistic, is contained within a multiple sequence alignment. A statistical model that accurately describes these evolutionary constraints for a given gene or a set of genes, should allow for the inference of the underlying physical molecular structure and interactions.

Recently, it was shown that a global statistical model of a protein family that captures both conservation and coevolution patterns in the family, to possess such quality. The strength of co-evolution term is correlated with residue-residue contacts in three-dimensional space. This means that not only can this information be used to predict contacts in proteins that have no structure, but also to better understand contacts in known structures. To assess the utility and the limitation of the method, we applied our method (GREMLIN) to predict residue-residue level interactions between proteins and within a protein. These contacts were used to predict the structure of 58 protein families and complexes. Nine of these structures have since been determined with traditional experimental methods and were found to be quite accurate. Most recently we extended the approach to small protein families by recruiting metagenomic sequences. Using this approach, we provided the predictions for over 600 protein families (>10% of PFAM).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would first like to acknowledge my undergraduate advisors Mark Fishbein and Susan Masta from PSU (Portland State University) for giving me the opportunity to volunteer in their labs and introducing me to science and phylogenetics. Fishbein lab: Kevin Weitemier, Kate Halpin, Mike Wilder, Margaret Parks, Diane Bland, Basma Saadoun and David Chuba. Masta lab: Erin Brandt, Spencer Smith, Shahan Derkarabetian, Kori Quatermass, Matthew Mulhern and Jason Bazzano. I would also like to acknowledge the Ronald E. McNair program at PSU for their support.

At University of Washington, my acknowledgement goes out to: MCB and the Biochemistry program and the labs that I rotated in: Ram Samudrala, David Baker and Jesse Bloom. I would also like to thank my committee members: Harmit Malik, Phil Bradley, Bill Noble and Ram Samudrala.

At the Baker lab: Hetunandan Kamisetty, Christof Angermüller, Wayne Mao, Vanessa Grey and Ivan Anishchenko for their co-evolution support. David Kim, Frank DiMaio, Ray Wang, James Thompson, Yifan Song, Chris Miles, Hahnbeom Park, Lei Shi, TJ Brunette, Kenneth Jung and Firas Kahatib for their protein structure prediction support. David La, Lance Stewart, Darwin Alonso, Possu Huang and the rest of the Baker lab for their expert advice!

MCB friends: Kwaku Opoku, Shannon Newman, Becky Scholz, Qing Feng, Zoi Villasana, Rob Lawrence, Carissa Pilling, Jackie Lang, Joe and Monica Sanchez, Jennifer Whiddon, Kris Blair, Katie Hooper, Nicole Iranon and Paul Biswajit! My roommate Lucia Huang.

Collaborators: Yi Shang, Robert Dempski, Lisa Kinch, Nick Grishin, Neha Varghese, Georgios Pavlopoulos, Nikos Krypides and Ryan Pavlovicz. Also Schara Safarian, Hartmut Michel, Damian Ekiert, Gira Bhabha, Stefan Schoebel, Tom Rapoport and Maofu Liao for sharing their experimental data. And of course the CASP community!

My Portland family: Inna, Leonid, Viktor, Slava, Sveta and Pavel Ovchinnikov.

My Seattle family: Qing Feng, Edgar and Allen.

And everybody else that I may have forgotten!

DEDICATION

I would like to dedicate the following thesis to the refugees and immigrants that truly make this world great. We are one people, one genetic code and one earth. Let us never allow the artificial country borders and walls to separate us and make us think we are different.

Chapter 1. INTRODUCTION

1.1 EARLY SEQUENCE ANALYSIS

Before 1955, it was widely thought that proteins were amorphous molecules. When Fredrick Sanger determined the first protein sequence of insulin (Ryle et al. 1955), it took the scientific world by storm. The amino acid sequence demonstrated that proteins have a defined chemical composition. As more protein sequences were determined, it opened the door to comparative analysis. Orthologous sequences from different organisms were compared (see figure 1) to measure evolutionary distances for phylogenetic studies, to identify positions under functional selection for active site prediction and to detect patterns of co-evolution (Zuckerkandl et al. 1962, Margoliash 1963, Zuckerkandl et al. 1965 and Wyckoff et al. 1968). Wyckoff mapped the difference between rat and bovine RNase onto the 3D protein structure, noting a spatial relationship between observed pair of substitutions. To account for this co-evolution pattern, in 1970, Fitch proposed an improved method for determining codon variability in genes. Fitch reasoned that a very restricted set of positions can accept mutations, but upon fixation other positions open up. Though without 3D structure or further mutational experiments (Yanofsky et al. 1964), it was not possible to predict spatial relationships at the time.

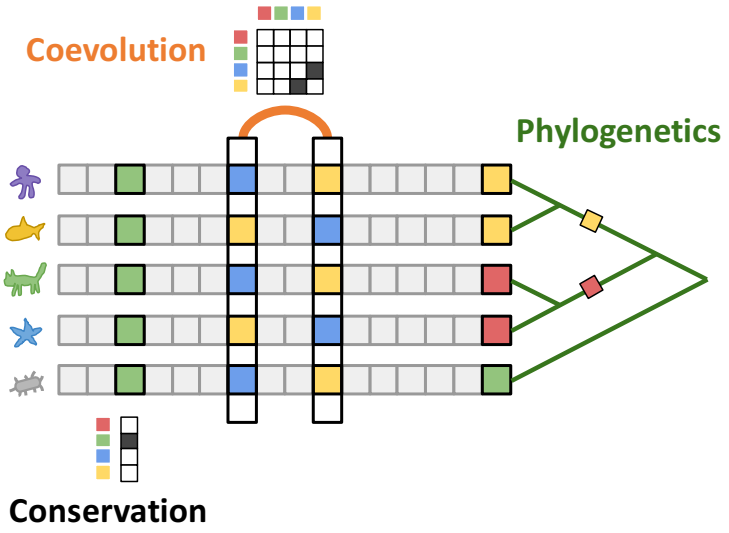


Figure 1.1. A cartoon illustrating the utility of a MSA (multiple sequence alignment). For clarity, only four characters are used (red, green, cyan and yellow). The MSA can be used to measure per position conservation, coevolution between positions and phylogenetic inference.

1.2 STATISTICAL MODEL OF PROTEIN EVOLUTION

Even though co-dependence between positions was noted early on, due to low number of sequences, initial efforts focused on general substitution models or models that assumed independence between positions. In 1966, Margaret Dayhoff compiled the first empirical substitute rate matrices for amino acids (Eck et al. 1966) using her atlas of protein sequence. This followed by position-specific scoring matrices (PSSM) (Stormo et al. 1982) and application of Hidden Markov models (HMM) to protein sequences (Krogh et al. 1994, Karplus et al. in 1997, Eddy et al. 1998), modeling for insertion and deletions.

In 1998, Alan Lapedes and co-workers proposed using a Markov Random Field (MRF) to account for both per position conservation and coupling (or co-evolution) between residue pairs (Lapedes et al. 1998 and 1999). Briefly, an MRF is defined as follows, where the probability of a given sequence X is:

$$P(X) = \frac{1}{Z}\exp\left(\sum_{i=1}^{L}\left[V_i(x_i) + \sum_{j>i}^{L} W_{i,j}\left(x_i, x_j\right)\right]\right)$$

L is the number of positions in a given multiple sequence alignment. x_i is the amino acid identity at position i. V_i is a vector of parameters encoding individual propensity for each amino acid at position i, and $W_{i,j}$ is a matrix of parameters modeling the statistical coupling in amino acid propensities between positions i and j of the protein. Z is the partition function to ensure the probabilities sum to 1. A graphical version of the model is shown in Figure 2 below.

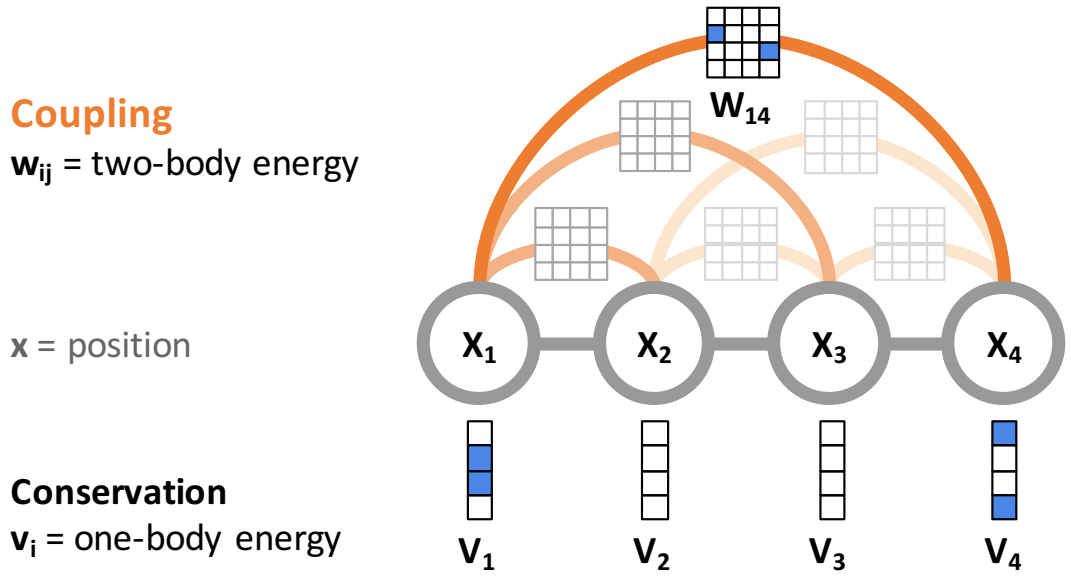


Figure 1.2. Graphical representation of a Markov Random Field.
For clarity, an alphabet of only 4 characters is s used. X_1 is the amino acid identity at position 1. V_1 is a vector encoding conservation at position 1. W_{14} is a matrix encoding the couplings at positions 1 and 4 of the sequence.

During the learning procedure the V and W parameters are optimized to maximize the probability of the given sequence. But there is a problem, to compute Z, all possible sequences of length L (20^L) must be considered, making the problem computationally intractable for computing the likelihood, let alone optimization. Partially due to this reason and the lack of sequences at the time, Lapedes' work was largely forgotten until a decade later when Thomas et al. proposed a greedy algorithm to learn an MRF for protein design by adding edges between nodes with high mutual information (Thomas et al. 2005). A few years later Weigt et al. used messages passing algorithm to learn the MRF across protein-protein interfaces (Weigt et al. 2008). Unfortunately, these learning algorithms did not provide guarantee of optimal model.

In 2010, Kamisetty et al. proposed using pseudo-likelihood (a different objective function), that eliminates the need to compute the global Z. This approach was first implemented in a method called GREMLIN (Generative REgularized ModeLs of proteINs) (Balakrishnan and Kamisetty et al. 2011). Briefly, the pseudo-likelihood of θ (parameters V, W), given a msa (multiple sequence alignment) is:

$$pll(\theta|msa) = \sum_{n=1}^{N}\sum_{i=1}^{L} \log \frac{\exp\left(V_i(x_i^n) + \sum_{\substack{j=1 \\ j\neq i}}^{L} W_{i,j}\left(x_i^n, x_j^n\right)\right)}{\sum_{c=1}^{21} \exp\left(V_i(c) + \sum_{\substack{j=1 \\ j\neq i}}^{L} W_{i,j}\left(c, x_j^n\right)\right)} - R(\theta)$$

Under certain conditions and assumptions, it was shown that θ that maximizes $pll(\theta|msa)$ is equal to the θ that maximizes $\mathcal{L}(\theta|msa)$ (Besag 1971). N is the number of sequences in the multiple sequence alignment. x_i^n is the amino acid identity at sequence n and position i. The global Z is replaced with a local Z. Instead of computing every possible sequence of length L, only one position is considered at a time, while all other characters remain fixed. The 21 characters (c) are the 20 amino acids plus gap (or deletion). R is the regularization or penalty term to promote a sparse network:

$$R(\theta) = \lambda_v \sum_{i=1}^{L} ||V_i||_2^2 + \lambda_w \sum_{\substack{i,j=1 \\ j\neq i}}^{L} ||W_{i,j}||_2^2$$

λ_v is 0.01 and λ_w is 0.2(L-1). The regularization can be used to encode additional prior information, such as P(contact|PSIPRED, sequence separation) or P(contact|SVMCON) (Kamisetty et al. 2013). PSIPRED uses neural networks to predicts the secondary structure given a PSSM window (McGuffin et al. 2000). SVMCON uses support vector machines to predict contacts given two PSSM windows (Cheng et al. 2007). Since R is outside the $\sum_{n=1}^{N}()$ loop, this effectively decreases the penalty as the number of sequences increase. Following optimization, each $W_{i,j}$ is L2-normalized to obtain a single value. The matrix of normalized values is then corrected for effects of entropy using Average Product Correction (Dunn et al. 2008). Remarkably, the strength of these values was found to correlate to physical residue-residue contacts in protein 3D structures solved by experimental methods such as Xray or NMR (Balakrishnan and Kamisetty et al. 2011), indicating the approach can be used for contact prediction. See Figure 3 below for an example.

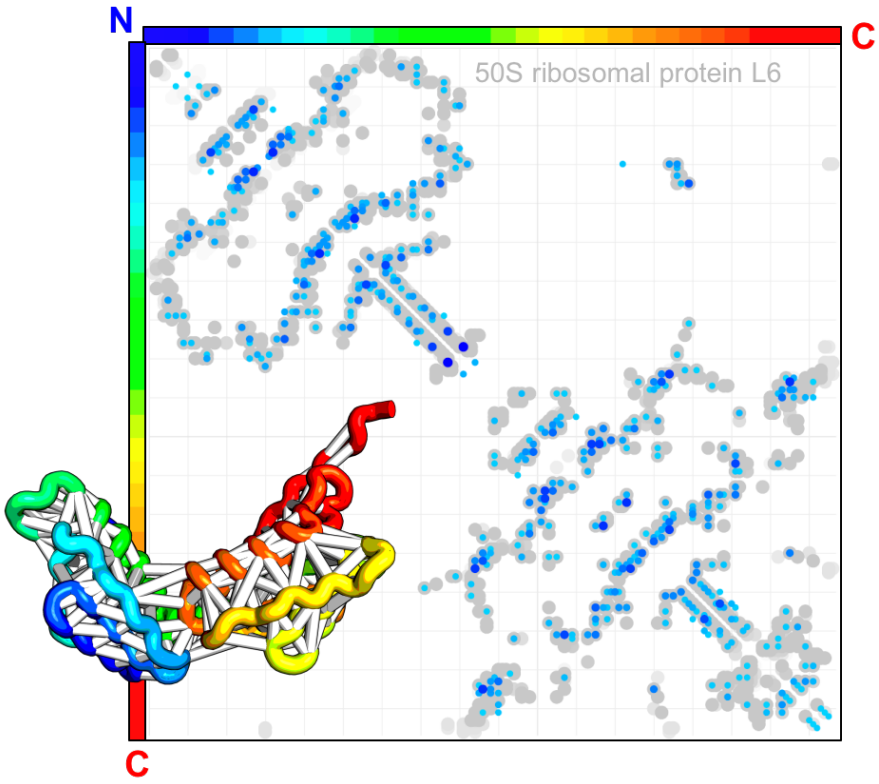


Figure 1.3. Comparing predicted contacts to XRAY contacts.
The top 3L/2 (L=length) co-evolving contacts are shown as white lines connecting residues in the ribbon diagram of a protein in the bottom left corner. The protein is colored in rainbow (blue to red, from n to c terminus). The contact map in the background is a 2D representation of the 3D structure. In grey are the contacts found in the XRAY crystal structure. Contacts are defined when the minimal distance between any two heavy atoms is within 5 angstroms. The coupling residues are in shades of blue, depending on the strength. Here we see that nearly all the blue contacts overlay the grey.

1.3 CONTACT PREDICTION

In parallel to the development of statistical models for phylogenetics, sequence generation and predicting effects of mutations, there was also interest in the statistical models for residue-residue contact prediction. These efforts were further motivated when it was demonstrated that one could easily recover a 3D structure from 2D contact maps, even from noise-corrupted contact maps (Vendruscolo et al.1997) or from sparse contacts (one contact for every seven residues) (Skolnick et al. 1997). In a more recent CASP10 experiment, it was shown that one correct contact for every 12 residues is enough to generate fold-level accuracy models (Kim et al. 2013). All that was missing was a method that can predict accurate contacts.

Due to lack of sequences, the initial methods for contact prediction resorted to “local” statistical models. MI is one of the first application, in which coevolution was measured as the sum of marginal entropies of two positions minus the joint entropy of both positions (Cover et al. 1991, Chiu et al. 1991). Other approaches include correlated mutations (Gobel et al. 1994, Shindyalov et al. 1994) and statistical coupling analysis (SCA) (Lockless et al. 1999). SCA subsamples the MSA according to the amino acid (AA) distribution of one position then checks to see how that affects the AA distribution in other positions. If the distribution is affected significantly, it is thought to be coupled. One of the flaws in these local approaches is that each pair of positions are evaluated independent of all other positions, hence a “local” statistical model. The problem comes from indirect correlations. If position A is coupled to B, and B is coupled to C, if one was to analyze the coupling of A to C independent of B, there is likely to be some indirect correlation detected. Global statistical models on the other hand, such as MRFs, consider all edges simultaneously, with a penalty for the addition of new edges. This promotes sparse number of edges.

Since development of GREMLIN (described in section 1.2), two faster alternatives to learning the direct couplings from MRFs were introduced: Direct Coupling Analysis (DCA) (Morcos et al. 2011) and Protein Sparse InverseCOVariance (PSICOV) (Jones et al. 2012). These two methods accomplish the task of separating direct from indirect correlations by estimating the inverse covariance matrix through approximate moment matching. Later it was shown that the former approach (GREMLIN), optimized for contact prediction, results in substantially more accurate contacts (Kamisetty et al. 2013, Ekeberg et al. 2013). Going from MIc (mutual information with corrections, Jeong et al. 2012) to mfDCA/PSICOV increased the contact prediction accuracy by 10%. Going from mfDCA/PSICOV to GREMLIN increased the accuracy by another 10%. See figure 4 below for details.

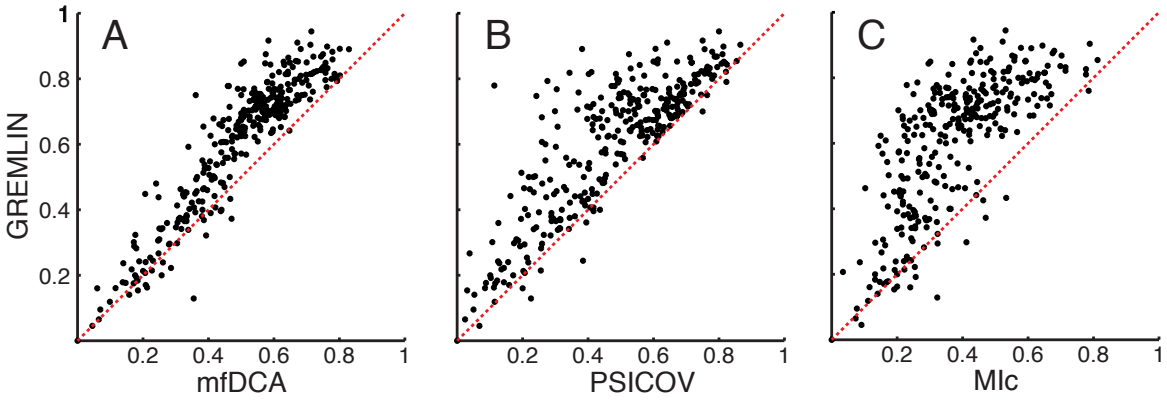


Figure 1.4. Accuracy of contact prediction.
Comparison of GREMLIN with mfDCA (A), PSICOV (B) and MIc (C). Each point corresponds to a protein, the axes indicate the accuracy of the top ranked L/2 Cβ−Cβ contacts predicted by the indicated methods.

1.4 PROTEIN STRUCTURE PREDICTION

The most successful methods in protein structure prediction have been those that copy fragments from homologous structures, based on local sequence-structure relationships (Han et al. 1996). As more protein structures are experimentally solved and added to the PDB (Protein databank), the more likely a larger homologous fragment can be identified, in which case it becomes a "homology modeling" problem (Song et al. 2013). For *de novo* structure prediction (when no homolog can be identified in the PDB) the short fragments (typically 3-9 amino acids in length) not only speed-up conformational sampling, but also restrict the sampling to protein-like space (Simons et al. 1997, Jones et al. 1997). Given the large conformational space a protein can adopt, *de novo* protein structure prediction has been largely limited to small and simple folds (Bradley et al. 2005). Challenges to adequate sampling are at both local and global levels. Local structure can be modelled incorrectly if the fragments are incorrect (due to incorrect secondary structure prediction), these in return can influence the global structure, where a helix or sheet might need to break to sample the correct fold. Other global issues include complex topologies with many non-local interactions (not covered by fragments). Any non-local contact information can drastically reduce the search space and provide a gradient for global sampling.

Residue-residue contact information such as from NMR experiments has been shown to drastically reduce the search space. For proteins up to 400 amino acids, backbone–backbone NOEs combined with methyl–methyl NOEs from I, L, and V residues can be used with the ROSETTA energy function and fragment-based sampling to generate accurate models (1.1–4.1

Å RMSD) (Raman et al. 2010, Lange et al. 2012). In other words, the ROSETTA energy function combined with sampling in protein-like space (through fragments from the PDB), allows resolving the sparse and ambiguous restraints from NMR experiments.

Co-evolution methods provide another source of residue-residue contact data (see section 1.3). In 2007 with SCA and again in 2011 using PSICOV, Taylor et al. showed that there was enough sequence data to predict the correct topology of a protein structure using contacts and a quick enumeration of different folds (Bartlett et al. 2008, Taylor et al. 2012). This was followed by a fragment-based approach (Nugent et al. 2012) for transmembrane proteins and MD-like approach using contacts derived from DCA (Sulkowska et al. 2012 and Marks et al. 2012). One of the disadvantages of using the MD-like approach is that false contacts can distort the model resulting in spaghetti-like structures.

One way to overcome the false contacts from distorting the models and to keep the gradient during minimization, is to use progressively smaller number of contacts (that are more reliable) or to randomly group the contacts into ambiguous groups (Lange et al. 2012). The different restraint sets can be used to generate models, but then each trajectory can be re-ranked based on total number of contacts made. Though these approaches should be useful to eliminate false random contacts, they are unlikely to work for homo-oligomeric structures with large interfaces, or structures with conformational change (figure 5), where an extensive number of contacts are involved. To overcome this problem, we use sigmoidal restraints during sampling instead of harmonic (see Figure 6), where whole patches of contacts can be effectively turned off if they are inconsistent with another patch of contacts made, though this could require a large amount of sampling.

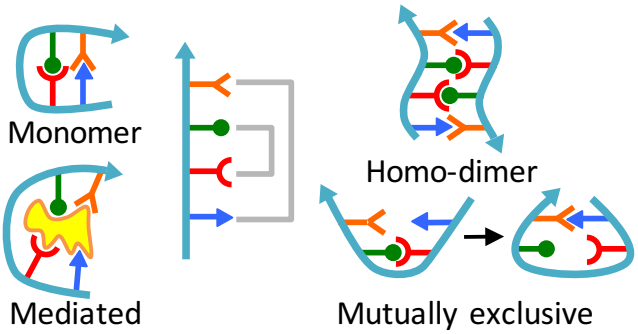


Figure 1.5. Source of contacts.
Natural selection optimizes contacts that are within the monomer, at the homo-oligomeric interface, ligand mediated and mutually exclusive set made at different conformational stages of

function or folding. The same predicted contacts (grey lines in center) can be satisfied a couple different ways (as illustrated).

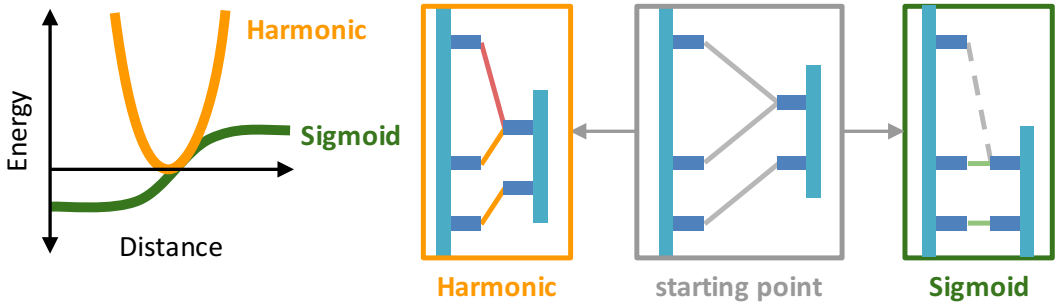


Figure 1.6. Using sigmoidal restraints to find self-consistent contacts.
On the left is a plot of the functional forms of the restraint types. On the right is a cartoon illustrating the effects of the two different restraint types on three given contacts in grey.

For my thesis, I focus on making models using a significantly more accurate contact prediction method GREMLIN in combination with the ROSETTA structure prediction pipeline. For details see chapter 2 (for protein-protein docking) and chapters 3 to 4 (for protein modeling).

1.5 REFERENCES

1 Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I., & Langmead, C. J. (2011). Learning generative models for protein fold families. Proteins: Structure, Function, and Bioinformatics, 79(4), 1061-1078.

2 Bartlett, Gail J., and William R. Taylor. "Using scores derived from statistical coupling analysis to distinguish correct and incorrect folds in de-novo protein structure prediction." *Proteins: Structure, Function, and Bioinformatics* 71, no. 2 (2008): 950-959.

3 Besag, Julian. "Efficiency of pseudolikelihood estimation for simple Gaussian fields." *Biometrika* (1977): 616-618.

4 Bradley, Philip, Kira MS Misura, and David Baker. "Toward high-resolution de novo structure prediction for small proteins." *Science* 309, no. 5742 (2005): 1868-1871.

5 Cheng, Jianlin, and Pierre Baldi. "Improved residue contact prediction using support vector machines and a large feature set." *BMC bioinformatics* 8, no. 1 (2007): 113.

6 Chiu, David KY, and Ted Kolodziejczak. "Inferring consensus structure from nucleic acid sequences." *Bioinformatics* 7, no. 3 (1991): 347-352.

7 Cover, Thomas M., and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

8 Dunn, Stanley D., Lindi M. Wahl, and Gregory B. Gloor. "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction." *Bioinformatics* 24, no. 3 (2007): 333-340.

9 Göbel, Ulrike, Chris Sander, Reinhard Schneider, and Alfonso Valencia. "Correlated mutations and residue contacts in proteins." Proteins: Structure, Function, and Bioinformatics 18, no. 4 (1994): 309-317.

10 Han, Karen F., and David Baker. "Global properties of the mapping between local amino acid sequence and local structure in proteins." *Proceedings of the National Academy of Sciences* 93, no. 12 (1996): 5814-5818.

11 Henikoff, Steven, and Jorja G. Henikoff. "Amino acid substitution matrices from protein blocks." *Proceedings of the National Academy of Sciences* 89, no. 22 (1992): 10915-10919.

12 Eck, Richard V., and Margaret O. Dayhoff. "Atlas of protein sequence and structure." (1966).

13 Eddy, Sean R. "Profile hidden Markov models." *Bioinformatics (Oxford, England)* 14, no. 9 (1998): 755-763.

14 Ekeberg, Magnus, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models." *Physical Review E* 87, no. 1 (2013): 012707.

15 Jeong, Chan-Seok, and Dongsup Kim. "Reliable and robust detection of coevolving protein residues." *Protein Engineering, Design & Selection* 25, no. 11 (2012): 705-713.

16 Jones, David T. "Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs." *Proteins: Structure, Function, and Bioinformatics* 29, no. S1 (1997): 185-191.
17 Jones, David T., Daniel WA Buchan, Domenico Cozzetto, and Massimiliano Pontil. "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments." *Bioinformatics*28, no. 2 (2012): 184-190.
18 Karplus, K., Sjölander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L. and Sander, C., 1997. Predicting protein structure using hidden Markov models. *Proteins Structure Function and Genetics*, *29*(s 1), pp.134-139.
19 Kamisetty, Hetunandan, Sergey Ovchinnikov, and David Baker. "Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era." *Proceedings of the National Academy of Sciences* 110, no. 39 (2013): 15674-15679.
20 Krogh, Anders, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. "Hidden Markov models in computational biology: Applications to protein modeling." *Journal of molecular biology* 235, no. 5 (1994): 1501-1531.
21 Kim, David E., Frank DiMaio, Ray Yu-Ruei Wang, Yifan Song, and David Baker. "One contact for every twelve residues allows robust and accurate topology-level protein structure modeling." *Proteins: Structure, Function, and Bioinformatics* (2013).
22 Lapedes, A. S., B. G. Giraud, L. C. Liu, and G. D. Stormo. *A maximum entropy formalism for disentangling chains of correlated sequence positions*. No. LA-UR-98-1094. Los Alamos National Lab., NM (US), 1998.
23 Lapedes, Alan S., Bertrand G. Giraud, LonChang Liu, and Gary D. Stormo. "Correlated mutations in models of protein sequences: phylogenetic and structural effects." *Lecture Notes-Monograph Series* (1999): 236-256.
24 Lapedes A, Giraud B, Jarzynski C. 2012. Using sequence alignments to predict protein structure and stability with high accuracy. arXiv 1207.2484. http://arxiv.org/abs/1207.2484
25 Lockless, Steve W., and Rama Ranganathan. "Evolutionarily conserved pathways of energetic connectivity in protein families." Science 286, no. 5438 (1999): 295-299.
26 Lange, Oliver F., Paolo Rossi, Nikolaos G. Sgourakis, Yifan Song, Hsiau-Wei Lee, James M. Aramini, Asli Ertekin et al. "Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples." *Proceedings of the National Academy of Sciences* 109, no. 27 (2012): 10873-10878.
27 Margoliash, E., 1963. Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences*, *50*(4), pp.672-679.
28 Marks, Debora S., Thomas A. Hopf, and Chris Sander. "Protein structure prediction from sequence variation." *Nature biotechnology* 30, no. 11 (2012): 1072-1080.
29 McGuffin, Liam J., Kevin Bryson, and David T. Jones. "The PSIPRED protein structure prediction server." *Bioinformatics* 16, no. 4 (2000): 404-405.
30 Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of USA 108:E1293–E1301.
31 Morcos, Faruck, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. "Direct-coupling analysis of residue coevolution captures native contacts across many protein families." *Proceedings of the National Academy of Sciences* 108, no. 49 (2011): E1293-E1301.
32 Nugent, Timothy, and David T. Jones. "Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis." *Proceedings of the National Academy of Sciences* 109, no. 24 (2012): E1540-E1547.
33 Ryle, A.P., Sanger, F., Smith, L.F. and Kitai, R., 1955. The disulphide bonds of insulin. *Biochemical Journal*, *60*(4), p.541
34 Raman, Srivatsan, Oliver F. Lange, Paolo Rossi, Michael Tyka, Xu Wang, James Aramini, Gaohua Liu et al. "NMR structure determination for larger proteins using backbone-only data." *Science* 327, no. 5968 (2010): 1014-1018.
35 Simons, Kim T., Charles Kooperberg, Enoch Huang, and David Baker. "Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions." *Journal of molecular biology* 268, no. 1 (1997): 209-225.
36 Stormo, Gary D., Thomas D. Schneider, Larry Gold, and Andrzej Ehrenfeucht. "Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli." *Nucleic acids research* 10, no. 9 (1982): 2997-3011.
37 Sułkowska, Joanna I., Faruck Morcos, Martin Weigt, Terence Hwa, and José N. Onuchic. "Genomics-aided structure prediction." *Proceedings of the National Academy of Sciences* 109, no. 26 (2012): 10340-10345.

38 Skolnick, Jeffrey, Andrzej Kolinski, and Angel R. Ortiz. "MONSSTER: a method for folding globular proteins with a small number of distance restraints." *Journal of molecular biology* 265, no. 2 (1997): 217-241.
39 Song, Yifan, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, T. J. Brunette, James Thompson, and David Baker. "High-Resolution Comparative Modeling with RosettaCM." *Structure* 21, no. 10 (2013): 1735-1742.
40 Taylor, William R., David T. Jones, and Michael I. Sadowski. "Protein topology from predicted residue contacts." *Protein Science* 21, no. 2 (2012): 299-305.
41 Thomas J, Ramakrishnan N, Bailey-Kellogg C. Graphical models of residue coupling in protein families. In: BIOKDD '05: Proceedings of the 5th International Workshop on Bioinformatics, ACM, New York, NY, 2005, 12–20.
42 Vendruscolo, Michele, Edo Kussell, and Eytan Domany. "Recovery of protein structure from contact maps." *Folding and Design* 2, no. 5 (1997): 295-306.
43 Weigt, Martin, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. "Identification of direct residue contacts in protein–protein interaction by message passing." *Proceedings of the National Academy of Sciences* 106, no. 1 (2009): 67-72.
44 Wyckoff, H. W. "Discussion." In *Brookhaven Symp. Biol*, vol. 21, pp. 252-257. 1968. < https://digital.library.unt.edu/ark:/67531/metadc170990/m2/1/high_res_d/metadc67245.pdf>
45 Yanofsky, Charles, Virginia Horn, and Deanna Thorpe. "Protein structure relationships revealed by mutational analysis." *Science* 146, no. 3651 (1964): 1593-1594.
46 Zuckerkandl, E. and Pauling, L., 1962. Molecular disease, evolution and genetic heterogeneity.
47 Zuckerkandl, E. and Pauling, L., 1965. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, *97*, pp.97-166.

Chapter 2. ROBUST AND ACCURATE PREDICTION OF RESIDUE-RESIDUE INTERACTIONS ACROSS PROTEIN INTERFACES USING EVOLUTIONARY INFORMATION

A version of this chapter has been previously published as:

Ovchinnikov, Sergey, Hetunandan Kamisetty, and David Baker. "Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information." *Elife* 3 (2014): e02030.

2.1 ABSTRACT

Do the amino acid sequence identities of residues that make contact across protein interfaces covary during evolution? If so, such covariance could be used to predict contacts across interfaces and assemble models of biological complexes. We find that residue pairs identified using a pseudo-likelihood-based method to covary across protein–protein interfaces in the 50S ribosomal unit and 28 additional bacterial protein complexes with known structure are almost always in contact in the complex, provided that the number of aligned sequences is greater than the average length of the two proteins. We use this method to make subunit contact predictions for an additional 36 protein complexes with unknown structures, and present models based on these predictions for the tripartite ATP-independent periplasmic (TRAP) transporter, the tripartite efflux system, the pyruvate formate lyase-activating enzyme complex, and the methionine ABC transporter.

2.2 INTRODUCTION

Recent work has demonstrated the accuracy of coevolution-based contact prediction for monomeric proteins using a global statistical model (Thomas et al., 2008) to distinguish between direct and indirect couplings (Marks et al., 2011; Morcos et al., 2011; Hopf et al., 2012; Nugent and Jones, 2012; Jones et al., 2012; Lapedes et al., 2012; Marks et al., 2012; Sułkowska et al., 2012; Kamisetty et al., 2013). While early approaches relied on estimating an inverse covariance matrix (Marks et al., 2011; Morcos et al., 2011; Jones et al., 2012), more recent studies have shown that a pseudolikelihood-based approach (Balakrishnan et al., 2011) results in more accurate predictions (Ekeberg et al., 2013; Kamisetty et al., 2013) for a range of alignment sizes and protein lengths.

In contrast to this rich body of work for monomeric proteins, relatively little is known about the utility of such statistical models in predicting protein–protein interactions. The more general problem of predicting if two proteins interact with each other has been studied extensively using a wide variety of approaches (de Juan et al., 2013; Hosur et al., 2012; Zhang et al., 2012; Shoemaker and Panchenko, 2007, Valencia and Pazos, 2002, Ochoa and Pazos, 2010). Amino acid residue coevolution has been used to predict residue–residue interactions across

interfaces with local statistical models (Pazos et al., 1997; Halperin et al., 2006). As noted above, the accuracy of these models is reduced by the confounding of direct and indirect correlations (Lapedes et al., 1999; Weigt et al., 2009); the application of global statistical models to coevolution-based contact prediction across interfaces has been limited to the case of the histidine-kinase/response-regulator two component system (Burger and van Nimwegen, 2008; Weigt et al., 2009; Schug et al., 2009; Dago et al., 2012).

In this study, we examine residue–residue covariation across protein–protein interfaces using a pseudo-likelihood-based statistical method. In a large set of complexes of known structure, we find that covarying pairs of positions are almost always in contact in the three-dimensional structure, provided there are sufficient aligned sequences. We find further that significant residue–residue covariance occurs frequently between physically interacting protein pairs but very rarely between non-interacting pairs, and hence should be useful for predicting whether two proteins interact. We use the pseudo-likelihood method to predict contacts across protein-interfaces for 36 evolutionarily conserved complexes of unknown structure and present structure models for four of the complexes particularly well constrained by these data.

2.3 RESULTS

For a single protein family, it is straightforward to generate a multiple sequence alignment and subsequently identify covarying residue pairs. To identify covarying residue pairs between two proteins A and B is not as easy: only organisms that contain an ortholog of protein A and protein B contribute, and in generating the alignments the protein A and protein B sequences for each organism must be properly paired. To simplify the ortholog identification problem, we focus on pairs of genes with conserved chromosomal locations separated in the genome by fewer than 20 other annotated genes. We then build GREMLIN global statistical models for sequences in the paired protein families. The models have ‘one-body’ parameters for each amino acid at each position in the two proteins, and ‘two-body’ parameters for each pair of amino acids at each pair of positions in the two proteins. These parameters are obtained by maximizing the pseudo-likelihood of the observed sequence pairs, rather than their likelihood, which makes the quite formidable estimation tractable. In the following sections, we investigate the structural contexts of residue pairs with large values of these two-body coupling parameters.

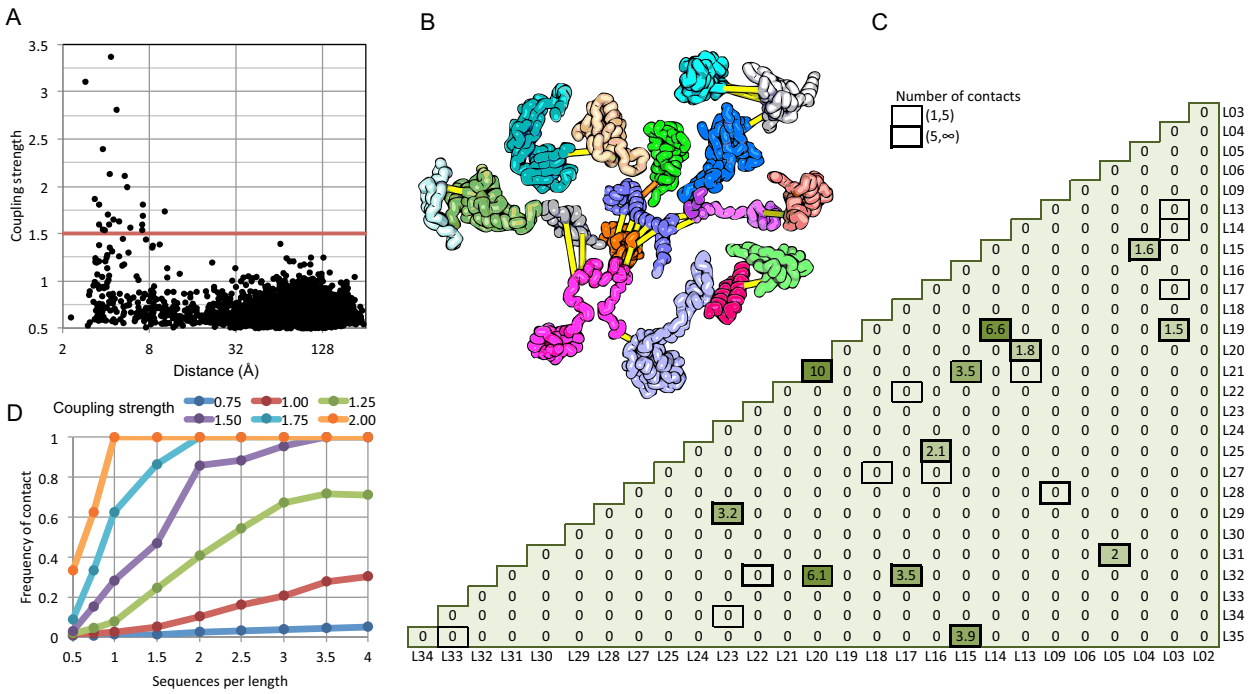


Figure 2.1. Residue pairs with high normalized coupling strengths are in contact in the 50S ribosomal subunit.
(A) Coupling strengths and inter-residue distances for each residue pair in the 50S subunit (black dots). Residue pairs with coupling strength greater than 1.5 are nearly always less than 8 Å apart. (B) Locations of coevolving (high coupling strength) residue pairs in the protein component of the 50S subunit. The monomers have been pulled apart slightly for clarity. Lines connect residue pairs with coupling strength greater than 1.5; yellow, distance less than 8 Å; orange, distance less than 12 Å. (C) Protein pairs with strong inter-residue covariation (colors) make contact in the three-dimensional structure (black boxes). For each protein pair, the sum of the coupling strength greater than 1.5 for each pair of 50S subunit proteins is indicated; black boxes indicate contacts in the crystal structure. (D) Dependence of contact prediction accuracy on coupling strength and the number of sequences in the alignments. For each of the indicated coupling strength cutoffs (colors), the frequency of contact in the 50S structure (y axis) was computed for sub alignments with different sequence depths (x axis).

2.3.1 *Residue–residue covariation in the bacterial 50S ribosomal unit*

We began by studying residue–residue coupling parameters in the bacterial 50S ribosomal subunit— the largest evolutionarily conserved bacterial multiprotein complex with an atomic resolution structure. For each individual protein in the complex, we constructed multiple sequence alignments by querying the UniProt sequence database (Wu et al., 2006) for homologous sequences. For every pair of proteins in the complex, we then constructed a paired multiple sequence alignment (‘Materials and methods’). For each such paired alignment, we built

a GREMLIN global statistical model, computed normalized coupling strengths from the two body coupling parameters, and ranked inter protein residue pairs based on these scores (‘Materials and methods’). A coupling strength larger than one indicates higher than average coupling between two residues.

We find that in the 50S ribosomal subunit only a small fraction of residue pairs coevolve, as indicated by coupling strengths (y axis of Figure 1A) greater than 1.5. Remarkably, the two residues in each of these pairs are almost all within 8 Å of each other in the 50S crystal structure (Figure 1A) and all are within 12 Å. The locations of the covarying residue pairs in the 50S structure (with the individual proteins pulled apart for clarity) are shown in Figure 1B; yellow lines indicate distances less than 8 Å and orange lines, distances less than 12 Å. For the 50S ribosome, the GREMLIN model was built using sequence data from ~1500 non-redundant genomes; Figure 1D suggests that for complexes with such large numbers of aligned sequence, residue–residue interactions across interfaces can be predicted with quite high confidence based on amino acid sequence covariation.

For a large protein–protein complex, can the sum of the coupling strengths between pairs of proteins in the complex be used to distinguish directly interacting and non-interacting protein pairs? In the 50S subunit, every pair of proteins with summed coupling strengths (numbers in Figure 1C) greater than 1.5 interacts with each other (boxes in Figure 1C). There are, however, several instances of protein pairs that contact in the 50S subunit for which no covariance is observed; clearly not every interaction will be identified by the sum of the coupling strengths, for example between two proteins that are held together primarily by the ribosomal RNA.

How many aligned sequences are required for accurate contact prediction? To assess the dependence on alignment depth, we generated paired sub-alignments with varying numbers of sequences for every pair of 50S proteins and recomputed coupling strengths for each sub-alignment. For each alignment depth, we calculated the fraction of residue pairs within 12 Å for different ranges of coupling strengths. We find that the greater the number of aligned sequences, the lower the value of the coupling strength above which residue pairs are likely to be in contact in the structure (Figure 1D). For example, if the number of aligned sequences is greater than the sum of the lengths of the two proteins, residue–residue contact predictions are likely to be accurate if the coupling strength is 2 or greater (Figure 1D: orange dots), while if there are twice as many sequences, contact predictions are accurate above a coupling strength of 1.5 (the cutoff

shown in Figure 1A). A sigmoidal function of the coupling strength and the number of sequences per position in the complex accurately fits the observed contact frequency data ('Materials and methods' and Figure 1—figure supplement 1); we refer to the fitted values as GREMLIN scores for the remainder of the paper.

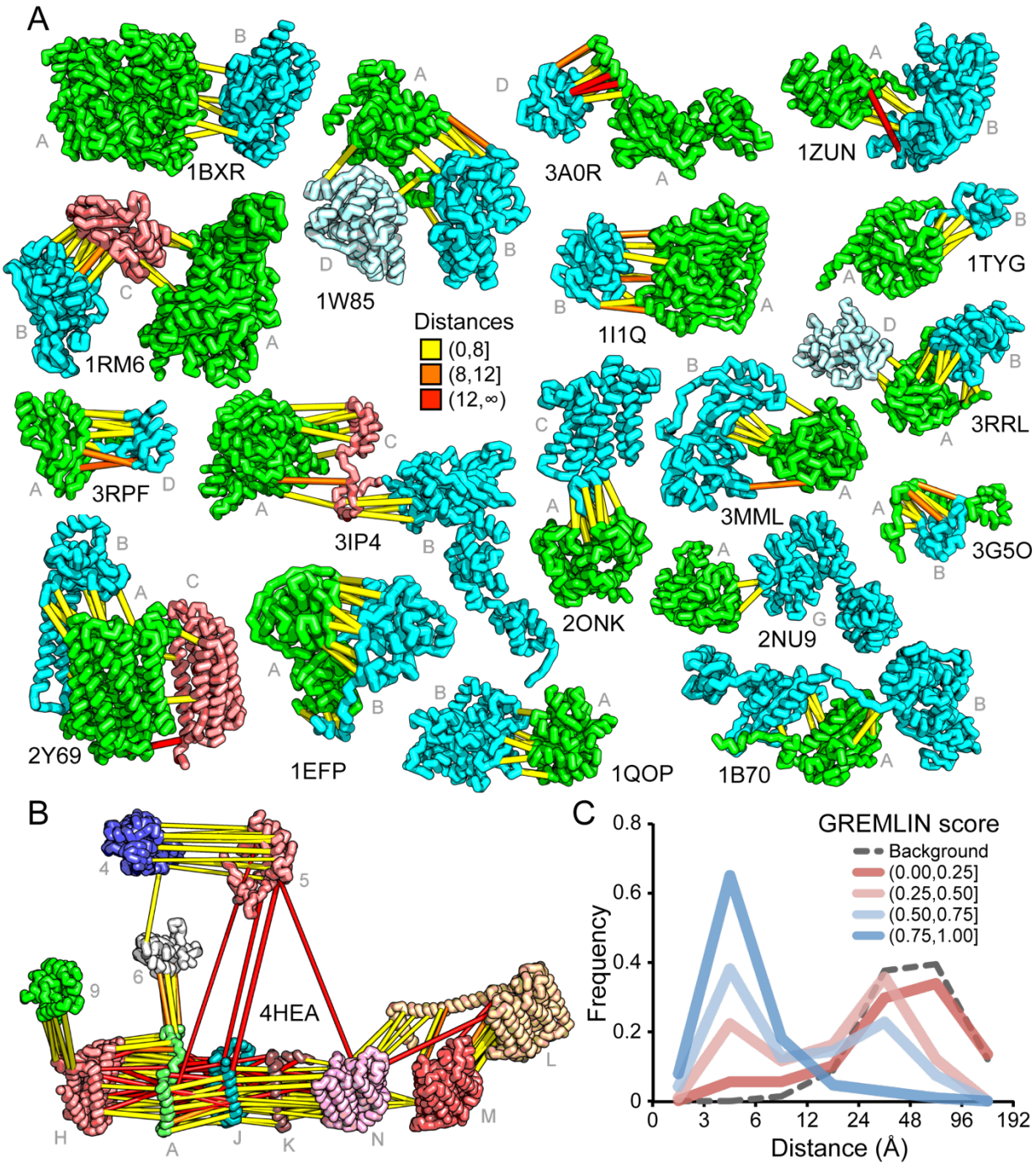


Figure 2.2. Residue covariation in complexes with known structures.

(A) Residue-pairs across protein chains with high GREMLIN scores almost always make contact across protein interfaces in experimentally determined complex structures. All contacts with GREMLIN scores greater than 0.6 are shown; the structures are pulled apart for clarity. Labels are according to chains in the PDB structure. (B) Complex I of the electron transport chain has an unusually large number of highly co-varying inter residue pairs not in contact in the crystal structure of 4HEA; these contacts may be formed in different state of the complex. Residue pairs within 8 Å are in yellow, between 8 Å and 12 Å in orange, and greater than 12 Å, in red. Distances are the minimal distances between any side chain heavy atom. Labels are according to chains in 4HEA. (C) Dependence of inter-residue distance distributions on GREMLIN score. All residue– residue pairs between subunits in the benchmark set were grouped into four bins based on their GREMLIN score (colors), and the distribution of residue–residue distances (x axis) within each bin computed from the three dimensional structures. See Figure 2—source data 1 for the table of all the interfaces used in the calculation.

2.3.2 *Bacterial complex benchmark*

We next generated paired-alignments for all *E. coli* gene-pairs that had conserved intergenic distances across genomes deposited in the UniProt ('Materials and methods'). As the 50S results (Figure 1D) suggested that alignment depths greater than the average of the lengths of the two proteins were required for accurate prediction, we focused on paired alignments with at least this number of sequences—1126 gene pairs in total excluding the ribosomal proteins. For each of these 1126 pairs, we generated GREMLIN global statistical models and determined the coupling strength for each residue pair.

For 64 of the 1126 gene pairs, at least one pair of residues had GREMLIN score >0.85. For 28 of the 64 pairs three-dimensional structures have been determined experimentally, and the locations of the residue pairs with GREMLIN score >0.6 for several of these complexes are shown in Figure 2A (pairs within 8 Å are in yellow, between 8 Å and 12 Å in orange, and greater than 12 Å, in red). Almost all pairs with GREMLIN scores greater than 0.6 are in contact in the complex structures, with the notable exception of the NADH dehydrogenase subunits (Figure 2B). The complex is thought to undergo a cascade of conformational changes during electron transfer (Baradaran et al., 2013); the high GREMLIN score contacts not made in the solved structure may provide insight into the nature of these changes. As observed for the 50S complex (Figure 1C), the existence of one or more high GREMLIN scores between two proteins provides evidence that the proteins interact: 44% (28/64) of the protein pairs with high GREMLIN scores form a complex which has been solved crystallographically compared to 8% (78/1126) over the whole set.

YIAM_YIAN (3.9L)

21_F	246_F	1.00
91_I	25_L	1.00
42_R	52_D	0.97
14_A	268_L	0.93
87_L	26_L	0.89
59_D	273_K	0.89
17_S	245_A	0.87
28_Y	234_L	0.84
43_Y	48_V	0.82
10_A	268_L	0.79
64_Q	17_G	0.76
56_A	276_S	0.75
9_L	272_A	0.75
46_V	52_D	0.72
13_L	271_A	0.67
50_F	51_A	0.66
95_L	47_L	0.62
46_V	51_A	0.60
10_A	269_I	0.60

YIAN_YIAO (3.6L)

168_S	196_Y	1.00
296_E	62_K	0.89
166_S	221_E	0.72
237_P	197_T	0.59

YIAM_YIAO (5.0L)

33_S	186_N	0.56

PFLA_PFLB (0.7L)

66_F	650_T	0.92
59_E	643_K	0.67
62_T	570_N	0.54
70_S	688_G	0.27

MDTP_MDTN (2.4L)

431_A	146_E	1.00
223_S	146_E	0.96
227_H	142_P	0.90
231_A	150_S	0.74
431_A	148_F	0.45
435_R	142_P	0.44

METI_METQ (2.0L)

80_R	193_D	0.99
80_R	192_D	0.91
87_I	182_L	0.70
87_I	66_F	0.56
174_Q	67_N	0.50
87_I	75_A	0.43

UMUC_UMUD (1.0L)

415_S	38_I	1.00
421_V	54_V	0.86
132_H	76_S	0.85
404_R	33_Y	0.72
183_D	127_V	0.66
78_A	34_V	0.60

IF1_SECY (1.4L)

32_V	343_A	0.89

YOHJ_YOHK (1.9L)

64_C	75_Y	1.00
61_N	79_E	1.00
115_S	94_I	1.00
79_G	38_M	1.00
110_V	101_V	0.99
106_V	101_V	0.99
99_S	115_A	0.97
83_M	42_I	0.96
92_Q	116_S	0.93
104_T	184_A	0.89
105_L	161_F	0.84
40_I	146_I	0.84
71_M	34_L	0.83
65_Y	79_E	0.81
113_W	169_M	0.80
16_V	14_I	0.78
74_L	65_L	0.78
82_V	42_I	0.76
101_A	157_L	0.75
39_S	70_V	0.72
115_S	176_A	0.70
106_V	105_T	0.68
115_S	179_L	0.66
100_C	150_C	0.65
18_I	72_A	0.64
109_L	169_M	0.63
110_V	97_I	0.60
110_V	213_I	0.60

CYOC_NUOK (3.8L)

49_V	80_L	0.96
83_Y	29_L	0.68
75_L	73_A	0.68
65_E	89_N	0.65
101_W	29_L	0.61
42_I	62_Y	0.60

ATP6_ATPF (2.3L)

74_K	34_E	1.00
149_V	10_Q	0.93
77_T	33_I	0.92
53_G	13_A	0.86
155_L	11_A	0.78
50_V	13_A	0.72
255_I	139_S	0.68
243_I	20_F	0.64
49_S	10_Q	0.64
263_Y	21_C	0.64
239_V	16_L	0.63
111_W	10_Q	0.62

FECI_FECR (3.3L)

162_E	14_R	0.96
133_Q	57_R	0.95
158_A	18_H	0.61

FTSI_FTSW (2.0L)

39_L	313_V	0.98
42_V	312_V	0.94
47_V	309_Y	0.80
43_A	309_Y	0.74
34_A	357_L	0.62

FLGB_FLGC (2.4L)

34_D	13_A	1.00
34_D	107_S	0.99
110_A	117_E	0.98
98_N	11_G	0.98
121_S	129_T	0.96
34_D	103_V	0.96
113_S	108_A	0.89
27_A	114_A	0.88
117_Q	128_K	0.80
111_D	122_V	0.78
113_S	125_M	0.71
27_A	27_N	0.63
131_M	69_E	0.61
23_Q	37_P	0.60
93_P	45_K	0.60

FLIP_FLIQ (2.2L)

185_I	84_L	0.99
54_I	55_I	0.90
229_V	77_V	0.89
203_V	24_L	0.68
205_M	66_G	0.67
213_P	55_I	0.66
243_F	72_L	0.65
149_L	67_P	0.64
188_T	80_L	0.63

FLGH_FLGI (0.6L)

52_F	133_V	0.99
82_L	257_S	0.59

FLHB_FLIR (1.1L)

181_A	188_L	0.98
185_A	192_T	0.64

FLIP_FLIR (1.5L)

78_L	187_A	0.90
97_T	13_L	0.71

MLAD_MLAE (1.4L)

14_L	13_G	0.93
14_L	24_G	0.76
8_I	6_L	0.65
58_V	6_L	0.65
65_V	219_F	0.63

CCMC_CCME (0.8L)

49_Q	104_R	0.95

CCMA_CCMB (1.3L)

95_E	16_R	0.87

MREB_NIFU (0.6L)

21_N	57_R	0.97
149_I	108_I	0.72

QMCA_YBBJ (1.5L)

145_E	115_R	0.91
36_R	113_H	0.72

PTPC1_PTPD (1.2L)

130_K	58_K	1.00
210_L	189_I	1.00
34_I	122_C	0.98
126_D	28_R	0.98
232_G	180_L	0.93
123_T	27_E	0.86
180_H	258_S	0.85
125_A	32_G	0.85
134_T	53_L	0.72
125_A	61_L	0.66
104_S	71_V	0.64
12_L	223_I	0.60

YADG_YADH (5.4L)

97_Q	22_R	0.94
49_G	93_E	0.93
106_Y	99_P	0.84
90_N	15_K	0.70
97_Q	18_H	0.68
82_L	97_V	0.67
104_G	14_A	0.64

RBSA_RBSC (5.6L)

365_S	188_Y	1.00
112_K	188_Y	0.97
301_A	211_L	0.95
93_Q	204_G	0.79
371_E	197_R	0.66
105_E	198_Y	0.64

APAG_PPIC (0.8L)

68_V	73_E	0.97

DDPB_DDPC (8.7L)

148_W	268_G	1.00
279_L	231_T	1.00
20_G	135_I	0.99
327_Y	180_L	0.97
327_Y	177_G	0.93
233_E	292_K	0.91
221_R	283_D	0.91
229_E	212_P	0.83
28_I	147_A	0.77
222_Q	217_Q	0.75
13_G	127_A	0.73
268_L	137_L	0.67
126_R	293_A	0.65
143_S	279_N	0.65
140_T	272_L	0.63
24_I	135_I	0.62
233_E	185_Y	0.59

ENGB_NDK (1.0L)

193_L	60_F	0.87

CLPP_CLPX (1.3L)

96_F	271_G	0.95
40_E	267_G	0.72

APPB_APPC (1.2L)

91_I	474_S	1.00
101_P	68_A	0.99
95_C	481_L	0.95
88_V	473_F	0.94
95_C	477_M	0.92
342_I	98_M	0.90
91_I	477_M	0.77
332_S	90_D	0.70

RNFE_RNFG (1.8L)

188_L	160_V	0.90
85_M	27_N	0.90
67_I	12_L	0.89
146_G	171_T	0.66

NRFC_NRFD (1.2L)

130_Y	92_S	0.93
129_Q	76_T	0.79
124_L	87_H	0.69
154_F	55_T	0.63

TOLQ_TOLR (4.9L)

149_I	27_V	0.91
177_A	27_V	0.91
177_A	25_L	0.90
177_A	26_L	0.84
139_Y	19_V	0.83
142_L	20_P	0.77
179_A	24_V	0.76
167_V	27_V	0.76
179_A	26_L	0.73
177_A	18_I	0.69
133_V	102_N	0.67
177_A	19_V	0.67
156_L	19_V	0.66
142_L	22_L	0.66
200_N	11_D	0.63
134_G	19_V	0.59

ATKA_ATKC (0.7L)

190_I	30_G	1.00
448_P	91_N	1.00
249_A	175_V	0.99
186_A	26_T	0.99
183_L	26_T	0.84
249_A	167_L	0.79
11_T	170_Y	0.78
198_F	24_L	0.63
238_A	83_G	0.62
244_F	55_I	0.60

FEOA_FEOB (0.9L)

41_R	378_F	0.94

YEJA_YEJB (3.1L)

466_D	325_R	0.90

Figure 2.3. Predicted residue–residue interactions across protein interfaces of unknown structure.

Strongly co-evolving residue pairs for complexes without known structure that had at leastone prediction with GREMLIN score greater than or equal to 0.85. Each row shows the residue pairs, their sequence identity and the GREMLIN score. Structure models for complexes highlighted in red are shown in Figure 5. Full dataset is provided with the deposited data.

A

Name	Complex	Chain 1	Chain 2	iRMSD	Fnat
Cytochrome c oxidase CO1/CO3	2y69AC	3dtuA	2y69C	0.4	0.99
3-oxoadipate-CoA transferase	3rrlAB	1k6dA	3rrlB	0.4	0.99
4-hydroxybenzoyl-CoA reductase	1rm6BC	1dgjA	1rm6C	0.6	0.93
Dihydroorotate dehydrogenase	1ep3AB	1dorA	1ep3B	0.9	0.89
Pyruvate dehydrogenase	1w85AB	1w85A	1ik6A	1.0	0.91
PTS-dependent dihydroxyacetone kinase	3pnlAB	3pnkA	2btdA	1.0	0.89
ABC transporter	2onkAC	2it1A	2onkC	1.4	0.88
Tryptophan synthase	1qopAB	2ekcA	1v8zA	1.6	0.83
Anthranilate synthase	1i1qAB	1k0eA	1i1qB	1.8	0.98
Succinyl-CoA synthetase	2nu9AG	1oi7A	2nu9G	1.9	0.83
Histidine kinase/response regulator	3a0rAD	3a0rA	3a10A	2.0	0.84
NADH dehydrogenase ND6/NDA	4hea6A	3ias6	4heaA	2.3	0.77
GatCAB	3ip4AC	2gi3A	3ip4C	3.2	0.78
Sulfate adenylyltransferase	1zunAB	1zunA	1jnyA	3.6	0.77
Phenylalanyl-tRNA synthetase	1b70AB	3tupA	1b70B	4.3	0.73
Thiazole synthase/sulfur carrier	1tygAB	1xm3A	1tygB	4.8	0.42
Allophanate hydrolase	3mmlAB	3mmlA	2phcB	12.1	0.72
F1-ATP synthase gamma/epsilon	3oaaHG	1bsnA	3oaaG	16.5	0.50

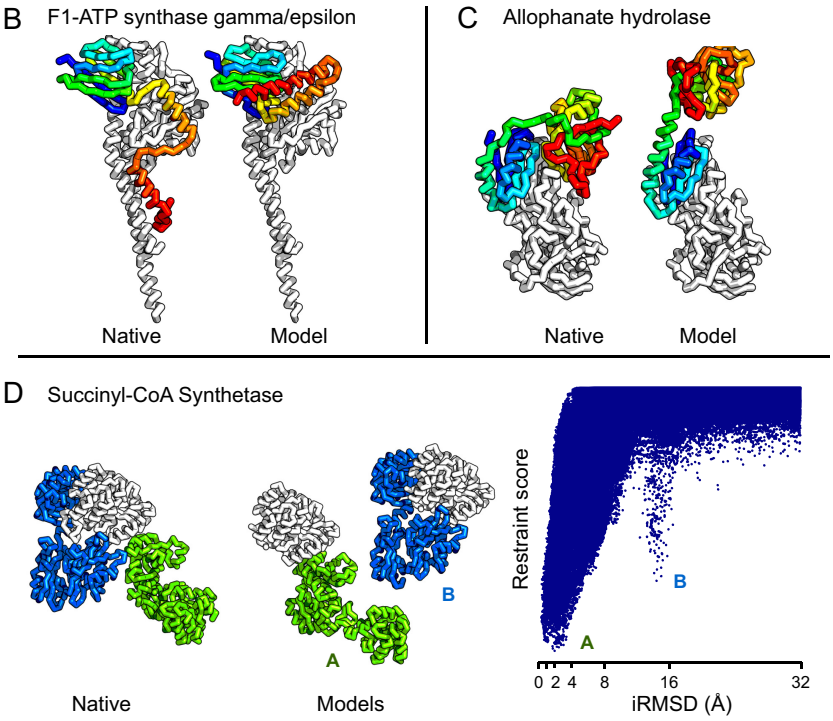


Figure 2.4. Contact guided protein–protein docking on a benchmark set of 18 protein complexes.

(A) Structure models for each complex were generated by docking structures of its constituents, at least one of which (blue) was not from the structure of the complex guided by coevolution derived distance restraints. The interface C-alpha RMSD (iRMSD) of the structural model with the lowest energy to the experimentally determined structure and the fraction of native contacts are shown. Structure models for cases in red are shown in B and C and D. (B and C) Comparison between native and docked structure for the two largest failures in the benchmark: the large iRMSD is due to large conformational changes in the monomers upon docking but the interface

is still modeled correctly in the region not involved in conformational change. (D) Multiple minima in the docking landscape (right) correspond to distinct interfaces in the complex (left).

2.3.3 *Contact predictions for complexes of unknown structure*

The results with the 50S ribosome and the protein pairs in the benchmark suggest that interactions can be accurately predicted across protein–protein interfaces given a sufficient number of aligned sequences. In Figure 3, we provide residue–residue contact predictions for the 36 of the 64 complexes with currently unknown structure (the *E. coli* gene sequences were clustered, and hence each complex may represent multiple *E. coli* gene pairs). These predictions should contribute to the determination of the structures of these biologically important complexes.

2.3.4 *From contacts to structural models*

Are the predicted contacts useful in assembling models of the protein complex from models of each component? We evaluated this on a docking test set containing 18 protein complexes from the benchmark set where at least one component (or a close homolog) had a known structure in the apo form (‘Materials and methods’, docking test-set). We developed a docking protocol that used the predicted contacts as distance restraints and sampled the space of physically plausible structures to generate models of the protein–protein complex. The model with the best restraint score had an interface that was within 4 Å (in root mean square deviation) of the native interface in 14 of the 18 cases and had more than half the native contacts in 16 of the 18 cases (Figure 4A, Figure 4—figure supplement 1). Two of the cases in which the iRMSD (interface root-mean-square deviation) was the highest (bottom of table in A) are illustrated in Figure 4B–C: the high iRMSD is due to large changes in the conformation of one of the monomers upon binding; despite these changes the binding interface is reasonably accurately identified. Conformational changes that hinder the rigid-body docking protocol from sampling the bound conformation also occurred for thiazole synthase/ sulfur carrier and phenylalanyl-tRNA synthase with iRMSD of 4.8A and 4.3A, respectively. In Figure 4D, a second energy minimum corresponds to a second interface in the complex with a different homooligomer subunit. In the absence of conformational changes, predicted contact guided docking is very accurate. The same protocol, on a positive control set of known bound structures of 41 protein-pairs (including 15 protein-

pairs from the NADH electron transport complex), generated models that were within 2 Å of the native complex structure in 38 cases and within 4 Å in all but one case (Figure 4— source data 1, Figure 4—figure supplement 2).

Taken together, these results suggest that in cases with small conformational change, the docking protocol can recover the entire interface to high accuracy and in cases where binding is accompanied by a large conformational change, the protocol recovers the largest intact and/or unobstructed interface.

Of the complexes with unknown structure listed in Figure 3, we selected four cases with two or more high GREMLIN score (≥0.6) contact predictions across the interface that had experimentally determined structures for most of the subunits (‘Materials and methods’) and generated structural models of the complexes. These models provide the basis for formulating hypotheses about the structure/function of the complex, but we emphasize they are not experimentally determined structures; in particular the assumption in the modeling procedure that there are not large backbone rearrangements could be incorrect—in such cases the overall organization of the complex is still likely to be correct but the details of the interfaces could be considerably in error.

2.3.5 *The TRAP complex*

The tripartite ATP-independent periplasmic (TRAP) transporters are composed of three proteins: two integral membrane proteins YIAM and YIAN, and one periplasmic protein YIAO (Mulligan et al., 2011). The structure of the periplasmic domain is known, but the membrane portion is unknown. To generate a model of the three-dimensional structure of the complex, we built YIAM models using Rosetta de novo structure prediction (Simons et al., 1999; Raman et al., 2009) guided by the intramonomer predicted contacts, and models for YIAN and YIAO using RosettaCM comparative modeling. For YIAN the homologous structure of 4f35 (Mancusso et al., 2012) was used. The three monomer structure models were then assembled using PatchDock (Duhovny et al., 2002) and RosettaRelax (Conway et al., 2014) guided by the predicted intersubunit contacts (‘Materials and methods’). In the resultant model of the complex (Figure 5), YIAO interacts with both of the membrane components; this is supported by a number of intersubunit contacts (yellow lines).

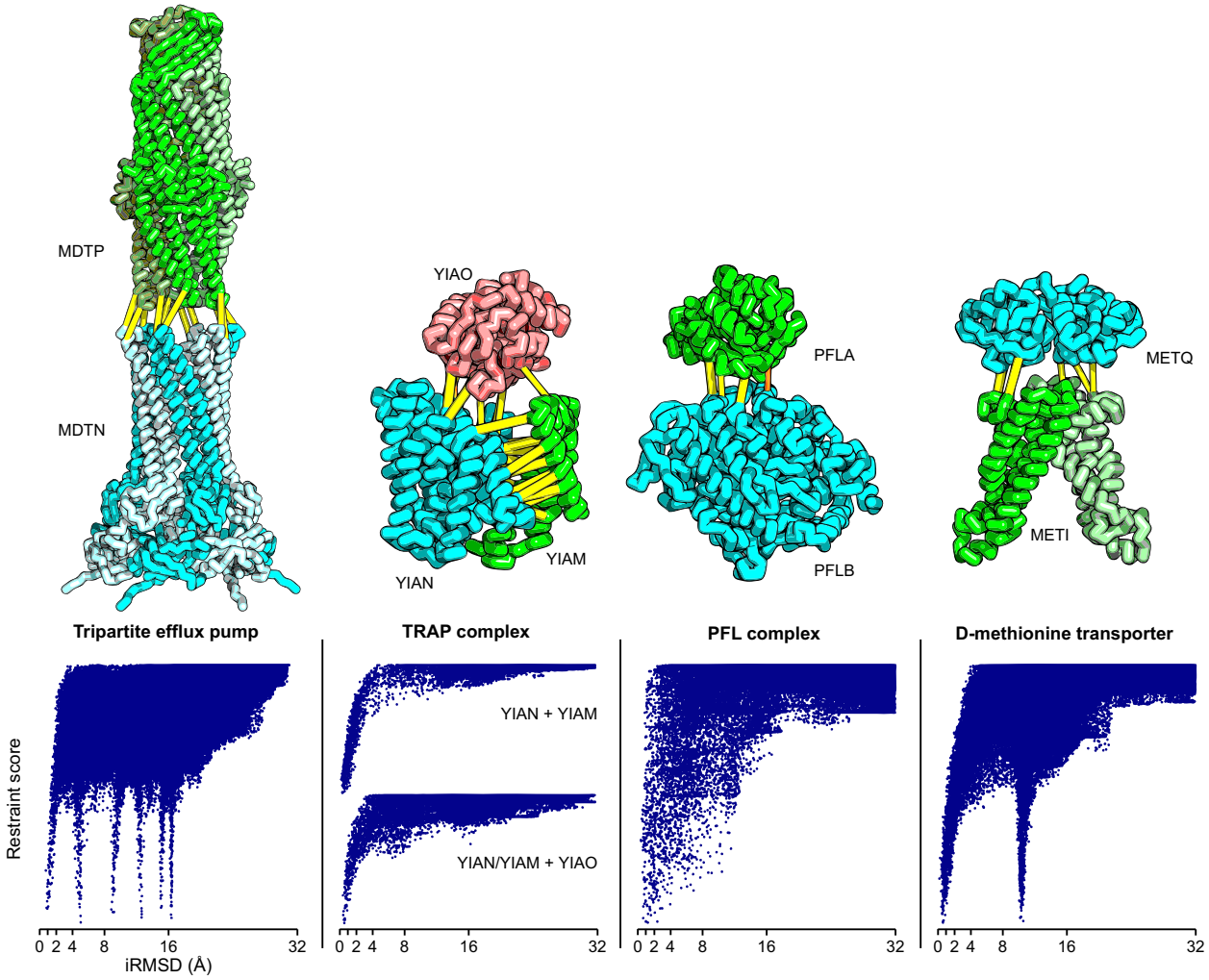


Figure 2.5. Structure models for complexes with unknown structures.
Residue pairs with GREMLIN scores ≥ 0.60 are connected by yellow bars; the structures are pulled apart for clarity. For METQ-METI and PFLA-PFLB GREMLIN scores ≥ 0.3 are shown. For each docking calculation the docking energy landscape is shown, with iRMSD to the selected model on the x-axis. The multiple minima correspond to permutations of the labels on the subunits of the homo-oligomer complex. Predicted structures of each complex are provided with the deposited data.

2.3.6 *Tripartite efflux system*

Tripartite efflux complexes span both the inner and outer membrane, and are widely used in bacteria to pump toxic compounds out of the cell. The mode of interactions between the outer membrane factor and the membrane fusion protein is unresolved, with reports suggesting either a tip-to-tip interaction, the insertion of one into the other, or a multistage interaction with an initial tip-to-tip interaction, followed by sliding one through the channel of the other (Long et al., 2012). We generated homology models for the subunits based on the alignments to 1yc9

(Federici et al., 2005) and 3fpp (Yum et al., 2009) and docked them to generate models of the multidrug resistance protein complex. The predicted residue–residue contacts for this family of complexes support the tip-to-tip interaction (Figure 4; yellow lines); the coevolution data did not provide any evidence to support the insertion model.

2.3.7 *Pyruvate formate lyase-activating enzyme complex*

Pyruvate formate-lyase (PFL) catalyzes the reaction of acetyl-CoA and formate from pyruvate and CoA in the Fermentation pathway. Formate acetyltransferase 1 or Pyruvate formate-lyase 1 (PFLB) is activated by Pyruvate formate-lyase 1-activating enzyme (PFLA). The structure of the complex is unknown, but the structures of the individual proteins have been solved (PDB ids: 3c8f [Becker and Kabsch, 2002] and 1h16 [Vey et al., 2008]). We carried out rigid body docking calculations with these two proteins guided by GREMLIN predictions. Interestingly, the region that undergoes conformational change in the activating enzyme upon substrate binding (3c8f -> 3cb8 [Becker and Kabsch, 2002]) is in the region we predict to be in contact with PFL.

2.3.8 *D-methionine transport system*

D-methionine transporter is an ATP-driven transport system that transports methionine. We docked the *E. coli* structure of METI (3tui, chain A and B, Johnson et al., 2012) with a RosettaCM model of METQ based on 3k2d (Yu et al., 2011). The resulting docked model is consistent with the top ranked GREMLIN predictions (Figure 5).

2.4 DISCUSSION

Our results demonstrate unequivocally that there is strong selective pressure at protein–protein interfaces beyond simple residue conservation, and that co-evolving residue pairs are nearly always in contact in the protein complex. Not all contacting residues across protein interfaces likely co-evolve nor all protein–protein interfaces. Nevertheless, as illustrated in Figures 1 and 2, there is clearly sufficient coevolutionary signal to significantly constrain models of a large number of protein complexes.

There is a notable contrast in the utility of intra-monomer and intersubunit predicted contacts for structure modeling. We found previously (Kamisetty et al., 2013) that contacts could

be predicted with high accuracy for monomeric proteins, provided there were sufficient aligned sequences, but in such cases there was almost always already a structure of a family member from which comparative models could be built, limiting the utility of the predicted contacts in structure prediction (Though predicted contacts can be useful in modeling allosteric changes in protein structures [Hopf et al., 2012; Morcos et al., 2013]). In contrast, here we find that more than half of the complexes for which the protein families of the constituent subunits are sufficiently large for accurate contact prediction do not currently have three-dimensional structures. Hence, while predicted contacts can be very accurate for both monomeric globular proteins and for protein–protein complexes, they are more useful for structure modeling for the latter due to the much poorer representation of protein complexes in the PDB.

While our approach of constructing a global statistical model from paired sequence alignments is generally applicable to any taxa, the current study focuses on prokaryotes and mitochondria. Doing so allows us to largely avoid the problem of distinguishing between paralogs by exploiting the operon architecture of bacterial genomes (Jacob et al., 2005). Constructing paired-sequence alignments for more complex genomic architectures is more involved and requires the ability to distinguish orthologs from paralogs, the subject of active research (Remm et al., 2001; Datta et al., 2009). Protocols for generating paired sequence alignments more generally are an important area for development in this area.

2.5 MATERIALS AND METHODS

2.5.1 *Individual alignment generation*

Multiple sequence alignments were generated for each of the 4303 *E. coli* protein genes as identified by EcoGene 3.0 (Zhou et al., 2013) using HHblits (-n 8 -e 1E-20 -maxfilt ∞ -neffmax 20 -nodiff -realign_max ∞), and HHfilter (-id 100 -cov 75) in the HHsuite (version: 2.0.15, Remmert et al., 2011). To reduce redundancy, we constructed HMMs from each MSA and clustered genes based on the HHΔ (Kamisetty et al., 2013), a measure of HMM–HMM similarity: a pair of genes was assigned to the same cluster if the HHΔ is less than 0.5. This procedure resulted in 2340 non-redundant gene clusters.

For the benchmark set, a new alignment was generated using the sequence associated with each PDB. For the 50S ribosome and NADH dehydrogenase, we used Thermus

thermophilus HB8 sequences from PDB structures 3uxr (Bulkley et al., 2012) and 4hea (Baradaran et al., 2013) respectively. For paralogous NADH dehydrogenase chains L, M, and N, we used an e-value of 1E-60 in the alignment generation protocol. In addition to complexes from the *E. coli* analysis, we also include the GatCAB amidotransferase complex in our benchmark set, using sequences from the PDB structure 3ip4 (Nakamura et al., 2010). For cases where the PDB sequence length was much longer than average coverage, we modified the coverage filter to 50% of query. The sequences were then realigned using clustal omega v1.2 (--iterations 2 --full-iter) (Sievers et al., 2011). Residues not present in the query sequence were dropped from subsequent analysis.

Paired alignment generation We construct alignments of paired protein sequences $[x_1, x_2, \ldots, x_p; x_{p+1}, \ldots, x_{p+q}]$ from the same genome with positions 1:p and p+1:p+q corresponding to the first and second proteins respectively. We refer to such a multiple sequence alignment of paired sequences as a paired alignment.

For gene families with a single copy in each genome such as the ribosomal proteins, constructing paired alignments is straightforward as sequence pairs from the same genome can simply be concatenated. While the process of generating paired alignments in general is complicated in the presence of multiple paralogs of a gene in a single genome, in prokaryotes, co-regulated genes are often co-located on the genome into operons. We exploit this property to avoid paralogous genes when creating paired sequences by restricting to gene pairs that have small, conserved intergenic distances. A similar approach was used to construct a database of fusion proteins in prokaryotic genomes (Suhre and Claverie, 2004). Defining Δgene as the number of annotated genes between a gene pair, we only consider pairs with Δgene conserved in 60% of genomes and less than 20. To allow for ambiguity in annotation, if the second or third most common intergenic distance is within 1 of the mode, these gene-pairs are included in the conservation calculation. Given that most UniProt accession IDs are serially assigned in a genome (UniProt Accession), Δgene can be rapidly evaluated by looking at the difference in accession ids. The paired alignment is then filtered to reduce redundancy to 90% sequence identity and to remove positions that have more than 75% gaps.

2.5.2 *Identification of protein complex structures*

To identify protein pairs in the same complex structure, a HMM was constructed for each *E. coli* protein using hmmbuild from the already generated HHblits alignments. We then used hmmsearch to scan PDB sequences in the S2C database (Wang et al.; Both hmmbuild and hmmsearch are part of the HMMER v3.1b package [Eddy, 2009]). Only hits with e-value less than 1E-10 were considered. Protein pairs found in the same complex structure (PDB file) were considered to be in contact if a Cα atom in one structure was within 12 Angstroms of a Cα atom in the other.

2.5.3 *Gremlin model construction from paired alignments*

GREMLIN constructs a global statistical model of the paired alignment, assigning a probability to every amino-acid sequence in the paired alignment:

$$p(X_1, X_2 \ldots, X_p; X_{p+1} \ldots X_{p+q}) = \frac{1}{Z}\exp(\Sigma_1{}^{p+q}[v_i(X_i) + \Sigma_{j=1}{}^{p+q} w_{i,j}(X_i, X_j)])$$

where, the v_i are vectors encoding position-specific amino-acid propensities and the $w_{i,j}$ are matrices encoding amino-acid coupling between positions i and j. These parameters are obtained from the aligned sequences by maximizing the regularized pseudo-likelihood (Balakrishnan et al., 2011) of the alignment as described in (Kamisetty et al., 2013):

$$v, w = \arg\max \Sigma_1{}^{N}\Sigma_1{}^{p+q} \log P(X_i | X_1 .. X_{i-1} X_{i+1} .. X_{p+q}) + R(v, w)$$

where, each term in the summation is a conditional distribution capturing the probability of a particular amino-acid at a position in the context of the entire protein sequence and R(v,w) is a regularization term to prevent over-fitting.

Previous approaches (Morcos et al., 2011; Jones et al., 2012) estimated v, w using an approximate moment matching approach (Kamisetty et al., 2013) by inverting a generalized covariance matrix. These rely on a Gaussian-like approximation to the global partition function. Unlike these approaches, estimation via the pseudo-likelihood avoids this approximation relying instead on local partition functions (Balakrishnan et al., 2011; Ekeberg et al., 2013; Kamisetty et al., 2013). The resulting global optimization problem can be efficiently solved using standard

convex optimization techniques and provides estimates for each vector vi and matrix wij (Kamisetty et al., 2013).

2.5.4 *Ranking residue pairs with gremlin scores*

To reduce the w_{ij} matrices to single values reflecting the strength of the coupling between positions i and j, we first compute s_{ij}, their vector 2-norm (the square root of the averages of the squares of the individual matrix elements). We correct for differences in s_{ij} due to sequence variability at different positions using the row and column averages of these values:

$$s_{ij}{}^{corr} = \frac{s_{ij} - < s_{kj} >_k < s_{ik} >_k}{< s_{kl} >_{kl}}$$

where brackets indicate averages taken over the indices outside the brackets in a manner similar to that of Average Product Correction (APC, Dunn et al., 2008). Unlike the APC, we account for differences in the rates of evolution in the two protein families by computing the averages only over the positions of the proteins corresponding to positions i and j: if i and j are both in the first (second) protein, the averages are computed over the positions in the first (second) protein; if i is in the first protein and j in the second, the column average is computed only over the positions of the first protein and the row average, only over the positions of the second protein. We then compute a normalized coupling strength, ncs_{ij}, by dividing the $s_{ij}{}^{corr}$ by the average of the top 3L/2 $s_{ij}{}^{corr}$ values across the two proteins (since there are roughly 3L/2 contacts for a protein of length L [Kamisetty et al., 2013; SI]).

As illustrated in Figure 1D, the relation between normalized coupling strength and contact frequency varies with the ratio of the number of aligned sequences to the length of the protein complex. We also observed that residues were more frequently in contact for a given coupling strength when the top score for that complex was high. To account for these dependencies, we constructed a model that estimates the probability of being in contact based on the bacterial 50S ribosomal complex:

$$\mathrm{GremlinScore}(x, N/L) = 1/(1 + \exp(-\sigma(x - \mu))$$

where

$$\mu = m^{N/L + 1} + c$$

and x is $\sqrt{ncs_{ij}}$ for the top scoring contact in each complex and $\sqrt{ncs_{ij}}$ scaled by the Gremlin score of the top contact in all other cases. The values of m, c, and σ (0.47, 0.96, and 9.77 respectively) were determined by a non-linear fit to the observed frequencies in the 50S

ribosomal data from Figure 1D. This function accurately accounts for the observed contact frequencies (Figure 1—figure supplement 1).

Conversion of gremlin scores to distance restraints We converted coupling strengths into residue-pair specific distance restraints and included them in the Rosetta structure prediction program. We use sigmoidal distance restraints of the form:

$$\text{restraint(d)} = \frac{\text{weight}}{1 + \exp(-\text{slope}(\text{d} - \text{cutoff}))} + \text{intercept}$$

where, d is the distance between the constrained atoms and the weight is proportional to ncsij. The restraints were introduced between Cβ atoms (Cα in the case of glycine) in the reduced-atom representation of Rosetta (centroid mode) and as ambiguous distance restraints (Lange et al., 2012) between side-chain heavy atoms (cutoff of 5.5 and slope of 4) in the full-atom stage of Rosetta. For the centroid mode, restraints used the amino acid pair specific C β-C β cutoff and slopes, as described in Kamisetty et al., 2013 SI Table III. These distance restraints supplement the Rosetta all atom energy; the combination ensures the sampling of physically realistic structures consistent with the contact predictions.

2.5.5 *Comparative modeling*

Comparative models were built using RosettaCM (Song et al., 2013) based on alignments to homologous structures generated using HHsearch (Remmert et al., 2011). For proteins that had missing density in regions predicted to be in contact, we used RosettaCM with co-evolution derived restraints to build the missing region before docking.

2.5.6 *De Novo modeling*

The Rosetta ab initio protocol consists of two stages: in the initial stage (‘centroid’) side-chains are represented by fixed center-of-mass atoms allowing for rapid generation and evaluation of various protein-like topologies; the second stage (‘full-atom’) builds in explicit side-chains and carries out all atom energy minimization (Simons et al., 1999; Raman et al., 2009). YIAM, a membrane protein, was modeled with the Rosetta membrane energy function (Yarov-Yarovoy et al., 2012, Barth et al., 2007). Strong repulsive interactions (Equation 1, weight: −100, cutoff: 35, slope: 2 and intercept: 100) were added between the center of the extracellular regions and the center of predicted intracellular regions, and strong attractive restraints (weight:100, cutoff:35,

slope:2 and intercept: 0) within predicted intracellular regions and extracellular regions, effectively constructing a membrane-like sampling space. We used the consensus output of MESSA (Cong and Grishin, 2012) to predict transmembrane regions. 100,000 models were generated and 20 models that best fit the restraints converged to a single cluster.

2.5.7 *Docking test set*

Jackhammer (part of HMMER v3.1b package; Eddy, 2009) was used to identify a subset of 18 complexes in the benchmark set where at least one of the proteins or a close homolog had a solved structure of its apo form. In cases where the structure was of a homologous protein (e-value < 1E-20) and where most of the interface residues were present, we generated a structural model of the target protein using comparative modeling. We only considered cases where at least one of the structures was unbound as the bound–bound docking problem is not representative of real world docking challenges (Betts and Sternberg, 1999). The positive control shown in Figure 4—source data 1 was run on all protein-pairs from the benchmark set, where at least two predicted inter contacts had a high GREMLIN score (>0.6).

2.5.8 *Complex assembly by protein–protein docking*

For each inter restraint pair that is in the top 3/2L predictions, we used PatchDock v1.0, with clustering parameters (rmsd 0.5; discardClustersSmaller 0) (Duhovny et al., 2002) to generate an ensemble of conformations that were then scored using all the restraints. For tripartite efflux pump, the surface segmentation parameters were further modified (low_patch_thr 0; prune_thr 0.1; flat 1), to allow for more diverse interfaces. The top 5 models by restraint score were energy-minimized in cartesian space using both inter and intra restraints with cycles of minimization and side chain repacking using Rosetta as described in Conway et al. (2014). The best scoring model by restraint score was then selected.

For fraction of native contact (Fnat) and interface root-mean-squared deviation (iRMSD) calculation, the interface residue–residue contacts are those where the minimal distance between any heavy side-chain atom is less than 5 Å. The Fnat calculation is performed as described in Kamisetty et al. (2013) SI Table III.

All structural figures were drawn with PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

2.6 Data Availability

The multiple sequence alignments used in the analysis and the full GREMLIN results for all the calculations described in the paper are provided at http://gremlin.bakerlab.org/complexes/ along with a web-server for paired-alignment generation, coevolution analysis and contact prediction/Rosetta restraint generation. The paired-alignments along with the PDB coordinates of the predicted structures are also available at Dryad: Ovchinnikov et al., 2014.

2.7 Acknowledgements

We thank Lei Shi and David La for their comments and helpful suggestions, and Rosetta@home participants for donating their computer time.

2.8 References

1. Balakrishnan S, Kamisetty H, Carbonell JG, Lee Su-I, Langmead CJ. 2011. Learning generative models for protein fold families. Proteins: structure, Function, and Bioinformatics 79:1061–1078.
2. Baradaran R, Berrisford JM, Minhas GS, Sazanov LA. 2013. Crystal structure of the entire respiratory complex I. Nature 494:443–448.
3. Barth P, Schonbrun J, Baker D. 2007. Toward high-resolution prediction and design of transmembrane helical protein structures. Proceedings of the National Academy of Sciences of the United States of America 104:15682–15687.
4. Becker A, Kabsch W. 2002. X-ray structure of pyruvate formate-lyase in complex with pyruvate and CoA. How the enzyme uses the Cys-418 thiyl radical for pyruvate cleavage. Journal of Biological Chemistry 277:40036–40042.
5. Betts MJ, Sternberg MJE. 1999. An analysis of conformational changes on protein–protein association: implications for predictive docking. Protein Engineering 12:271–283.
6. Bulkley D, Johnson F, Steitz TA. 2012. The antibiotic thermorubin inhibits protein synthesis by binding to inter-subunit bridge B2a of the ribosome. Journal of Molecular Biology 416:571–578.
7. Burger L, van Nimwegen E. 2008. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. Molecular Systems Biology 4:165.
8. Cong Q, Grishin NV. 2012. MESSA: MEta-server for protein sequence analysis. BMC Biology 10:82.
9. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D. 2014. Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Science 23:47–55.
10. Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. 2012. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proceedings of the National Academy of Sciences of the United States of America 109:E1733–E1742.
11. Datta RS, Meacham C, Samad B, Neyer C, Sjölander K. 2009. Berkeley PHOG: PhyloFacts orthology group prediction web server. Nucleic Acids Research 37:W84–W89.
12. de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. Nature Reviews Genetics 14:249–261.
13. Duhovny D, Nussinov R, Wolfson HJ. 2002. Efficient unbound docking of rigid molecules. In: Berlin, editor. Algorithms in bioinformatics. Springer: Heidelberg. p. 185–200.

14. Dunn SD, Wahl LM, Gloor GB. 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics 24:333–340.
15. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Informatics 23:205–211.
16. Ekeberg M, Lövkvist C, Lan Y, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Physical Review E 87:012707.
17. Federici L, Du D, Walas F, Matsumura H, Fernandez-Recio J, McKeegan KS, Borges-Walmsley MI, Luisi BF, Walmsley AR. 2005. The crystal structure of the outer membrane protein VceC from the bacterial pathogen Vibrio cholerae at 1.8 Å resolution. Journal of Biological Chemistry 280:15307–15314.
18. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. 2003. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. Journal of Molecular Biology 331:281–299.
19. Halperin I, Wolfson H, Nussinov R. 2006. Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins: structure, Function, and Bioinformatics 63:832–845.
20. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149:1607–1621.
21. Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Sander C, Bonvin AMJJ, Marks DS. 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. bioRxiv.
22. Hosur R, Peng J, Vinayagam A, Stelzl U, Xu J, Perrimon N, Bienkowska J, Berger B. 2012. A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. Genome biology 13:R76.
23. Jacob F, Perrin D, Sánchez C, Monod J. 2005. L'opéron: groupe de gènes à expression coordonnée par un opérateur [CR Acad. Sci. Paris 250 (1960) 1727–1729]. Comptes Rendus Biologies 328:514–520.
24. Johnson E, Nguyen PT, Yeates TO, Rees DC. 2012. Inward facing conformations of the MetNI methionine ABC transporter: Implications for the mechanism of transinhibition. Protein Science 21:84–96.
25. Jones DT, Buchan DWA, Cozzetto D, Pontil M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184–190.
26. Kamisetty H, Ovchinnikov S, Baker D. 2013. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences of the United States of America 110:15674–15679.
27. Lange OF, Rossi P, Sgourakis NG, Song Y, Hsiau-Wei L, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D. 2012. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proceedings of the National Academy of Sciences of the United States of America 109:10873–10878.
28. Lapedes A, Giraud B, Jarzynski C. 2012. Using sequence alignments to predict protein structure and stability with high accuracy. arXiv 1207.2484. http://arxiv.org/abs/1207.2484
29. Lapedes AS, Giraud BG, Liu LC, Stormo GD. 1999. Correlated mutations in models of protein sequences: phylogenetic and structural effects. Lecture Notes-Monograph Series 33:236–256.
30. Long F, Su CC, Lei HT, Bolla JR, Do SV, Yu EW. 2012. Structure and mechanism of the tripartite CusCBA heavy-metal efflux complex. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 1047–1058.
31. Mancusso R, Gregorio GG, Liu Q, Wang Da-N. 2012. Structure and mechanism of a bacterial sodium-dependent dicarboxylate transporter. Nature 491:622–626. doi: 10.1038/nature11542.
32. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. PLOS ONE 6:e28766.
33. Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. Nature Biotechnology 30:1072–1080.
34. Morcos F, Jana B, Hwa T, Onuchic JN. 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proceedings of the National Academy of Sciences of the United States of America 110:20533–20538.
35. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America 108:E1293–E1301.
36. Mulligan C, Fischer M, Thomas GH. 2011. Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea. FEMS Microbiology Reviews 35:68–86.
37. Nakamura A, Sheppard K, Yamane J, Yao M, Söll D, Tanaka I. 2010. Two distinct regions in Staphylococcus aureus GatCAB guarantee accurate tRNA recognition. Nucleic Acids Research 38:672–682.

38. Nugent T, Jones DT. 2012. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proceedings of the National Academy of Sciences of the United States of America 109:E1540–E1547.
39. Ochoa D, Pazos F. 2010. Studying the co-evolution of protein families with the Mirrortree web server. Bioinformatics 26:1370–1371.
40. Ovchinnikov S, Kamisetty H, Baker D. 2014. Data from: Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. Dryad Digital Repository.
41. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. Correlated mutations contain information about protein-protein interaction. Journal of Molecular Biology 271:511–523.
42. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins: structure, Function, and Bioinformatics 77:89–99.
43. Remm M, Storm CEV, Sonnhammer ELL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology 314:1041–1052.
44. Remmert M, Biegert A, Hauser A, Söding J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9:173–175.
45. Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. 2009. High-resolution protein complexes from integrating genomic information with molecular simulation. Proceedings of the National Academy of Sciences of the United States of America 106:22124–22129.
46. Shoemaker BA, Panchenko AR. 2007. Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLOS Computational Biology 3:e43.
47. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539.
48. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. 1999. Improved recognition of nativelike protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins: structure, Function, and Bioinformatics 34:82–95.
49. Song Y, DiMaio F, Wang RYR, Kim D, Miles C, Brunette TJ, Thompson J, Baker D. 2013. High-Resolution comparative modeling with RosettaCM. Structure 21:1735–1742.
50. Suhre K, Claverie JM. 2004. FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. Nucleic Acids Research 32:D273–D276.
51. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. 2012. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences of the United States of America 109:10340–10345.
52. Tamir S, Rotem-Bamberger S, Katz C, Morcos F, Hailey KL, Zuris JA, Wang C, Conlan AR, Lipper CH, Paddock ML, Mittler R, Onuchic JN, Jennings PA, Friedler A, Nechushtai R. 2014. Integrated strategy reveals the protein interface between cancer targets Bcl-2 and NAF-1. Proceedings of the National Academy of Sciences of the United States of America 111:5177–5182.
53. Thomas J, Ramakrishnan N, Bailey-Kellogg C. 2008. Graphical models of residue coupling in protein families. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 5:183–197.
54. UniProt Accession. UniProt User manual. http://www.uniprot.org/manual/accession_numbers. accessed September 9, 2013.
55. Valencia A, Pazos F. 2002. Computational methods for the prediction of protein interactions. Current Opinion in Structural Biology 12:368–373.
56. Vey JL, Yang J, Li M, Broderick WE, Broderick JB, Drennan CL. 2008. Structural basis for glycyl radical formation by pyruvate formate-lyase activating enzyme. Proceedings of the National Academy of Sciences of the United States of America 105:16137–16141.
57. Wang G, Dunbrack RL Jr. S2C: a database correlating sequence and atomic coordinate residue numbering in the Protein Data Bank. Dunbrack Lab. http://dunbrack.fccc.edu/s2c/. Accessed October 12, 2013.
58. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein–protein interaction by message passing. Proceedings of the National Academy of Sciences of the United States of America 106:67–72.
59. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Research 34:D187–D191.

60. Yarov-Yarovoy V, DeCaen PG, Westenbroek RE, Pan CY, Scheuer T, Baker D, Catterall WA. 2012. Structural basis for gating charge movement in the voltage sensor of a sodium channel. Proceedings of the National Academy of Sciences of the United States of America 109:E93–E102.
61. Yu S, Yeon Lee N, Park SJ, Rhee S. 2011. Crystal structure of toll-like receptor 2-activating lipoprotein IIpA from Vibrio vulnificus. Proteins: structure, Function, and Bioinformatics 79:1020–1025.
62. Yum S, Xu Y, Piao S, Sim SH, Kim HM, Jo WS, Kim KJ, Kweon HS, Jeong MH, Jeon H, Lee K, Ha NC. 2009. Crystal structure of the periplasmic component of a tripartite macrolide-specific efflux pump. Journal of Molecular Biology 387:1286–1297.
63. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490:556–560.
64. Zhou J, Rudd KE. 2013. EcoGene 3.0. Nucleic Acids Research 41:D613–D624.

2.9 SUPPLEMENTAL FIGURES

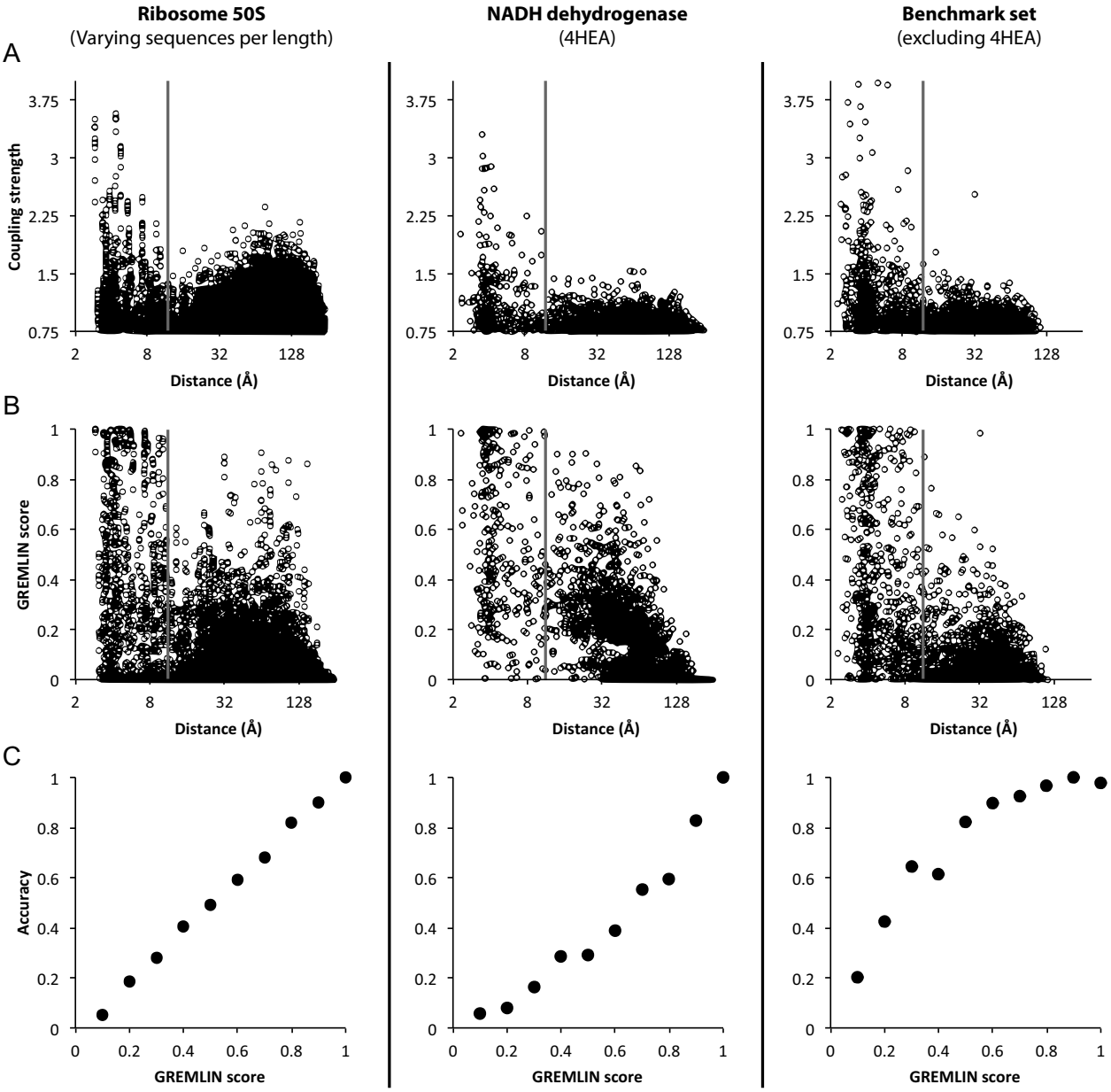


Figure 1—figure supplement 1. Determining GREMLIN scores from normalized coupling strengths. Top row: (**A**) Normalized Coupling strengths. (**B**) GREMLIN score obtained by fitting a sigmoidal function of normalized coupling strengths to observed frequencies on the 50S ribosome (left column) evaluated on the benchmark set (complexes from the NADH dehydrogenase, middle column and the remaining, right column). (**C**) The GREMLIN score is well-calibrated: the fraction of predictions with a Gremlin score of x that are correct (distance <12 Å) is roughly x (x in [0, 1]). The overall behavior is similar across the three datasets.

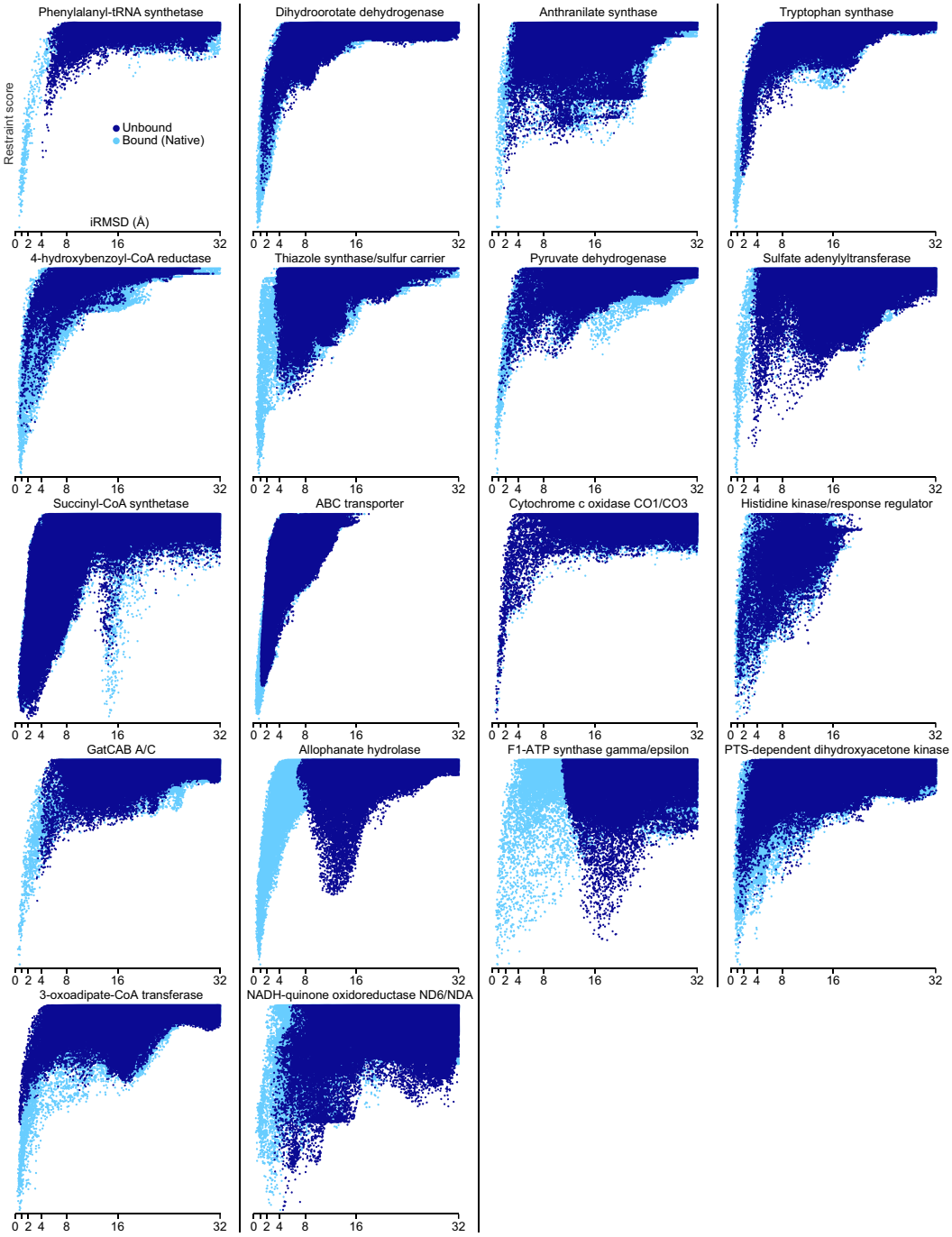


Figure 4—figure supplement 1. Docking landscapes showing iRMSD (x-axis) vs GREMLIN restraint score (y-axis). Each point represents a structure model generated by docking the subunits guided by the GREMLIN score. Dark blue points are from calculations in which at least one subunit was solved independently of the complex; light blue points, from positive control calculations in which both subunits are from the bound complex.

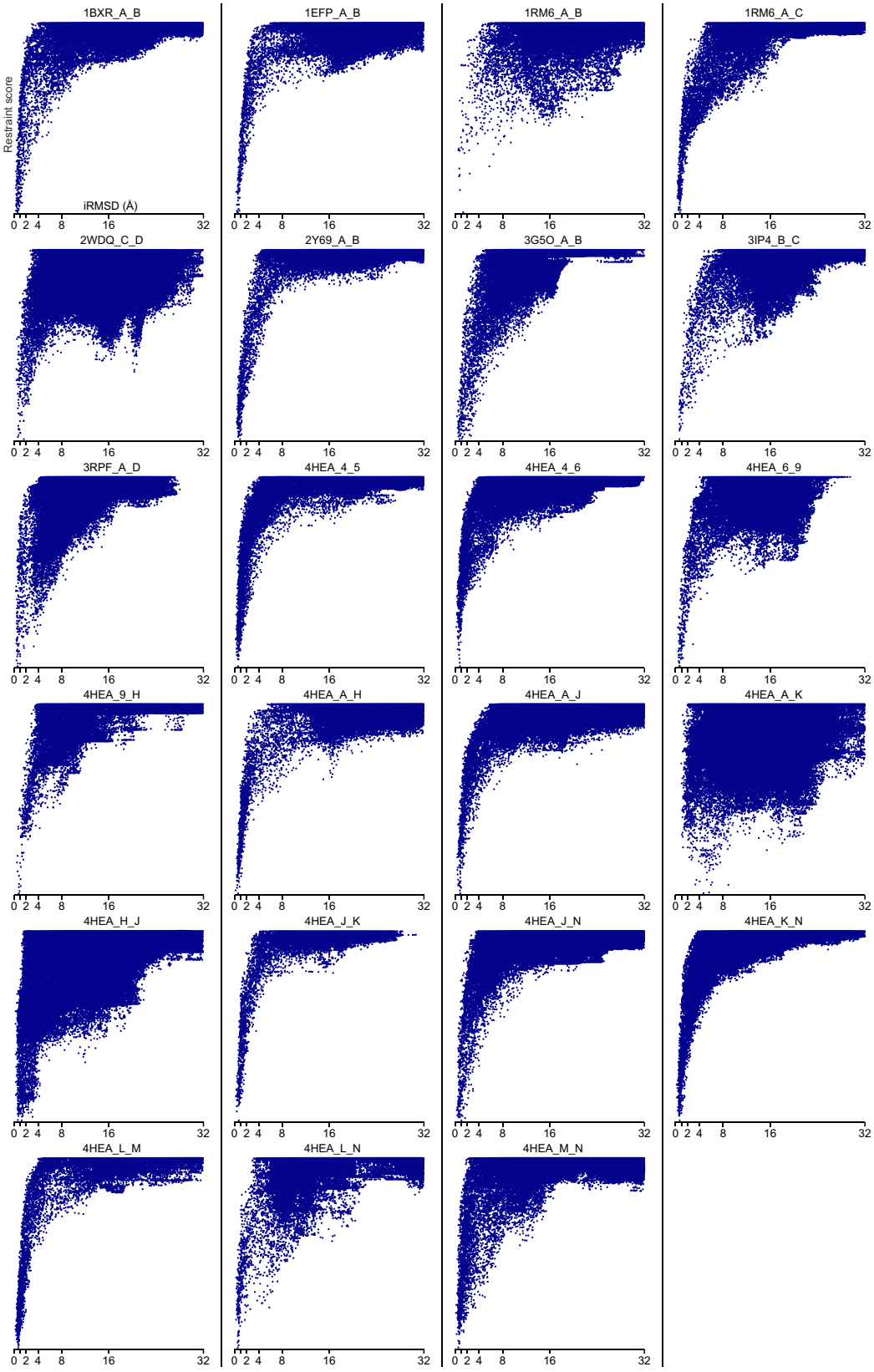


Figure 4—figure supplement 2. Bound set. Docking landscapes with GREMLIN restraint score. X-axis, iRMSD; y-axis GREMLIN restraint score.

Chapter 3. LARGE SCALE DETERMINATION OF PREVIOUSLY UNSOLVED PROTEIN STRUCTURES USING EVOLUTIONARY INFORMATION

A version of this chapter has been previously published as:

Ovchinnikov, Sergey, Lisa Kinch, Hahnbeom Park, Yuxing Liao, Jimin Pei, David E. Kim, Hetunandan Kamisetty, Nick V. Grishin, and David Baker. "Large-scale determination of previously unsolved protein structures using evolutionary information." *Elife* 4 (2015): e09248.

3.1 ABSTRACT

The prediction of the structures of proteins without detectable sequence similarity to any protein of known structure remains an outstanding scientific challenge. Here we report significant progress in this area. We first describe de novo blind structure predictions of unprecendented accuracy we made for two proteins in large families in the recent CASP11 blind test of protein structure prediction methods by incorporating residue–residue co-evolution information in the Rosetta structure prediction program. We then describe the use of this method to generate structure models for 58 of the 121 large protein families in prokaryotes for which three-dimensional structures are not available. These models, which are posted online for public access, provide structural information for the over 400,000 proteins belonging to the 58 families and suggest hypotheses about mechanism for the subset for which the function is known, and hypotheses about function for the remainder.

3.2 INTRODUCTION

Despite substantial efforts over decades, high-resolution structure prediction is currently limited to proteins that have homologs of known structure, or small proteins where thorough sampling of the conformational space is possible (<100 residues; even in this case, predictions can be very inaccurate). For roughly 41% of protein families, there is currently no member with known structure (Kamisetty et al., 2013). While high-resolution ab initio structure prediction has remained a challenge, considerable success has been achieved in generating high-accuracy models when sparse experimental data are available to constrain the space of conformations to be sampled. This additional information, in combination with a reasonably accurate energy function, has enabled the determination of high-resolution structures for much larger proteins (Raman et al., 2009; DiMaio et al., 2011; Lange et al., 2012).

Recent work has shown that residue–residue contacts can be accurately inferred from co-evolution patterns in sequences of related proteins (Marks et al., 2011; Morcos et al., 2011; Hopf et al., 2012; Jones et al., 2012; Marks et al., 2012; Nugent and Jones, 2012; Sułkowska et al., 2012; Kamisetty et al., 2013). While early approaches estimated these restraints by inverting a covariance matrix (Marks et al., 2011; Morcos et al., 2011; Jones et al., 2012), subsequent

studies have shown that a pseudo-likelihood (PLM)-based approach (Balakrishnan et al., 2011) results in more accurate predictions (Ekeberg et al., 2013; Kamisetty et al., 2013). Distance restraints derived from such predictions have been used to model a wide range of unknown protein structures (Hayat et al., 2014; Wickles et al., 2014; Abriata, 2015; Antala et al., 2015; Hopf et al., 2015; Tian et al., 2015) and protein–protein complexes (Ovchinnikov et al., 2014; Hopf et al., 2014). However, while the generated structures often recapitulate the fold of the target protein, it has not been clear whether such methods can yield high-accuracy models of complex protein structures.

3.3 RESULTS

3.3.1 *CASP11 predictions*

In the recent CASP11 (11th critical assessment of techniques for protein structure prediction) blind test of protein structure prediction methods, we predicted the structures of proteins from large families with no representatives of known structure by integrating co-evolution derived contact information from GREMLIN (Kamisetty et al., 2013) into the Rosetta structure prediction methodology (Simons et al., 1999; Rohl et al., 2004; Raman et al., 2009). Starting from an extended polypeptide chain, Monte Carlo + Minimization searches through conformations with local structure consistent with the local sequence were carried out, optimizing first a low-resolution energy function favoring hydrophobic burial and backbone hydrogen bonding, and second a detailed all atom energy function describing hydrogen bonding and electrostatic interactions, van der Waals interactions, and solvation (Das and Baker, 2008). In the first phase, sampling was carried out in internal coordinates (the backbone torsion angles), and hence, to avoid loss of sampling efficiency by early formation of contacts between residues distant along the sequence, predicted contact information was first added for residues close along the chain and subsequently for residues with increasing sequence separation. The contact information was implemented through residue–residue distance restraints whose strength and shape were functions of the strength of the evolutionary covariance between the residues (see ‘Materials and methods’). Large numbers of independent trajectories were carried out using the Rosetta@Home distributed computing project, and the lowest energy (Rosetta all atom energy plus evolutionary restraint fit) models were recombined and further optimized using a new

iterative version (see 'Materials and methods') of the RosettaCM hybridization protocol (Song et al., 2013; Kim et al., 2014). The five lowest energy structures were submitted as predictions to the CASP organizers. When several months later the actual structures of these proteins were revealed, the predictions were found to be considerably more accurate than any previous predictions made in the 20 years of CASP experiments for proteins over 100 amino acids that lack homologs of known structure. Two particularly striking examples are shown in Figure 1; the prediction for the complex 256 residue structure of T0806 is 3.6 Cα-RMSD from the crystal structure (2.9 Å over 223 residues), and the prediction for the 108 residue T0824 is 4.2 Cα-RMSD from the crystal structure (2.7 Å over 77 residues). The models accurately recapitulate the complex topologies of the proteins. Due to time restraints, the calculations could not be run to convergence during CASP; with additional sampling, the lowest energy model for T0806 has an RMSD of 2.1 Å over 245 residues to the experimentally determined structure. Both the co-evolution derived contacts and the new iterative hybridization protocol were critical to obtaining higher accuracy models: Rosetta calculations without constraints failed to converge (data not shown), and the ROBETTA server models generated without the hybridization step were considerably less accurate (11.6 vs 3.6 Cα- RMSD for T0806 and 14.0 vs 4.2 Cα-RMSD for T0824).

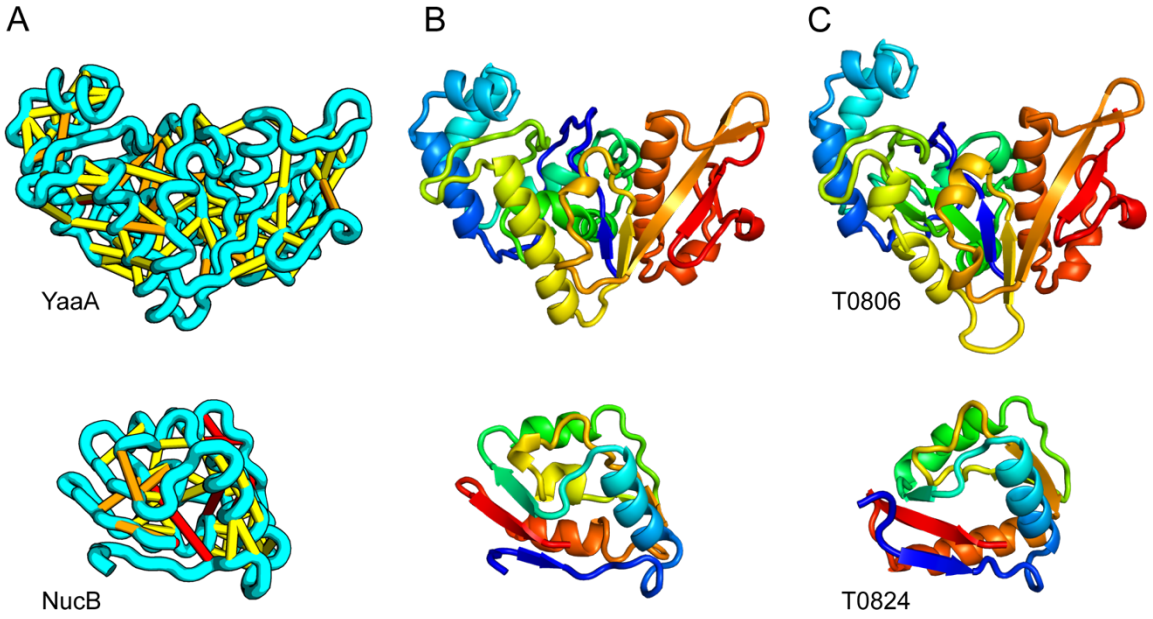


Figure 3.1. Accurate blind structure prediction of CASP11 targets T0806 and T0824.
(A) Location of the most strongly co-evolving residue pairs. Lines connect residue pairs with normalized coupling strength greater than 1.0; yellow, distance less than 5 A; orange, distance less than 10Å and red, greater than 10Å in the models. (B) CASP11 submitted models, colored

from N to C terminus (blue to red). (C) X-ray crystal structures. For T0806, the Cα RMSD over the full-length protein is 3.6Å and 2.9Å over 223 aligned residues. For T0824; the Cα RMSD over the full-length protein is 4.2Å and 2.7Å over 77 aligned residues. For statistics on all five models submitted during CASP, see Figure 1—source data 1

	T0806			T0824		
Model	Cα-RMSD	GDT-TS	Score	Cα-RMSD	GDT-TS	Score
1	**3.6**	**0.60**	**316.1**	4.4	0.52	102.9
2	3.9	0.58	298.7	**4.2**	**0.55**	**105.3**
3	4.9	0.58	282.8	9.5	0.34	87.9
4	7.5	0.34	242.6	8.5	0.34	84.9
5	7.6	0.36	245.8	12.3	0.22	87.8

Figure 3.1—source data 1. The five models submitted to CASP for targets T0806 and T0824. The Cα-RMSD and GDT-TS calculations are over the full-length sequence. The total GREMLIN score for the model is reported. The most accurate models have the best GREMLIN score.

3.3.2 *Prediction of structures for large protein families*

Having found that protein structures can accurately be modeled using co-evolution information, we set out to build models for representatives of large protein families in bacteria with no detectable structural homologs. To facilitate evaluation of such models, we developed a length-independent measure of the fit between a set of predicted contacts and a model: the ratio of the total GREMLIN score of the model to the score expected if it were the native structure (Rc, see 'Materials and methods').

We chose to focus on families with at least 4× (protein length) sequences to ensure that the predicted contacts have high accuracy (Kamisetty et al., 2013; at 4L sequences, the top 1.5L contacts are on average 50% correct). Families with detectable structure homologs were excluded using a sensitive sequence search method (HHsearch [Soding, 2005]). For computational efficiency, an initial scan was done using a single sequence, excluding families where the top hit had an e-value of 1 or greater to any protein of known structure. We identified 100 families satisfying these criteria in *Escherichia coli* (Gram-negative, Proteobacteria), and an additional 22, 5, and 4 families in *Bacillus subtilis* (Gram-positive, Firmicutes bacterium), *Halobacterium salinarum* (Euryarchaeota), and *Sulfolobus solfataricus* (Crenarchaeota), respectively (Supplementary file 1). For each of these top families, we carried out a more sensitive profile–profile sequence search against the Protein Data Bank (PDB) using HHsearch (Soding, 2005) and the fold recognition method SPARKS-X (Yang et al., 2011). We eliminated

families if the top HHsearch hit had an E-value less than 1E-04 and was consistent with GREMLIN contacts.

An alternative approach to structure modeling using predicted contacts is to search for weak fold recognition matches to known protein structures and determine if any of the hits fit the predicted contacts. This approach is not very effective for the families identified as described above; for only 4 of the 122 families with HHsearch E-values greater than 1E-04 did one of the top ten hits from HHsearch or SPARKS-X match the predicted contacts (have Rc values greater than 0.6).

Table 3.1. Transmembrane protein benchmark.
Column 1, PDB code (resolution of the crystal structure); column 2, protein name; column 3, sequences per length, after filtering to reduce the redundancy to 90%; column 4, RMSD of predicted structure to native structure; column 5, length of native structure modeled; column 6, RMSD over converged and constrained region; column 7, length of converged and constrained region; column 8, RMSD over TM-align structural alignment; column 9, length of structurally aligned region.

PDB	Name	seq/len	Full protein		Converged		Aligned	
			rmsd	length	rmsd	length	rmsd	length
4HE8_H (3.3)	NADH-quinone oxidoreductase subunit 8	17.3	4.9	269	2.1	183	2.2	234
1SOR_A (N/A)	Aquaporin-0	26.2	2.7	221	2.1	188	2.0	200
4Q2E_A (3.4)	Phosphatidate cytidylyltransferase	18.6	5.4	262	3.5	176	2.8	178
4HTT_A (6.8)	Sec-independent protein translocase protein	14.6	3.9	225	1.8	124	2.4	181
4P6V_E (3.5)	Na(+)-translocating NADH-quinone reductase subunit D	14.3	5.0	194	1.4	49	2.8	155
4J72_A (3.3)	Phospho-N-acetylmuramoyl-pentapeptide-transferase	19.9	6.6	323	3.1	251	2.4	237
3V5U_A (1.9)	Sodium/Calcium Exchanger	10.2	3.9	297	3.7	284	2.3	245
4PGS_A (2.5)	Uncharacterized protein YetJ	15.4	3.5	207	2.7	175	2.2	183
4QTN_A (2.8)	Vitamin B3 transporter PnuC	9.0	4.2	202	3.0	155	2.8	178
4OD4_A (3.3)	4-hydroxybenzoate octaprenyltransferase	22.8	3.9	275	3.4	242	2.8	231
4O6M_A (1.9)	CDP-alcohol phosphotransferase	13.3	4.1	188	4.0	165	2.3	159
4WD8_A (2.3)	Bestrophin domain protein	5.94	N/A	268	Not converged			
4F35_A (3.2)	Transporter, NadC family	14.5	N/A	434	Not converged			

Many of the families we identified with no homologs of known structure are transmembrane (TM) proteins. To evaluate the accuracy of our co-evolution-based structure prediction method on TM proteins, we tested it on a benchmark of 13 TM proteins with recently determined structures. Rather than evaluating the lowest energy five models as in the case of the CASP experiment, we instead selected the most central (see ‘Materials and methods’) low-energy model and eliminated positions not converged within the lowest energy models or not constrained by contact information (see ‘Materials and methods’). As shown in Table 1, for the

11 of the 13 proteins for which the structure prediction calculations converged, the RMSD of the predicted structure to the experimentally determined structure over the converged and constrained residues was below 4.0 Å (the RMSDs over the structurally aligned regions were all below 2.8 Å). Features such as kinked, discontinuous, and re-entrant helices as well as coils within the bilayer that complicate approaches to membrane protein structure prediction that assume the accuracy of a TM helix prediction were all recovered correctly (for example, the re-entrant helices of aquaporin; the power of fragment-based approaches to model such features was noted in Nugent and Jones, 2012).

We built models for representatives of the 121 families with unknown structures using the Rosetta coevolution-guided structure prediction protocol, eliminating from the lowest energy structures the nonconverged and non-constrained residues. The calculations converged for 58 of the 121 proteins (Table 2). Four targets had Rc values less than 0.7; these targets contain clusters of contacts that may be involved in homo-oligomeric formation. The models are very different from those generated using traditional profile search and threading methods: with the exception of five targets with TMscore of 0.5 (Table 2, columns 7–8), the structural similarity of the Rosetta models to the top ranked models generated by HHsearch/SPARK-X is very low. The intractability of modeling these families using profile–profile/fold recognition methods is reiterated by the very low similarity between the models that best fit the contacts produced by HHsearch and SPARKS-X (Table 2, column 9; Supplementary file 2).

Based on the benchmark, we expect that our monomeric protein models should be within 4.0 Å RMSD of the actual structure. Provided there are not large conformational changes upon docking, protein– protein complexes can be accurately assembled from crystal structures or comparative models of the constituent monomers using GREMLIN contact predictions (Ovchinnikov et al., 2014). Thus, the models of complexes we provide in this article are likely to be fairly accurate if the monomeric subunits are predicted accurately, but there is clearly more room for error in our more complex multi-subunit predictions.

Table 3.2. Comparison of fold recognition and Rosetta models for large protein families. Column 2: number of unique proteins in family; Column 3: negative log10 of E-value of top match found in HHsearch profile–profile search of PDB; Columns 4–6: fit to predicted contacts (Rc value) of best fitting of top 10 HHsearch hits (column 4), of best fitting of top 10 SPARKS-X hits (column 5), and Rosetta model (column 6). Native structures have Rc values ranging from 0.7 to 1.2 (Figure 17). Columns 7–9: structural similarity (TMscore) between Rosetta model (M) and best fitting HHsearch model, between Rosetta model and best fitting SPARKS-X model, and between best fitting HHsearch and SPARKS-X models. The Rosetta models fit the contacts as well as expected for native structures and are very different from best fitting HHsearch and SPARKS-X models. For RARD and YEIH, the HHsearch E-value is less than 1E-04, the recommended threshold for inclusion in the same Pfam clan (Xu and Dunbrack, 2012), but the fit with the co-evolutionary contacts was very poor (Rc < 0.3); these two cases are discussed in sections below. For FLIL and YAII, the Rc values for very weak HHSearch and SPARKS-X hits (E-values worse than 0.1) are greater than 0.6 but the contacts constrain only a portion of the structure.

Known function			Rc			TMscore		
Name	**#seq**	**Ev**	**HH**	**SP**	**M**	**M_HH**	**M_SP**	**HH_SP**
WECH: O-acetyltransferase	24750	-2.4	0.0	0.1	0.9	0.1	0.2	0.1
SATP: Succinate-acetateproton symporter	2298	-2.1	0.4	0.5	1.1	0.3	0.3	0.8
LSPA: Lipoprotein signal peptidase	8156	-2.0	0.2	0.1	1.0	0.2	0.3	0.3
YADH: ABC-type multidrug transport permease	42626	-2.0	0.1	0.1	0.7	0.3	0.2	0.2
YEBZ: Putative copper export protein	4067	-2.0	0.1	0.1	0.8	0.2	0.3	0.2
CRCB: Fluoride ion exporter	7829	-1.8	0.2	0.3	1.0	0.2	0.2	0.3
LPTG: Lipopolysaccharide export system permease	8101	-1.8	0.0	0.1	0.9	0.1	0.1	0.2
FTSW: Lipid II flippase	14900	-1.7	0.0	0.1	1.0	0.1	0.2	0.2
RFAL: O-antigen ligase	13535	-1.7	0.2	0.1	0.9	0.3	0.2	0.2
CCMB: Heme exporter protein B	2433	-1.6	0.1	0.1	0.7	0.2	0.2	0.2
MLAE: ABC transporter permease for lipid asymmetry	7662	-1.4	0.0	0.1	0.9	0.1	0.2	0.3
SULP: Sulfate permease	6647	-1.2	0.1	0.0	0.8	0.2	0.2	0.2
TOLQ: Biopolymer transport protein	9256	-1.2	0.1	0.1	0.7	0.2	0.2	0.2
LGT: Prolipoprotein diacylglyceryl transferase	8121	-1.1	0.1	0.2	1.0	0.2	0.3	0.3
Q97UR7: N-methylhydantoinase B (HyuB-3)	4491	-1.0	0.1	0.1	1.1	0.1	0.1	0.1
YGAZ: putative L-valine exporter	6435	-1.0	0.1	0.2	0.9	0.2	0.3	0.2
CCMC: Heme exporter protein C	5965	-0.8	0.1	0.1	1.1	0.2	0.2	0.2
YEDZ: Sulfoxide reductase heme-binding subunit	2247	-0.7	0.2	0.2	1.0	0.2	0.3	0.3
YIAM: TRAP transporter small permease protein	10715	-0.7	0.1	0.2	1.1	0.3	0.3	0.2
TTDA: Tartrate dehydratase, alpha subunit	4238	-0.6	0.0	0.1	1.2	0.1	0.1	0.1
UPPP: Undecaprenyl pyrophosphate phosphatase	7842	-0.6	0.0	0.1	1.0	0.2	0.2	0.2
PLSY: Probable glycerol-3-phosphate acyltransferase	6112	-0.4	0.1	0.2	1.1	0.2	0.4	0.2
FLIL: Flagellar protein	2690	-0.3	0.7	0.5	0.8	0.5	0.4	0.9
CYDB: Cytochrome bd oxidase 2	6864	0.0	0.1	0.1	1.0	0.2	0.2	0.1
CYDA: Cytochrome bd oxidase 1	6200	0.1	0.0	0.1	1.2	0.1	0.2	0.2
MOTA: Motility protein A, flagellar motor proton conductor	4734	0.3	0.1	0.1	0.9	0.1	0.1	0.2
SLYB: Outer membrane lipoprotein	1860	0.3	0.1	0.2	0.8	0.2	0.2	0.1
MRED: Rod shape-determining protein	1546	0.6	0.5	0.5	0.8	0.5	0.4	0.6
ZUPT: Zinc transporter	10517	0.6	0.1	0.1	0.8	0.2	0.1	0.2
YOHK: Putative effector of murein hydrolase LrgB	3941	2.3	0.2	0.1	0.9	0.4	0.2	0.2
PRSW: Membrane proteinase	2500	5.3	0.2	0.2	0.9	0.3	0.3	0.7
DDG: Lipid A biosynthesis palmitoleoyl acyltransferase	9430	5.8	0.4	0.1	1.0	0.4	0.2	0.2

Unknown function			Rc			TMscore		
Name	**#seq**	**Ev**	**HH**	**SP**	**M**	**M_HH**	**M_SP**	**HH_SP**
YQFA: UPF0073 inner membrane protein	7596	-2.6	0.1	0.4	1.1	0.2	0.5	0.3
YCED: Uncharacterized protein	1604	-2.5	0.1	0.2	0.9	0.2	0.2	0.2
YPHA: Inner membrane protein	2986	-2.2	0.1	0.4	1.0	0.2	0.3	0.2
YADS: UPF0126 inner membrane protein	5222	-1.9	0.1	0.1	0.9	0.2	0.3	0.2
YHHN: Uncharacterized membrane protein	2529	-1.9	0.1	0.2	0.9	0.2	0.3	0.2
YIDH: Inner membrane protein	1041	-1.9	0.1	0.2	0.6	0.3	0.3	0.2
YITE: UPF0750 membrane protein	8326	-1.7	0.1	0.1	0.9	0.2	0.3	0.3
HDED: Acid resistance membrane protein	2885	-0.6	0.1	0.2	0.8	0.2	0.2	0.2
YFIP: DTW domain-containing protein	3100	-1.5	0.2	0.2	0.9	0.2	0.2	0.1
YPJD: ABC-type uncharacterized permease	6180	-1.4	0.2	0.2	0.9	0.2	0.3	0.2
YJFL: UPF0719 inner membrane protein	1581	-1.3	0.1	0.1	0.7	0.2	0.3	0.3
YTEJ: Uncharacterized membrane protein	5733	-1.2	0.1	0.1	1.0	0.2	0.2	0.2
YIHY: UPF0761 membrane protein	10144	-0.9	0.1	0.1	0.9	0.1	0.2	0.2
YQAA: Inner membrane protein	2187	-0.9	0.1	0.3	1.0	0.2	0.4	0.3
YHID: Uncharacterized protein	4416	-0.7	0.2	0.2	1.0	0.2	0.1	0.2
YLOU: Uncharacterized protein	3738	-0.7	0.4	0.5	0.9	0.3	0.3	0.8
YGDD: UPF0382 inner membrane protein	3025	-0.6	0.5	0.3	1.0	0.3	0.2	0.4
YJCH: Inner membrane protein	1307	-0.5	0.3	0.2	0.8	0.4	0.2	0.2
YFCA: UPF0721 transmembrane protein	18846	0.0	0.1	0.1	1.0	0.2	0.3	0.2
YOHJ: Putative effector of murein hydrolase	3608	0.4	0.2	0.3	0.5	0.3	0.4	0.6
YHHQ: Inner membrane protein	3398	0.7	0.4	0.2	1.0	0.4	0.3	0.2
YAII: UPF0178 protein	3144	0.8	0.6	0.7	1.1	0.5	0.5	0.4
YUXK: Predicted thiol-disulfide oxidoreductase	1881	1.3	0.3	0.3	1.1	0.3	0.3	0.5
YICC: UPF0701 protein	4293	1.5	0.1	0.1	1.0	0.1	0.1	0.1
YEIH: UPF0324 inner membrane protein	4863	4.2	0.3	0.2	0.9	0.4	0.5	0.7
RARD: putative chloramphenical resistance permease	74507	6.3	0.1	0.1	1.0	0.3	0.3	0.2

The models are available at (External Database: http://gremlin.bakerlab.org/structures/). The biological implications of all of these structures cannot be explored in a single paper; here, we describe functional insights obtained from a subset of the models. These insights derive in part from the distribution of evolutionarily conserved residues in the models, as conserved sequence motifs tend to mark functional sites in structures (Zuckerkandl and Pauling, 1965; Villar and Kauvar, 1994; Pei and Grishin, 2001; Muth et al., 2012). As is evident in Figure 2, the conserved residues cluster quite strongly in the predicted structures. We describe first, hypotheses on mechanism for proteins of known function, and second, hypotheses on function for proteins with currently unknown function. In the following sections, the predictions are grouped by known biological functions assigned by Clusters of Orthologous Groups (Galperin et al., 2015). Hopf et al. (2012) also used co-evolution information to guide membrane protein structure prediction and function assignment; we compare to their conclusions in the two cases common to both studies.

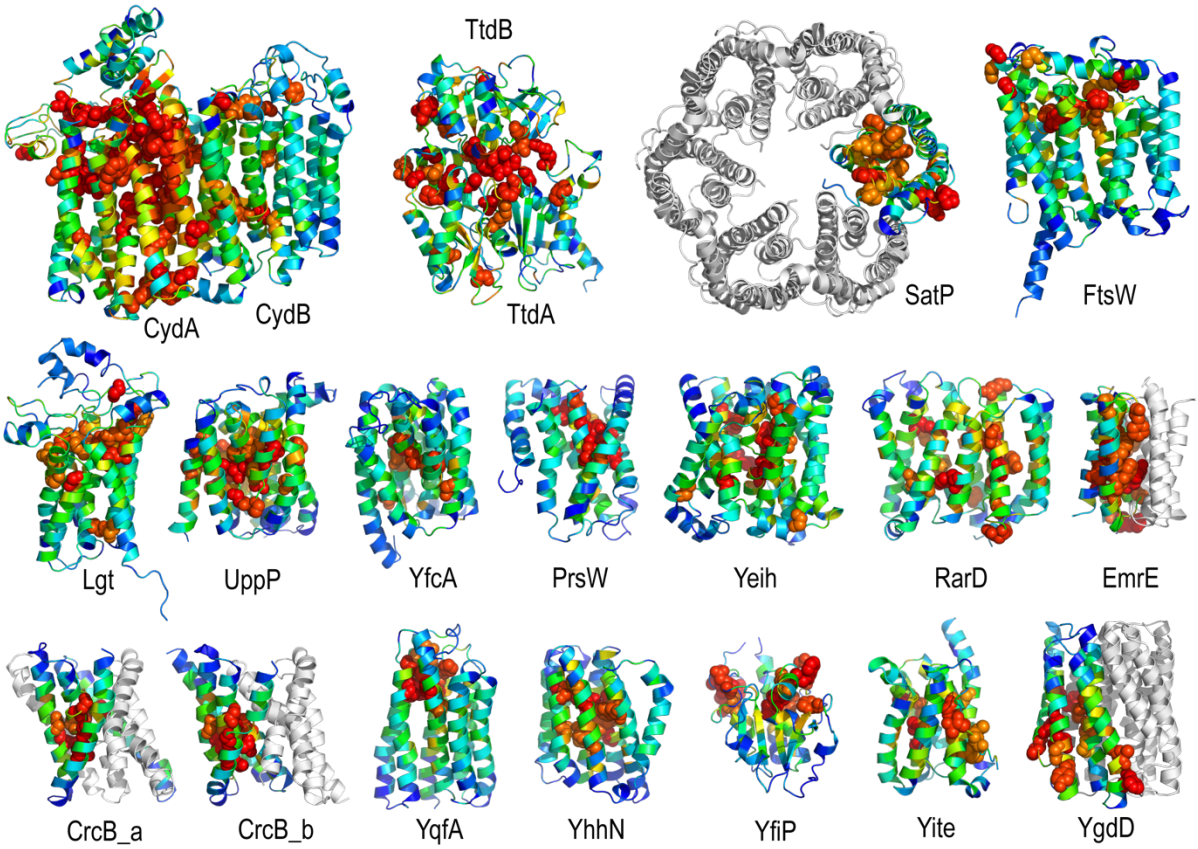


Figure 3.2. Conserved residues tend to cluster in the predicted structures.
Residue conservations from multiple sequence alignments were mapped to predicted structures using Al2Co (Pei and Grishin, 2001) and are colored in rainbow from blue (variable) to red (conserved). The most conserved residues (red or orange), displayed as spheres to highlight their positions, tend to line interaction surfaces and indicate potential functional sites.

3.4 BIOLOGICAL INSIGHTS FROM STRUCTURAL MODELS

3.4.1 *Energy production and transport*

Cytochrome bd-I ubiquinol oxidase generates a proton-motive force to power the adenosine triphosphate (ATP) synthase when oxygen is limited. The enzyme has two integral membrane subunits (CydA and CydB) with three hemes (heme b595, heme b558, and heme d) that mediate transfer of electrons from quinol to oxygen. Using the co-evolution-guided structure prediction protocol described above, we generated models for the structures of CydA and CydB, and then docked the subunits together using inter-protein predicted contacts as described in Ovchinnikov et al. (2014) to generate a model for the entire TM complex (Figure 3A,B). The models of CydA and CydB share the same fold—a duplicated four helix bundle unit—and form a pseudo

symmetric heterodimer. Structure comparisons of CydA and CydB to the PDB revealed nearly full-length structural similarity to polysulfide reductase TM domain (PsrC) (PDB: 2VPX), an enzyme complex responsible for the quinone-mediated reduction of polysulfide, and structure comparisons for the four helix bundle unit revealed strong similarities to cyt b561 (PDB: 4O7G). The b595 and b558 heme-binding sites of each CydA four helix bundle have been mapped experimentally by mutagenesis: H19 ligates heme b595, and H186 and M393, heme b558. Strikingly, in our model, these residues are aligned with conserved axial ligands in cyt b561 (Figure 3C). Residues ligating heme d have not yet been identified experimentally, but in our model, a third conserved CydA histidine, H126 structurally aligns to a known heme-binding site near the cytoplasmic surface of cyt b561. We hypothesize that this residue ligates heme d, which has been proposed to be on the periplasmic side (see Figure 3C), a location of heme d near the cytoplasm could explain the protonmotive force generated across the membrane. In addition to the heme-ligating residues, mutagenesis studies (Borisov et al., 2011) have identified residues involved in quinone binding and proton flow. In our model of the structure of CydA, the quinone-binding residues (Figure 3B, red spheres) cluster together, and the proton channel residues (Figure 3B, blue spheres) cluster together. Two additional conserved residues with no known function (R9 and R448) are near the quinone-binding Q loop on the periplasmic surface. Thus, the CydA model agrees with extensive mutagenesis data and places the cytochrome bd–I complex within the evolutionary context of other TM di-heme cytochromes.

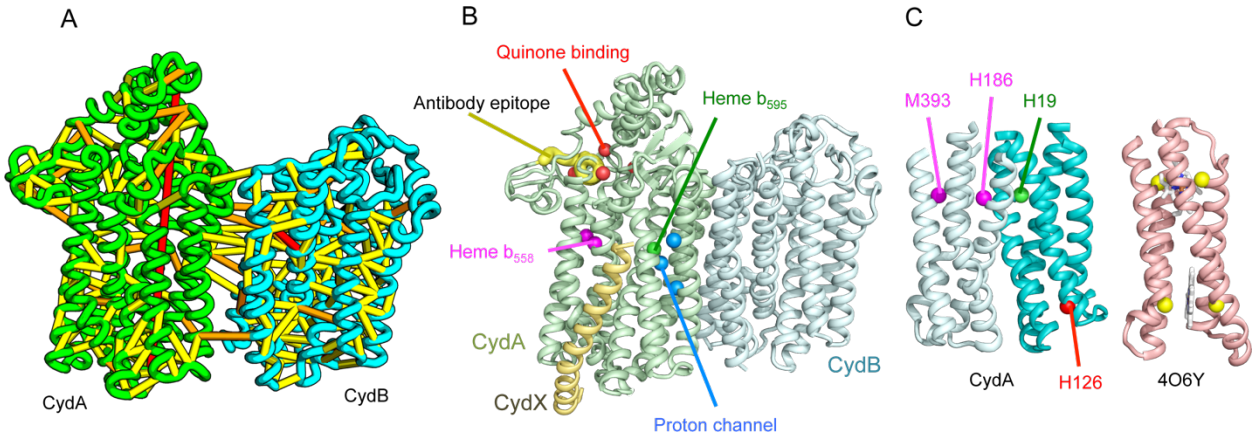


Figure 3.3. Predicted structure of the Cytochrome bd oxidase complex.
(A) Location of the top co-evolving residue pairs in our model. For clarity, the monomers have been pulled apart slightly. (B) Location of conserved and experimentally characterized residues (Borisov et al., 2011) on structure model. (C) Residues that coordinate heme in CydA are in the same location as histidines (yellow spheres) in Cytochrome b561 (PDB: 4O6Y). H126 (red

sphere) overlaps one of these histidines and is the proposed as a heme d coordination site. For clarity, both the model of CydA and the structure of Cyt b562 (4O6Y) are trimmed to highlight the four helix bundle(s).

L-tartrate dehydratase is used by *E. coli* under anaerobic conditions to convert L-tartrate (carbon source) to oxaloacetate. The enzyme is a hetero-tetramer, with two copies of TtdA and two copies of TtdB (Reaney et al., 1993). TtdA is homologous to the N terminus of a class I fumarase, and TtdB, to the C terminus of the fumarase. The structures of TtdA and the fumarase N-terminus have not been determined, but the structure of the fumarase C-terminal domain has been solved (PDB: 2ISB) and is structurally related to the swiveling domain from aconitase enzymes that perform similar chemistry (Lauble et al., 1994). The TtdB-like swiveling domain from aconitase (PDB: 1ACO) binds its substrate near the interface of the swiveling domain and another catalytic domain that binds 4Fe-4S. Given the importance of the adjacent catalytic domain as well as the domain interface in aconitase, we predicted the structure of TtdA and assembled it into the hetero-tetramer complex with a homology model of TtdB (Figure 4A). In our complex model, three conserved TtdA cysteines (C71, C190, and C277) cluster near the TtdB interface which maybe the 4Fe-4S cluster-binding site (Figure 4B). This potential active site also includes a conserved aspartate (D73) that might contribute to catalysis. The conserved TtdB H265 falls on the opposite side of the active site in our model and instead contributes to the active site of the second TtdB chain formed by the tetramer. Thus, our model suggests TtdA/TtdB forms an obligate tetramer that would not have been predicted by co-evolution or conservation alone.

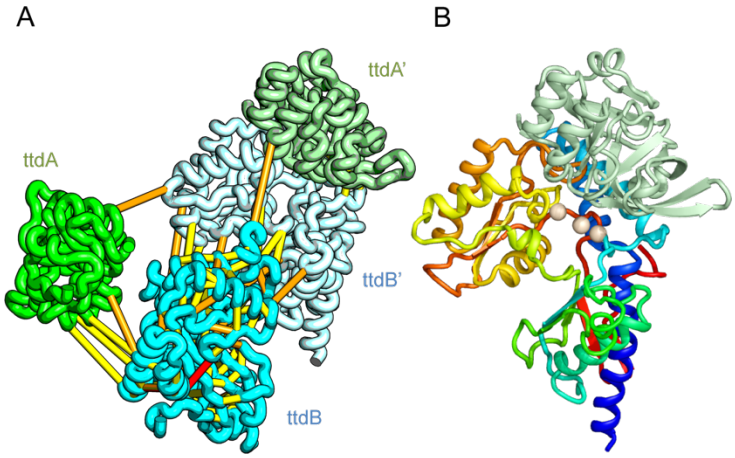


Figure 3.4. Predicted structure of the tartrate dehydratase heterotetramer composed of two copies each of ttdA and ttdB.

(A) Co-evolving residue pairs. The monomers have been pulled apart to reveal the contacts. (B) The ttdA subunit (rainbow) contains a 4Fe-4S cluster (white spheres) that is near the interface with ttdB (green).

SatP (Succinate-acetate/proton symporter) mediates the uptake of succinate and acetate in *E. coli* coupled to proton symport (Sa-Pessoa et al., 2013). Our predicted SatP structure (Figure 5A–C) is very similar to that of the proton-gated urea channel (Figure 5D). The urea channel assembles into a hexameric ring with each protomer forming a channel through the center of the 6TMH fold. Conserved residues line both the channel and the protomer interface and are important for proton gating and solute selectivity. Assembly of our SatP model into a hexameric ring satisfied predicted contacts not made in the monomer (Figure 5A,B). Residues that have been shown to influence the solute selectivity of SatP (Leu131 and Ala164) (Sa-Pessoa et al., 2013 ´) line the channel pore of our model (Figure 5C). Most of the conserved SatP residues line the channel at a similar depth as the constriction sites in UreI and are likely involved in similar gating and selectivity functions as their UreI counterparts. A cluster of conserved residues face the periplasmic surface and align to the periplasmic loop (PL1) in UreI that is thought to plug the channel in a proton-dependent manner (Hommais et al., 2004). The similarity of our model to the SatP fold is not only supported by mutagenesis data but also suggests the functional importance of multimeric assembly not revealed by co-evolution or conservation analysis alone.

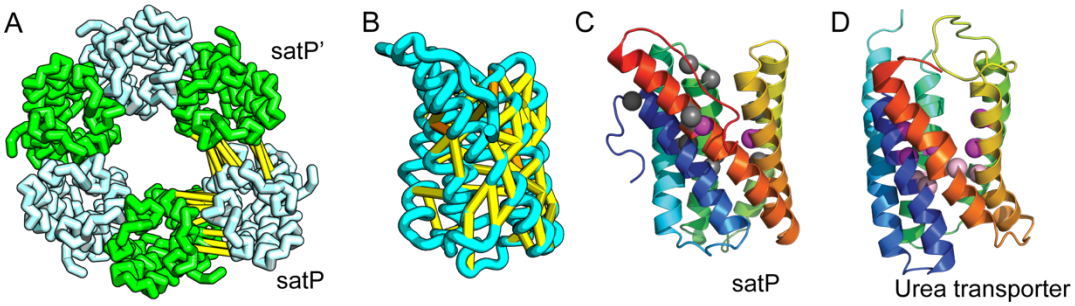


Figure 3.5. Succinate-acetate/proton symporter SatP (YaaH).
(A) Co-evolving residue pairs in homo-oligomer model. (B) Co-evolving residue pairs in SatP monomer model. (C) SatP co-evolution-based model places known acetate selective residues (magenta spheres) lining the channel. Conserved residues (gray spheres) line the periplasmic surface. The 6TMH channels are formed by threefold pseudo-symmetric TMH hairpins. (D) Proton-gated UreI channel protomer. C-alpha positions at the periplasmic constriction site (magenta spheres) and the cytoplasmic constriction site (pink spheres) are highlighted. The SatP model has the same fold as UreI (C vs D).

3.4.2 *Lipid and bacterial cell wall synthesis*

Bacterial cell wall synthesis involves multiple steps. FtsW is an integral membrane protein that is thought to transfer lipid-linked peptidoglycan precursors from the inner to the outer leaflet of the cytoplasmic membrane, where it interacts with the TM portion of peptidoglycan synthetase FtsI (Fraipont et al., 2011). Using co-evolutionary information for the FtsW family, and between it and FtsI, we generated a model of FtsW in complex with the TM domain of FtsI (Figure 6A,B). The FtsW model encompasses 10 TM helices, with the last seven (TMH4-TMH10) adopting a similar topology as TMH4-TMH7 and TMH10-12 of the TM domain of STT3 (PDB: 3WAK) (Figure 6C,D). STT3 is a dolichyl-diphosphooligosaccharide-protein glycosyltransferase that functions in N-glycan biosynthesis, transferring oligosaccharides from the membrane anchor dolichol-diphosphate to asparagine residues of proteins bound for secretion (Matsumoto et al., 2013). The FtsW substrate, lipid II, has a membrane anchor similar to that of the STT3 substrate donor: bactoprenol-pyrophosphate conjugated to disaccharide. A conserved DxH motif at the N-terminus of STT3 TMH4 coordinates a divalent metal ion in the active site. Two residues from the corresponding TMH4 of FtsW (R145 and K153) are essential for flippase activity, with the side chain of R145 overlapping the divalent metal in superpositions of the FtsW model with the STT3 structure. Other conserved FtsW residues line this site and probably contribute to function. The conserved FtsW/STT3 TMH core is similar in sequence to the *E. coli* O-antigen ligase RfaL, and our predicted structure for RfaL is similar in structure (Figure 6E). Thus, the structure model of FtsW suggest a potential active site analogous to that of the structurally related STT3 TMH core and unites the family with another bacterial cell wall biogenic enzyme.

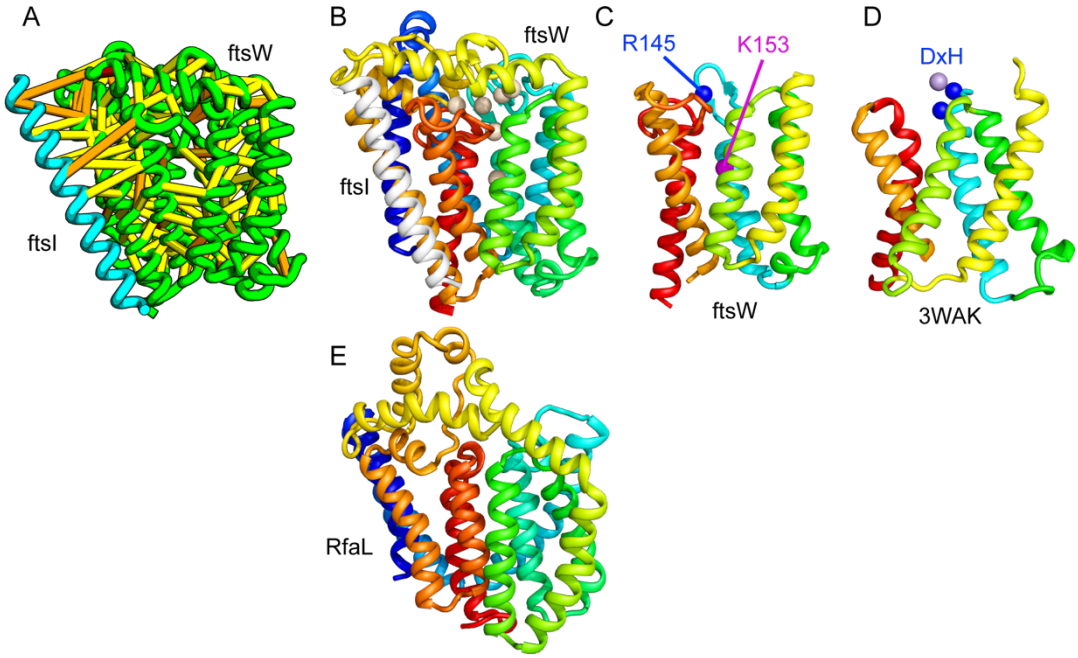


Figure 3.6. Lipid II flippase (FtsW) in complex with the transmembrane domain of Peptidoglycan synthase (FtsI).
FtsW is an essential cell division protein that transports lipids across the cytoplasmic membrane and is required for localization of FtsI. (A) Location of the top co-evolving residue pairs. (B) White spheres indicate conserved positions in FtsW that when mutated to alanine result in loss of flippase activity. (C, D) The last seven transmembrane (TM) helices of FtsW (TMH4-TMH10) adopting a similar topology as TMH4-TMH7 and TMH10-12 of the TM domain of STT3 (PDB: 3WAK). Both the model of FtsW and 3WAK was trimmed over the aligned helices for clarity. (C) Two residues from the corresponding TMH4 of FtsW (R145 and K153) are essential for flippase activity. (D) The side chain of R145 overlaps the residues that coordinate the divalent metal in the conserved DxH motif at the N-terminus of STT3 TMH4. (E) The model of RfaL adopts a similar fold as ftsW.

E. coli prolipoprotein diacylglyceryl transferase (Lgt) is an inner membrane protein that transfers the diacylglyceryl moiety from phosphatidylglycerol to an N-terminal cysteine residue that follows the signal peptide of prolipoproteins. Our predicted structure of Lgt has a novel seven trans-membrane helix (TMH) fold, with many of the conserved residues (Y26, R134, N146, E151, G154, R239, and E243) clustering near the proposed periplasmic surface to form a putative active site (Figure 7, white spheres); the activity of Lgt is lost or greatly reduced upon mutating these residues to alanine (Pailler et al., 2012). The topology and orientation of the TMHs in our structural model are consistent with a previously proposed topology model (Pailler et al., 2012).

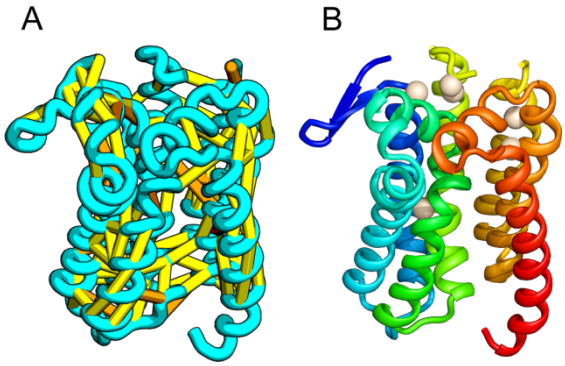


Figure 3.7. Prolipoprotein diacylglyceryl transferase (LGT).
(A) Predicted contacts indicated on model, (B) model with conserved positions at which alanine mutations result in loss in activity indicated in white spheres; five of these are clustered at the periplasmic end of the model.

UppP (undecaprenyl pyrophosphate phosphatase), an integral membrane protein with unknown structure, catalyzes the dephosphorylation of undecaprenyl pyrophosphate to form undecaprenyl phosphate, an essential carrier lipid for bacterial peptidoglycan cell wall synthesis (El Ghachi et al., 2005). In our UppP structure model, TMH1 and TMH5 form broken helices that enter and exit the membrane on the same side, placing both catalytic regions near to the core of the structure (Figure 8A). In contrast to a previously proposed model that assumed unbroken helices (Chang et al., 2014), our model has a twofold symmetry between the broken TMH1-TMH4 and the broken TMH5-TMH8 that is mirrored by an internal duplication present in the UppP sequence. A similarly duplicated TMH family of unknown function (YfcA) is distantly related to UppP by sequence. Our structure prediction calculations suggest that YfcA has the same fold (Figure 8B; the multiple sequence alignments used to make the predictions for UppP and YfcA do not share any sequences). The UppP model illustrates the ability of our method to model unusual structural features such as the broken TMH helices, which are typically difficult to model without precedence in existing structure templates.

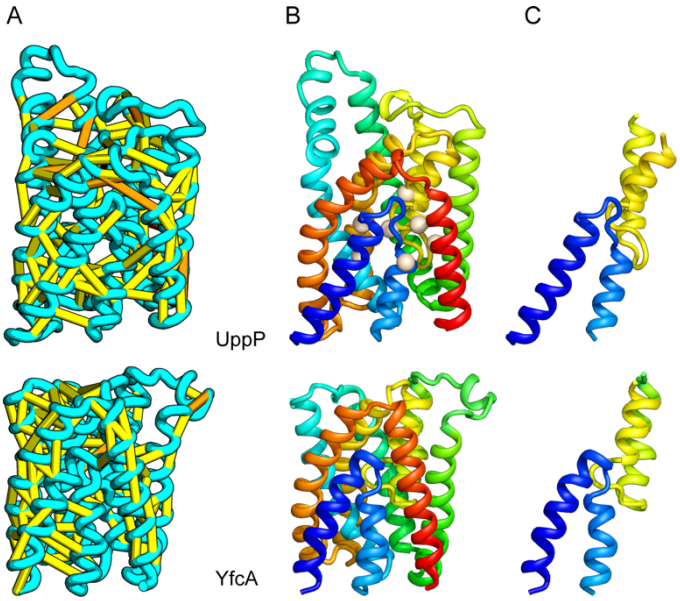


Figure 3.8. UppP catalyzes the dephosphorylation of undecaprenyl diphophate (UPP). A) Location of the top co-evolving residue pairs. B) Spheres in white indicate conserved residues experimentally shown to decrease activity to <1% (Chang et al. 2014); all these residues are in the core in the model. YfcA, a protein of unknown function is a very distant sequence homologue of UppP (they are in different PFAM families); C) the predicted structure of YfcA has the same fold as UppP with prominent broken helices (highlighted in blue and yellow).

3.4.3 *Proteases*

PrsW of *B. subtilis* is an intramembrane protease that cleaves site-1 anti-σ factor RsiW, a crucial step in the resistance to antimicrobial peptides (Ellermeier and Losick, 2006). PrsW belongs to a large superfamily of membrane proteins that includes putative bacteriocin-processing enzymes and the APH-1 subunit of gamma-secretase (Pei et al., 2011b). Our PrsW model has structural similarity to an archaeal type II CAAX prenyl protease (Manolaridis et al., 2013), mostly in a core of four TMHs (TMHs 3–6 in PrsW model and TMHs 4–7 in type II CAAX prenyl protease) (Figure 9C,D). The predicted active site residues in motifs EExxK (TMH3) and HxxxD (TMH6) of PrsW occupy structurally compatible positions as conserved residues in motifs EExxxR (TMH4) and HxxxN (TMH7) of type II CAAX prenyl protease. Another conserved histidine in the fifth TMH of PrsW (but absent in type II CAAX prenyl protease) is also located in the predicted active site of PrsW.

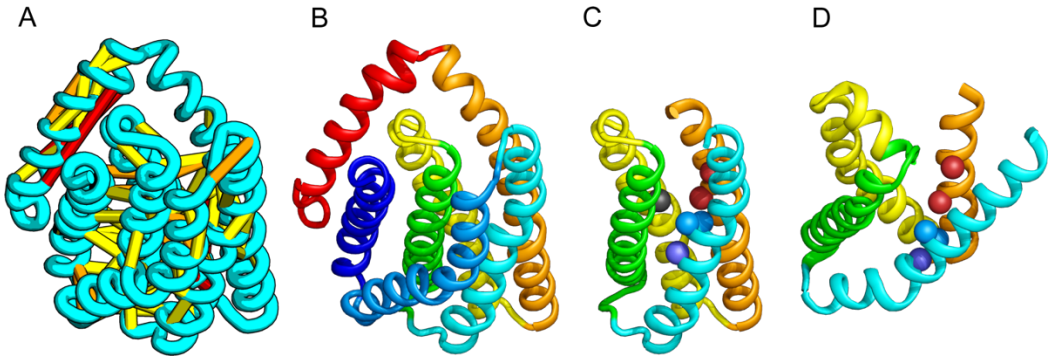


Figure 3.9. PrsW is an intramembrane protease that is crucial in the resistance to antimicrobial peptides.
(A) Location of the top co-evolving residue pairs. Our model of PrsW (B) contains a 4TMH substructure ([TMHs 3–6], (C) which is very similar to a substructure of type II CAAX prenyl protease [TMHs 4–7; D]). The predicted active site residues in PrsW motifs EExxK (TMH3; blue spheres, C) and HxxxD (TMH6; red spheres, C) are in positions similar to those of conserved residues in motifs EExxxR (TMH4; blue spheres, D) and HxxxN (TMH7; red spheres, D) of type II CAAX prenyl protease. Another conserved histidine in the fifth TMH of PrsW (but absent in type II CAAX prenyl protease) is also located in the predicted active site of PrsW (black sphere).

3.4.4 *Transporters*

The inner membrane protein YeiH is classified as a member of the CPA/AT transporter clan in PFAM and sequence search yields high confidence matches to sodium/proton antiporters (HHsearch e-value 2.2E-05). Remarkably, although our structural model of YeiH (Figure 10A,B) superimposes structurally with the structure of the antiporter NapA (PDB: 4BWZ; Lee et al., 2013), the connectivity of the core of the structure is completely different. The core domain of NapA contains two antiparallel discontinuous helices (TM4a, 4b and TM11a, 11b) that cross over each other (Figure 10C). In our YeiH model, the same hourglass-shaped assembly is formed by two pairs of broken helices (TM5, 6 and TM8, 9) that exit the membrane on the same side (Figure 10B). Discontinuous helices have been found in several transporter proteins and are frequently involved in ion binding (Screpanti and Hunte, 2007); a similar arrangement of broken helices is also observed in structures of CLC chloride channels and glutamate transporter Glt (Dutzler et al., 2002; Yernool et al., 2004). Like the UppP example, our model of YeiH highlights differences in the TM core that are hard to model with template-based homology modeling approaches.

The *E. coli* chloramphenical resistance protein RarD is an apparent duplication of the homodimeric *E. coli* EmrE drug transporter, and our predicted structure of RarD indeed adopts

an internal pseudo-symmetric fold. Our RarD model is structurally superimposable on the EmrE homodimer, but the first helices of the domains corresponding to the EmrE monomer are swapped (Figure 11B) The duplicated domains in RarD (called EamA domains) differ from EmrE by a critical helix insertion between helix-1 and helix-2 that causes helix-1 to adopt an inverted conformation in the membrane. The only way helix-1 in EamA can preserve the interactions seen in EmrE is to instead interact with the second copy of EamA (which is inverted in the membrane) as in our structure model (Figure 11C). The EamA protein, also composed of two EamA domains, was previously modeled (Hopf et al., 2012) but no structural similarity was reported to EmrE.

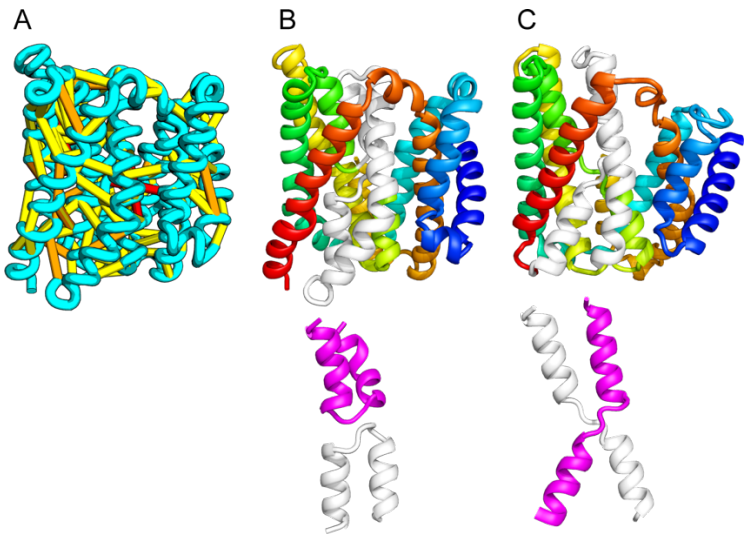


Figure 3.10. Our model of the inner membrane protein YeiH (A, B) is structurally similar to the structure of the antiporter NapA (C).
Lower panels: TM helices of core domains are highlighted in white and magenta: while these helices cross over each other in NapA (right), the core of our model of YeiH (left) is formed by two pairs of broken helices (TM5, 6 and TM8, 9) that exit the membrane on the same side.

EmrE is one of a small number of dual-topology TM proteins in which a single polypeptide chain can insert into the membrane in two opposite orientations, thus yielding inverted symmetric TMH topologies (Duran and Meiler, 2013). This inverted symmetry is fixed in the monomeric RarD structure. The proposed transport mechanism for EmrE involves switching the dimeric structure between alternate access states (Fleishman et al., 2006; Morrison et al., 2012). The homologous relationship between EmrE and RarD suggests the inverted symmetric RarD structure might also adopt alternate access states involving the two duplicated EamA halves. Our set of predicted structures includes a member of a second predicted dual-

topology protein, the dimeric fluoride transporter CrcB (Rapp et al., 2006), which adopts the dual topology (see External Database).

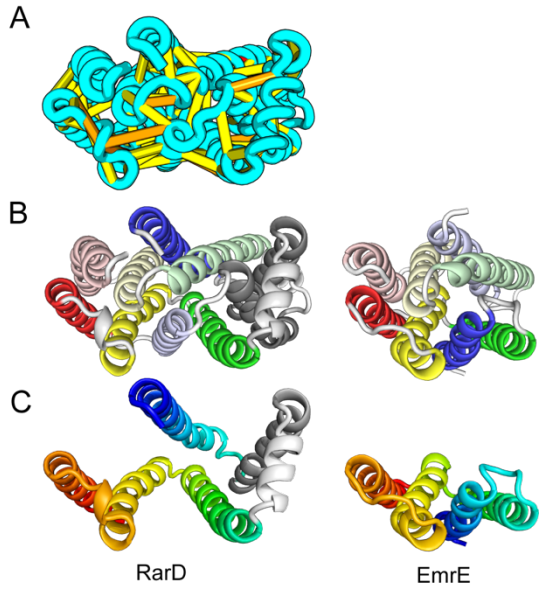


Figure 3.11. Our model of RarD has a similar architecture to EmrE but different fold.
(A, B) Full-length RarD and EmrE homodimer. (C) RarD internal repeat and EmrE monomer. The N-terminus helix (blue) is swapped in RarD relative to EmrE due to helix insertion (gray).

Internal pseudo-symmetry is often observed in the structures of TM proteins. Evolutionary pathways leading to such symmetry can involve gene duplication and fusion events, this is particularly likely when the symmetric single-chain protein has the same overall fold as a known homo-oligomer. While these duplication events can be revealed by the presence of internal sequence repeats, the tendency of the duplicated sequence to diverge and adopt alternate or specialized functions can mask detection of duplication events at the level of primary sequence. The sparseness of determined TM protein structures further complicates analysis of evolutionary folding pathways (Duran and Meiler, 2013). Our co-evolution-based structure models substantially increase known TM protein fold space (many have TMalign [Zhang and Skolnick, 2005] scores < 0.5 to any known structure, Supplementary file 2), populating it with new structures that reveal evolutionary folding pathways.

3.4.5 *Unknown function*

Protein structures with similar topology often have similar function. Building models allows detection of fold similarity to previously solved structures in the absence of significant sequence homology.

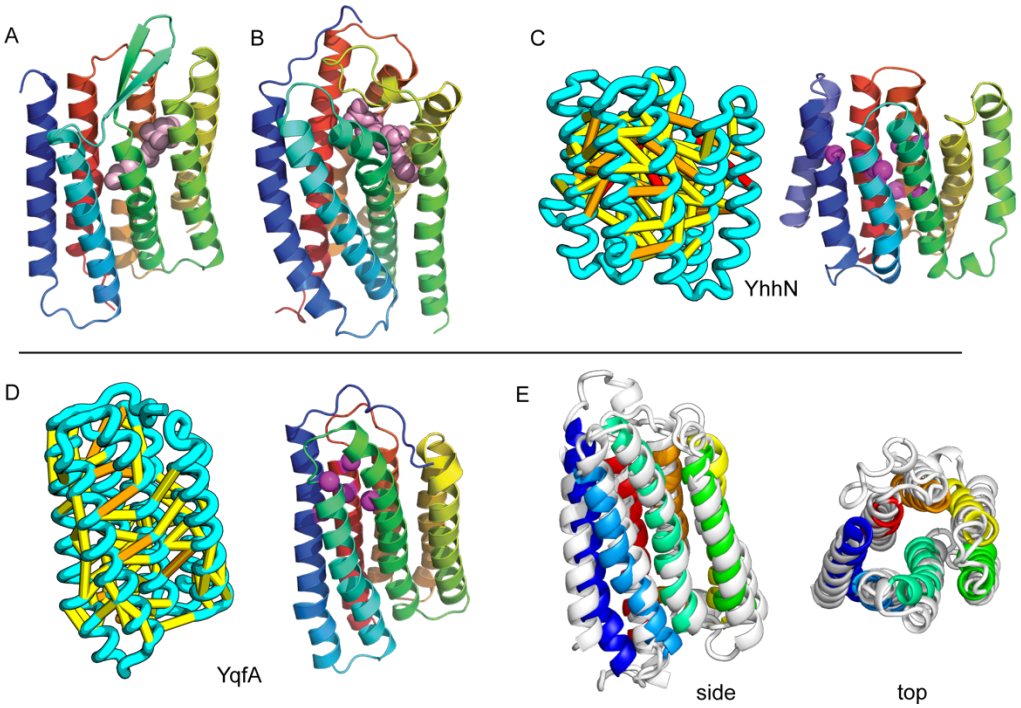


Figure 3.12. Predicted structures of YqfA and YhhN have topologies similar to G protein-coupled receptors (GPCR-like).
The core seven TM helix (TMH) fold exhibited by members of the GPCR superfamily is colored in rainbow from the N- to the C-terminus. (A) Bacteriorhodopsin binds retinal (pink spheres) in a pocket formed by TMH3-7 [PDB ID: 1m0k]. (B) The agonist (pink spheres) binding site of P2Y12 receptor is formed by the same set of helices [PDB ID: 4pxz]. (C) A co-evolution-based structure model for YhhN has the GPCR topology with an N-terminal TMH extension. Conserved residues that might form an active site (magenta spheres) cluster in a similar place as the YqfA catalytic residues. (D) Our co-evolution-based structure model for YqfA has a GPCR like topology and clusters residues that may form an active site (magenta spheres mark the Calpha position) in a region that corresponds to the GPCR ligand-binding pocket. (E) Side and top view of the TMalign superposition of YqfA model (in rainbow) over the recently released 3wxw (in white) human ortholog. The N- and C-terminal loops were trimmed for clarity. The TMalign score between the model and the homolog is 0.8.

Our models of *E. coli* proteins YqfA and YhhN have topologies similar to that of G-protein-coupled receptors (GPCRS). The YqfA sequence belongs to a large family of integral membrane proteins, with members in all three kingdoms of life. The eukaryotic members are seven-TM pass receptors for ligands such as prostaglandin and adipoQ (the progesterone-

adiponectin receptor (PAQR) was predicted to be bacteriorhodopsin-like by Hopf et al., 2012), while a bacterial member is associated with furfural tolerance through an unknown mechanism. The PAQR receptors belong to a larger superfamily of core seven TM-bound putative hydrolases identified as CREST (Pei et al., 2011a). The CREST superfamily is characterized by conserved motifs at the end of TMH2 (SxxxH), the beginning of TMH3 (D), and the beginning of TMH7 (HxxxH). In the YqfA model, the conserved CREST motifs likely form an active site and are in the same region as the ligand-binding pocket in GPCRs (Figure 12). During the preparation of this manuscript, the structure of a homolog of YqfA was released by the PDB; our model of YqfA is very similar to this structure (TMalign score of 0.80; Figure 12E), while the top hit of HHsearch and SPARKS-X models are not similar (TMalign score of 0.21 and 0.40). Our coevolution-based model of YhhN (Figure 12C) has the GPCR topology, but with an N-terminal TMH extension. A YhhN family member was recently shown to function as a lysoplasmalogenase that catalyzes hydrolysis of the vinyl ether bond of lysoplasmalogen in lipid metabolism. Conserved YhhN residues that might form the active site cluster in a similar place as the YqfA putative catalytic residues, lining what would be the ligand-binding site in GPCRs (Figure 12C,D).

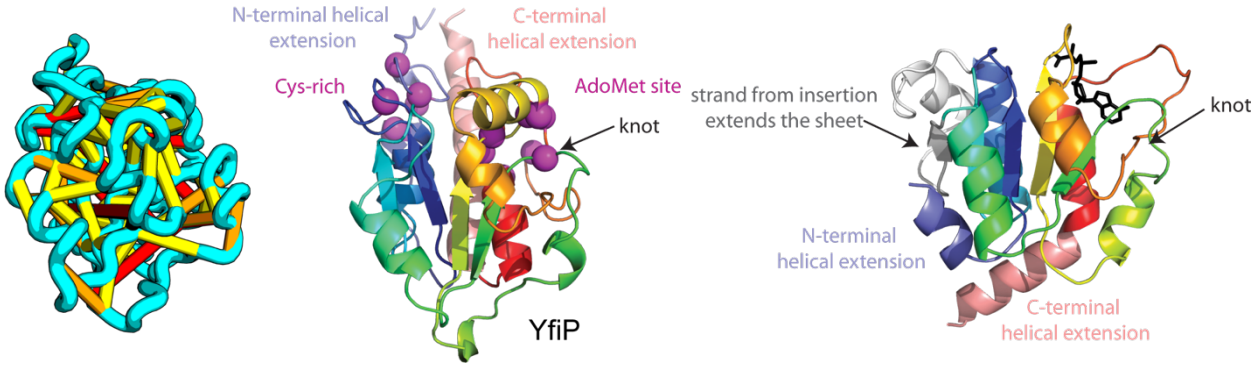


Figure 3.13. YfiP predicted structure has methyltransferase-like fold with knot.
Left: the top co-evolving residues pairs. Middle: conserved residues (magenta) surround the AdoMet-binding site and a conserved Cys could bind a Fe4S cluster. Right: 3nk7 methyltransferase bound to AdoMet.

Our model of YfiP, from a family of unknown function, contains a non-trivial trefoil knot topology characteristic of the alpha/beta knot methyltransferase (SPOUT) superfamily (Anantharaman et al., 2002). SPOUT structures utilize conserved residues to bind the AdoMet substrate in a binding cleft formed by the knot. In the predicted YfiP structure, conserved DTW

domain residues surround the AdoMet binding cleft, including Asp113, Thr115, Trp116, Pro87, Tyr145, Arg148, Thr158, and Glu160 (Figure 13). The predicted similarity of YfiP to the SPOUT methyltransferase fold substantiates a previously suggested role in rRNA processing (Burroughs and Aravind, 2014), as rRNA maturation requires extensive nucleotide modifications. The YfiP model is an example of a prediction of an unusual structure that is indicative of function and cannot be predicted by co-evolution or conservation alone.

Our model of YitE, from a protein family of unknown function, has an arrangement of secondary structures nearly identical to the aquaporin water channel fold (Figure 14), including the pseudo-symmetric repeat unit, but with completely different connectivity. The two half-helices that meet at the center of the protein are a key feature of water channels and critical for proton exclusion (Gonen et al., 2005). The YitE model does not have the 'NPA' motif in the half helix characteristic of water channels, but one of the half helices has an N pointing into the putative pore.

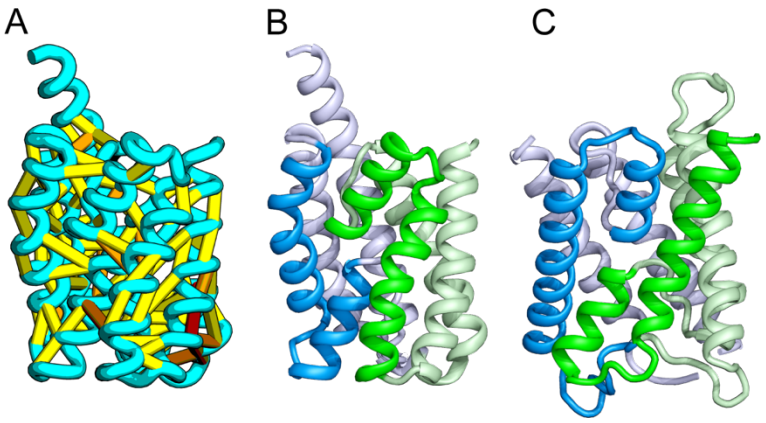


Figure 3.14. *Bacillus subtilis* YitE model.
(A) The top co-evolving residues. YitE (B) has architecture similar to aquaporin (PDB:2B6P) (C), including the internal pseudo-symmetry (blue vs green), but completely different connectivity.

The *E. coli* YgdD protein belongs to a family of unknown function (Pfam: DUF423) with members widely distributed in both bacteria and eukaryotes, including proteins from plants, fungi, and metazoans. The predicted contacts of YgdD are best accommodated in a homotrimer model (Figure 15A,B). Each YgdD molecule in the homotrimer has four TMHs with left-handed connections between helices 1, 2, 3 as well as between helices 2, 3, 4. Such a topology matches that of Membrane-Associated Proteins in Eicosanoid and Glutathione metabolism (MAPEG)

family proteins (Jakobsson et al., 1999; Hebert and Jegerschold, 2007). More strikingly, YgdD and MAPEG also exhibit the same overall homotrimer topology, both similar to the core of the hemecopper oxidase catalytic subunit (Pei et al., 2014) (Figure 15C,D). Sequence similarity searches of YgdD by HHsearch (Remmert et al., 2012) did not reveal significant hits to MAPEG proteins, but weak HHsearch matches to heme-copper oxidase members were found (e.g., 3mk7, chain A, HHsearch probability score: 41). The sequence alignment between the last two TMHs of YgdD and the last two TMHs of heme-copper oxidase members is consistent with the structural alignment between our predicted YgdD structure and the 3mk7 structure, suggesting that YgdD is evolutionarily related to heme-copper oxidases.

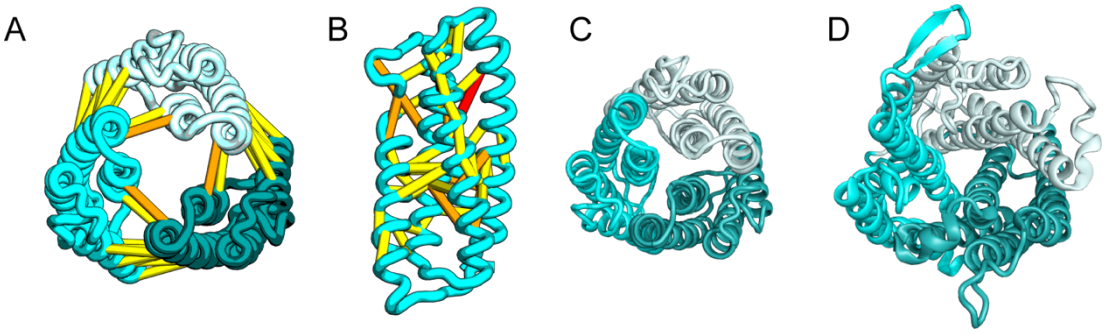


Figure 3.15. *Escherichia coli* protein YgdD.
Our model of the *E. coli* protein YgdD trimer (C) is based on predicted contacts satisfied within the monomer (B) and between monomers in the homo-trimer (A). Structural similarity to heme copper oxidase (D) along with a weak HHpred sequence match over part of the protein suggests that YgdD is evolutionarily related to heme–copper oxidases.

3.5 DISCUSSION

The models presented in this article for 58 large protein families which cannot be accurately modeled using comparative modeling or fold recognition methods cover a significant fraction of the prokaryotic sequences for which structural information was previously unavailable. Each of these families have thousands of members (Table 2, column 2), hence these models have quite broad impact. The analyses of a small subset of these structures provided here only begins to uncover the wealth of information relating to function they contain. In addition to the new structure-based interpretation of existing sequence conservation and mutational data the models enable, they illustrate the complex transformations occurring in membrane protein evolution: for example, the changes in YeiH and RarD structural element connectivity compared to previously known structures. With the advent of sensitive sequence profile–profile comparison methods,

much of protein structure modeling has been reduced to sequence alignment, and indeed for functional interpretation often much can be learned simply by draping the query sequence on a homologous structure; in contrast, as illustrated in the examples above, in the co-evolution-guided de novo structure prediction case, structure modeling is critical to functional insight.

Large-scale genome sequencing is having an unanticipated impact on protein structure modeling, enabling accurate protein structure and protein complex modeling using co-evolution-based predicted contacts. The importance of this approach to structural biology over the next decade will depend on the balance between two opposing trends: as more organisms are sequenced, the number of protein families with sufficient sequences for accurate modeling will increase, but as more structures are determined, there are fewer families for which accurate models cannot be produced by reliable comparative modeling methods. An increase in the number of eukaryotes sequenced—for example, by projects such as the recent Tara Ocean expedition (Bork et al., 2015; Sunagawa et al., 2015)—would make it possible to accurately model a large number of eukaryote-specific protein families of considerable biological interest. Because of the comparative difficulty in experimental structure determination, it is likely that co-evolution based prediction will continue to have the most impact for membrane proteins.

In this article, we present models for half of the large protein families in prokaryotes which do not currently have structures. The value of a comparable number of structures of eukaryotic protein families may justify the investment in genome sequencing of a diverse set of ~400 simple eukaryotes. For proteins not belonging to sufficiently large or diverse families but for which functional selections have been developed, it should be possible to develop experimental sequence covariation data sets by library generation, functional selection, and next generation sequencing. Significant resources were invested in the Protein Structure Initiative (PSI), with the initial goal ‘to make the three-dimensional, atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences (Burley et al., 2008)’. It is notable that structure models can now be generated for exactly the original class of proteins targeted by the PSI—large protein families without any available information—but at a small fraction of the cost.

3.6 MATERIALS AND METHODS

3.6.1 *Multiple sequence alignment generation*

Protein-coding genes were extracted from the *E. coli* (AUP000000625), *B. subtilis* (AUP000001570), *H. salinarum* (AUP000000554), *S. solfataricus* (AUP000001974) reference genomes in the UniProt proteome database (UniProt Consortium, 2014). Each protein from these proteomes was scanned against the PDB using HHsearch (-ssm 0, from HHsuite v. 2.0.15; [Remmert et al., 2012]) to identify proteins with no homologs of known structures (e-value of the top hit >1). We used two versions of the PDB database, one from 01 January, 2012 and one from 31 January, 2015. For the subset that had no hits in 2015, a multiple sequence alignment (MSA) was generated using Jackhmmer (-E 1E-20 -N 8, [Eddy, 2009]) and the uniref90 database (Suzek et al., 2007) from January, 2015. The alignments were filtered using HHfilter (-id 90 -cov 75), and positions that had more than 75% gaps were removed. To reduce redundancy, we constructed hidden Markov models (HMMs) using HHmake from each MSA and clustered the HMMs based on HHΔ (Kamisetty et al., 2013), a measure of HMM–HMM similarity. Families were assigned to the same cluster if the HHΔ was less than 0.5. The shortest *E. coli* protein was selected in each cluster; if no *E. coli* protein was in the cluster, a representative from *B. subtilis*, *H. salinrum*, or *S. solfataricus* was selected. Families for which the (number of sequences)/(length of representative protein) were greater than four were selected for modeling as described below. If the GREMLIN-predicted contacts (see below) were sparse and primarily between residues close along the linear sequence, the alignment was regenerated at an e-value 1E-40 cutoff, and the GREMLIN calculation repeated. If this resulted in too few sequences, the family was discarded (this eliminated six families).

TM protein domains that had a hit (e-value < 1E-20) in 2015, but no hit (e-value > 1) in 2012 were selected for the TM benchmark. Alignments were created using the UniProt sequence associated with the PDB, and trimmed at the N and C termini to match the crystal structure. We also include aquaporin (PDB: 1SOR_A) to test our protocol in modeling reentrant helices.

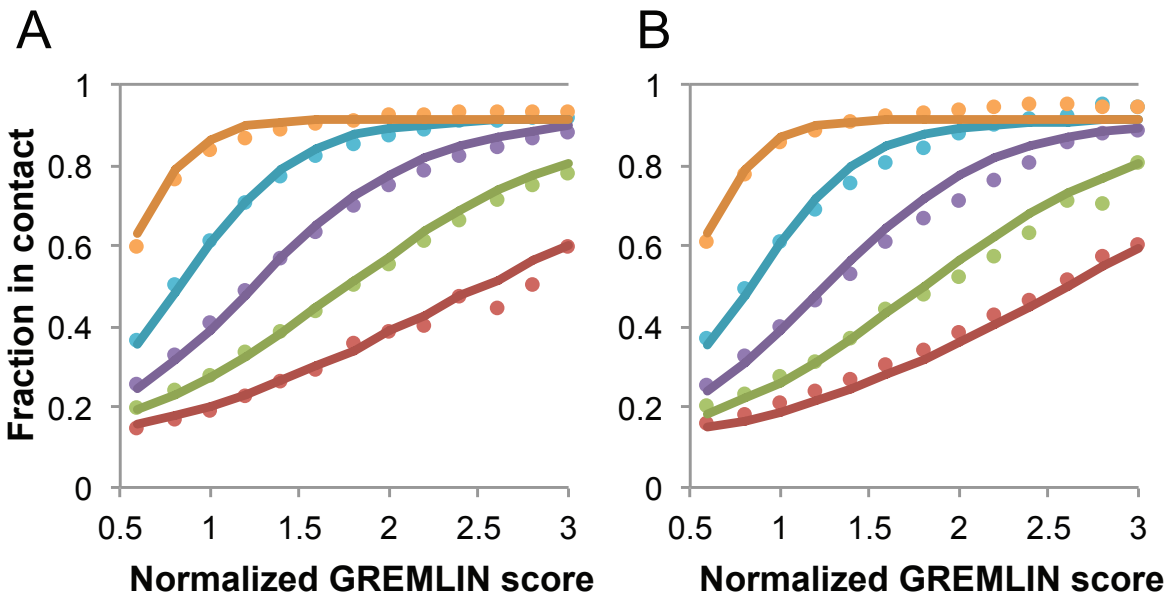


Figure 3.16. Dependence of the accuracy of predicted contacts on the normalized GREMLIN score (sco), the effective number of sequences (seq), the length (len), and the sequence separation (sep).

Contacts are defined based on amino acid specific Cβ-Cβ distance cutoffs as described in SI Table 3 in Kamisetty et al. (2013). (A) Observed vs predicted accuracies over a large data set of proteins of known structure with deep alignments (Supplementary file 3), sub sampled to different extents (seq/√(len) = 4 (red), 8 (green), 15 (purple), 32 (cyan), and 96 (orange)). Circles represent observed contact prediction accuracies, solid lines, a fit to a sigmoid function of the normalized coupling value, the number of sequences, the length, and the sequence separation (see Figure 16—figure supplement 1 and Figure 16—figure supplement 2). (B) Observed vs predicted accuracies in an independent data set of variable length alignments for 7047 pdb chains (Supplementary file 3), using maximum number of sequences obtained with HHblits as opposed to subsampling a large alignment. Circles again represent observed contact prediction accuracies; solid lines, the predicted accuracy using the model obtained by fitting to the data in (A). The contact prediction accuracy is correctly modeled for the independent data set, justifying its use on the unknown cases described in this article. The Equation use to calculate P(contact|sco,seq,len,sep) is

$$P(\text{contact}|\text{sco},\text{seq},\text{len},\text{sep}) \approx \frac{0.89\left(1-P(\text{contact}|\text{sep})\right)}{1+\exp\left(-0.58\left(\frac{\text{seq}}{\sqrt{\text{len}}}\right)^{0.50}\left(\text{sco}-5.46\left(\frac{\text{seq}}{\sqrt{\text{len}}}\right)^{-0.53}\right)\right)}+P(\text{contact}|\text{sep})$$

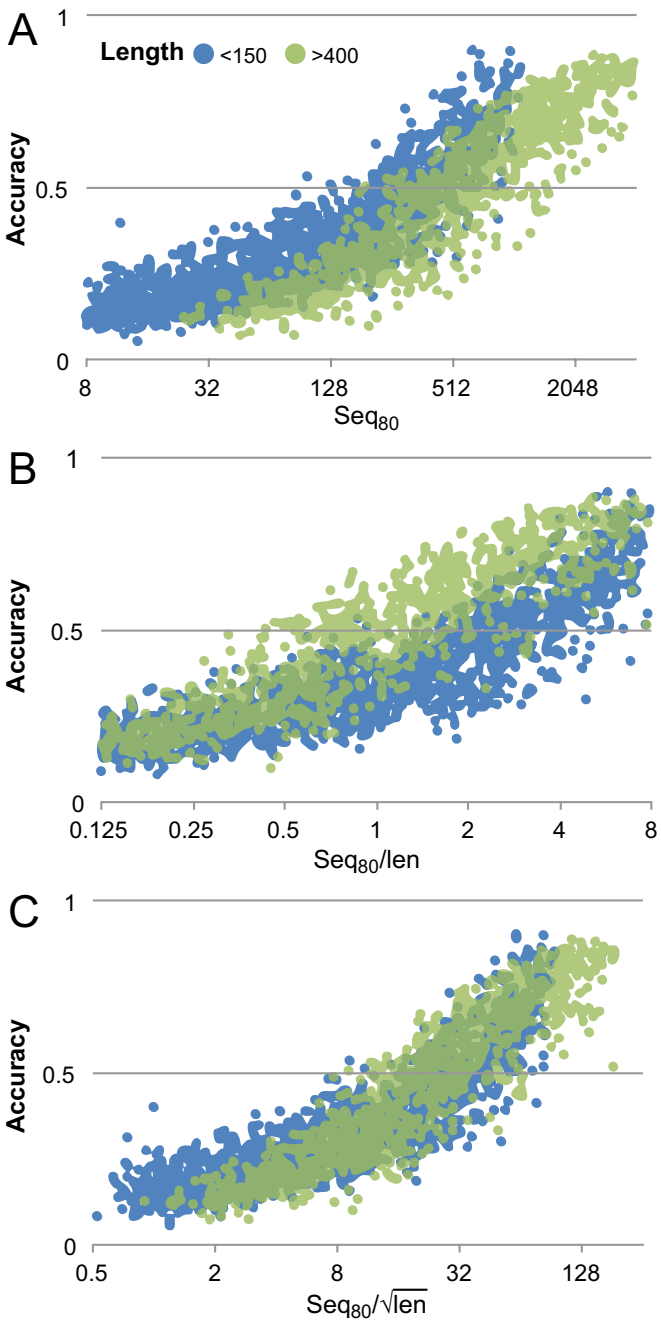


Figure 3.16—figure supplement 1. Contact prediction accuracy is better correlated with (#sequences/sqrt(length)) than with (#sequences/length). Accuracy is computed for the top 3L/2 GREMLIN predictions, with sequence separation ≥3, based on Cβ-Cβ amino acid specific distance as described in SI Table 3 in Kamisetty et al. (2013). The number of sequences after reducing the redundancy to 80% is shown. A set of 7047 pdb chains (see Supplemental file 3) was divided into two groups by length (less than 150 and greater than 400). (A) Larger proteins with similar number of sequence were less accurate then the smaller proteins. (B) #Sequences/length as often used does not accurately account for length dependence. There is a clear separation between the blue and green distributions. (C) #Sequences/√length better accounts for the length dependency. The blue and green distributions overlap.

3.6.2 *Contact prediction*

GREMLIN (v2.01) was used to learn a global statistical model of the sequences in large families using pseudolikelihood optimization (Balakrishnan et al., 2011). We previously reported that the accuracy of contact prediction using the residue–residue coupling values obtained from the model-fitting procedure is dependent on the number of sequences per length and the relative score (Kamisetty et al., 2013). To account for these dependencies, we constructed a model (Figure 16) that estimates the probability of being in contact using a pdb30 data set from PISCES (resolution limited to 2.5A or better, from 04 January, 2014; [Wang and Dunbrack, 2003]), with length of at least 100 residues. MSAs were generated for each of the 10,358 pdb chains using HHblits (-n 8 -e 1E-20 -maxfilt ∞ -neffmax 20 -nodiff -realign_max ∞), and HHfilter (-id 90 -cov 75) in the HHsuite (Remmert et al., 2012). The 3392 pdb chains with more than 10 sequences per length were subsampled to create MSAs with varying number of sequences, which were used to estimate probability of contact (Figure 16A). CCMPRED v0.1, a parallel implementation of GREMLIN (Seemayer et al., 2014), was used for the subsampled alignments. For CCMPRED, the default maximum number of iterations was modified to 100 to ensure convergence. The remaining 7047 pdb chains with less than 10L sequences were saved as a test set (Figure 16B, Figure 17). The top 3L/2 scores of residue pairs with sequence separation 3 or greater were normalized by rescaling the range so that the minimal value is 0.5 and average value is 1.0. Contact prediction accuracy was found to be a simple function of the residue–residue normalized coupling value, the number of sequences, the length (Figure 16—figure supplement 1), and the sequence separation as shown in Figure 16. The sigmoidal fit to these observed frequencies was used to estimate the probability of each contact being formed in the native structure.

To evaluate the significance of a match between predicted contacts and a model, we determined the expected total GREMLIN score over all contacts with sequence separation of 6 or greater using P (contact). To evaluate the fit of a particular model to a predicted contact set, we take the ratio of the actual total GREMLIN score of the model to the expected total score computed as above; we refer to this ratio of observed and expected contact scores as ‘Rc’ throughout the text. As shown in Figure 17, Rc ranges from 0.7 to 1.2 for native proteins, and from 0 to 0.3 when contact maps and structures are randomly paired. The Rc was evaluated over

the shortest overlap of the two lengths (contact map length vs pdb length). For homo-oligomer complexes, the Rc score includes all chains across all bio units.

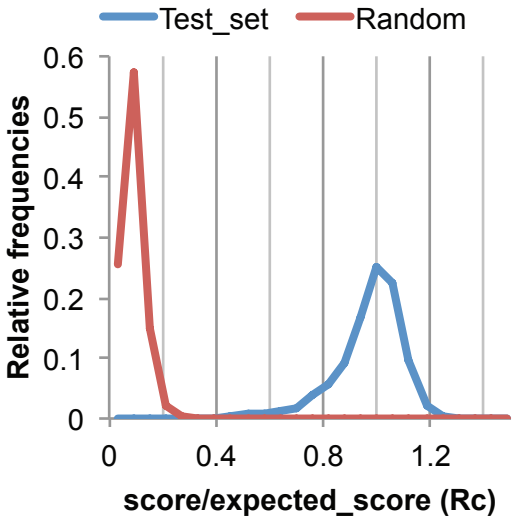


Figure 3.17. The Rc metric used to assess fit of predicted contacts to a model.
The expected total GREMLIN score if the structure was native was estimated by summing sco*P(contact|sco, seq, len, sep) over all contacts with sep ≥6. To evaluate the fit of a particular model to a predicted contact set, we take the ratio of the actual total GREMLIN score of the model to the expected total score computed as above; we refer to this ratio of observed and expected contact scores as 'Rc'. Blue line: the distribution of Rc in native structures with 4L–10L sequences; Red line: distribution of Rc after randomly reassigning contact predictions to structures. Rc values less than 0.7 are very infrequently observed for native structures; we use this value as a cutoff to evaluate the fit of a predicted contact set to a model.

3.6.3 *Co-evolution restraints and Rosetta energy function*

Residue-pair-specific distance restraints for use in the Rosetta structure prediction calculations were generated based on the normalized GREMLIN scores. Distance restraints were implemented as sigmoidal functions of the form: restraint(d) = weight(1 + exp(−slope(d − cutoff)) + intercept), where d is the distance between the constrained Cβ atoms (Cα in the case of glycine), the distance cutoffs and slopes are amino acid pair specific (SI Table 3 in Kamisetty et al., 2013), and the weight is the normalized Gremlin score multiplied by three to give the contact restraints roughly the same total dynamic range as the Rosetta energy. These distance restraints supplement the Rosetta energy function; the combination ensures the sampling of physically realistic structures consistent with the contact predictions. For TM proteins, the Rosetta energy function was modified to reflect the exposure of non-polar residues in the membrane-spanning

regions: the lazaridis-karplus solvation energy term weight was set to zero (fa_sol = 0.00), and to compensate for the short range repulsion implicit in the solvation model, the Lennard-Jones repulsive and attractive terms were given equal weights. We found this simple approach was equally effective and considerably less computationally intensive than the RosettaMembrane approach, which requires estimating the TM region for energy evaluation.

3.6.4 *Model generation*

The Rosetta ab initio protocol (Simons et al., 1999; Rohl et al., 2004) was used to generate 10,000 independent models guided by the covariance-derived restraints. For the benchmark set, fragments database from 2011 was used; for aquaporin, the fragments were filtered to remove any homologs with e-value < 1. After the generation of fragments for aquaporin, we examined the PDB files that contributed the most to the fragment set and verified that they did not contain aquaporin-like structures. The models generated by Rosetta ab initio were refined with an iterative version of the RosettaCM (Song et al., 2013) hybridization protocol used to refine models generated with contact information in CASP10 (Kim et al., 2014). In each iteration, 20 models are produced by recombination and minimization. In addition to the recombination of secondary structure chunks in the input models, fragment insertion was allowed in all positions. Iterations were continued until the procedure converged. For a 200-residue protein, the average runtime to produce a single model is about 30 min for RosettaAB and about 20 min for RosettaCM.

If in the initial Rosetta ab initio calculations, the top 10 models selected by restraint score converged (average pairwise TMscore [Zhang and Skolnick, 2004] > 0.8), the top five models were input directly into the iterative RosettaCM hybridization protocol. If models converged over substructures (average pairwise TMscore between 0.5 and 0.8), the top 10 models were first expanded by recombination to a population of 1000 structures, and the top five models were input into the RosettaCM hybridization protocol. If the Rosetta ab initio calculations did not converge (average pairwise TMscore < 0.5), we carried out an additional 10,000 Rosetta ab initio trajectories; if the top models did not converge, we considered the structure of the protein not accurately predictable using our approach. 15 of the 121 families were eliminated at this stage. We also eliminated families for which the models generated by the hybridization protocol did not satisfy the predicted contacts; 37 additional families were eliminated at this step.

Proteins over 400 amino acids for which there was little convergence of the lowest energy generated models were parsed into multiple domains (<200 residues) guided by the predicted contact information keeping overlaps of at least 50 residues between each domain, and Rosetta ab initio was used to generate models for each domain separately. If the overlapping regions in each domain converged during modeling, these were used to assemble the full model, otherwise the domains were trimmed to converged residues and docked using RosettaDock (Chaudhury et al., 2011).

If models converged overall in the Rosetta ab initio calculations but specific sets of contact restraints were consistently violated, we explored the possibility that the violations correspond to interactions between monomers in a homo-oligomer. To test for oligomeric contacts, docking was performed between two copies of the model using RosettaDock guided by the co-evolution-derived constraints

3.6.5 *Elimination of non-converging and unconstrained regions*

We developed a simple measure of convergence and contact violation after the hybridization protocol to trim regions with higher chance of being in error. The top five percent of the models were selected based on the sum of the Rosetta all atom energy and the contact restraint score and superimposed using THESEUS v3.1 (Theobald and Wuttke, 2006). The mean square deviation of the Cα coordinates of each residue was computed, and after smoothing with a Gaussian spanning three residues before and after the central residue, residues with MSD > $2Å^2$ were trimmed. We also eliminated residues in regions in which there were either very few contact restraints, or the majority of the restraints were violated.

For the benchmark set, the model closest to the average of the lowest energy 5% models was selected, and the RMSD to the native structure was computed over (1) the full length of the protein, (2) the converged and constrained residues, and (3) the residues structurally aligned using TM-align. The latter alignments are longer and more accurate, but selection of the subset of residues requires knowledge of the native structure; this is not the case for (2).

3.6.6 *Starting structures for homology modeling*

Our models provide starting templates to model any member of these families using comparative modeling, which requires relatively little computer time. To evaluate the protein space our 58

models cover, we carried out Jackhmmer search (-E 1E-20 -N 8, [Eddy, 2009]) with uniref100 database (Suzek et al., 2007) from January, 2015 and eliminated identical sequences and sequences aligned over less than 75% of the protein; the number of remaining sequences for each of the 58 families is listed in Table 2.

3.6.7 *Comparison to EvFold server*

To compare our method to the EvFold method, we submitted the alignments in our membrane protein benchmark to the EvFold web server. We compared the accuracy of the best of the 50 models generated by the server to our single selected model for each family in Tables 3, 4. The Rosetta models are considerably more accurate, but the EvFold server is orders of magnitude faster. We compare to the server results rather than to the results in the previously published Evfold paper as this would be unfair since there were fewer available sequences at the time the paper was written and the contact prediction method (mfDCA) was somewhat less accurate. For the server comparison, we chose to predict contacts using the PLM option as this is very similar to GREMLIN.

Table 3.3. Comparison of methods on CASP11 targets.
*Full-length Cα-RMSD and GDT-TS calculation based on the best of five models submitted to CASP11 from BAKER and Jones-UCL groups. For Evfold, the values for best of 50 models generated by the web server are reported, sorted by full-length Cα-RMSD. For the comparison, the alignments used during CASP11 were provided as input to the Evfold-web server, with PLM option selected. For T0824, the minimal number of sequence limit was set to 0 to allow Evfold-web server to run. PLM, pseudo-likelihood.

	BAKER*		Jones-UCL*		Evfold-webserver	
Targets	Cα-RMSD	GDT-TS	Cα-RMSD	GDT_TS	Cα-RMSD	GDT-TS
T0806	3.6	60.4	6.8	34.3	8.2	30.0
T0824	4.2	55.3	9.2	41.4	8.1	32.6

Table 3.4. Comparison of methods on transmembrane benchmark set.
The Cα-RMSD and GDT-TS calculations are over the full sequence. For Evfold web server results, we report the best Cα-RMSD of 50 models returned. For the comparison, the alignments we used were provided as input to the Evfold webserver, and the pseudo-likelihood method was selected.

	BAKER		Evfold-webserver	
Targets	Cα-RMSD	GDT-TS	Cα-RMSD	GDT-TS
4HE8_H	4.9	54.5	5.3	50.3
1SOR_A (Aquaporin)	2.7	69.7	6.1	44.5
4Q2E_A	5.4	45.6	12.9	21.7
4HTT_A	3.9	60.6	6.4	41.8
4P6V_E	5.0	56.6	7.4	31.8
4J72_A	6.6	67.1	12.9	33.8
3V5U_A	3.9	58.8	4.6	47.1
4PGS_A	3.5	66.3	4.6	48.1
4QTN_A	4.2	59.6	4.9	51.4
4OD4_A	3.9	55.6	4.1	53.4
4O6M_A	4.1	64.0	11.2	33.0

3.7 DATA AVAILABILITY

External Database: http://gremlin.bakerlab.org/structures/ the external database provides the models generated and links to the GREMLIN web server output. The web server output provides the predicted contacts, overlay of the predicted contacts on the top 10 HHsearch pdb hits, restraints used in modeling and the alignment used for contact prediction. We also provide contact predictions for all protein coding genes with enough sequences for all of the four model organisms used in this paper. The models are also available at Dryad (Ovchinnikov et al., 2015).

3.8 ACKNOWLEDGEMENTS

We thank Per Jr. Greisen, Robert Gennis, Ranjani Murali, and Schara Safarian for comments and analysis of CydA/CydB complex, Andrew HJ Wang, and Jason Chou for comments on UppP, Tamir Gonen for comments on YitE, and Mark A Wilson and Rick Lewis for permission to disclose the unpublished structures of YaaA (T0806) and NucB (T0824). We also thank Rosetta@home and Charity engine participants for donating their computer time, Nicole Silvester from ENA and the CASP11 organizers. This work was funded by the NIH/NIGMS and the Welch Foundation.

3.9 REFERENCES

1. Abriata LA. 2015. An homology-and coevolution-consistent structural model of bacterial copper-tolerance protein CopM supports function as a ‘metal sponge’ and suggests regions for metal-dependent interactions with other proteins. bioRxiv:013581.
2. Anantharaman V, Koonin EV, Aravind L. 2002. SPOUT: a class of methyltransferases that includes spoU and trmD RNA methylase superfamilies, and novel superfamilies of predicted prokaryotic RNA methylases. Journal of Molecular Microbiology and Biotechnology 4:71–75.
3. Antala S, Ovchinnikov S, Kamisetty H, Baker D, Dempski RE. 2015. Computation and functional studies provide a model for the structure of the Zinc transporter hZIP4. The Journal of Biological Chemistry 290:17796–17805.
4. Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ. 2011. Learning generative models for protein fold families. Proteins 79:1061–1078.
5. Borisov VB, Gennis RB, Hemp J, Verkhovsky MI. 2011. The cytochrome bd respiratory oxygen reductases. Biochimica et Biophysica Acta 1807:1398–1413.
6. Bork P, Bowler C, de Vargas C, Gorsky G, Karsenti E, Wincker P. 2015. Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. Science 348:873.
7. Burley SK, Joachimiak A, Montelione GT, Wilson IA. 2008. Contributions to the NIH-NIGMS protein structure initiative from the PSI Production centers. Structure 16:5–11.
8. Burroughs AM, Aravind L. 2014. Analysis of two domains with novel RNA-processing activities throws light on the complex evolution of ribosomal RNA biogenesis. Frontiers in Genetics 5:424.
9. Chang HY, Chou CC, Hsu MF, Wang AH. 2014. Proposed carrier lipid-binding site of undecaprenyl pyrophosphate phosphatase from Escherichia coli. The Journal of Biological Chemistry 289:18719–18735.
10. Chaudhury S, Berrondo M, Weitzner BD, Muthu P, Bergman H, Gray JJ. 2011. Benchmarking and analysis of protein docking performance in Rosetta v3.2. PLOS ONE 6:e22477.
11. Das R, Baker D. 2008. Macromolecular modeling with rosetta. Annu Rev Biochem 77:363–382.
12. DiMaio F, Terwilliger TC, Read RJ, Wlodawer A, Oberdorfer G, Wagner U, Valkov E, Alon A, Fass D, Axelrod HL, Das D, Vorobiev SM, Iwa¨ı H, Pokkuluri PR, Baker D. 2011. Improved molecular replacement by density- and energy-guided protein structure optimization. Nature 473:540–543.
13. Duran AM, Meiler J. 2013. Inverted topologies in membrane proteins: a mini-review. Computational and Structural Biotechnology Journal 8:e201308004.
14. Dutzler R, Campbell EB, Cadene M, Chait BT, MacKinnon R. 2002. X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. Nature 415:287–294.
15. Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. Genome Informatics 23:205–211.
16. Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: using ¨ pseudolikelihoods to infer Potts models. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics 87: 012707.
17. El Ghachi M, Derbise A, Bouhss A, Mengin-Lecreulx D. 2005. Identification of multiple genes encoding membrane proteins with undecaprenyl pyrophosphate phosphatase (UppP) activity in Escherichia coli. The Journal of Biological Chemistry 280:18689–18695.
18. Ellermeier CD, Losick R. 2006. Evidence for a novel protease governing regulated intramembrane proteolysis and resistance to antimicrobial peptides in Bacillus subtilis. Genes & Development 20:1911–1922.
19. Fleishman SJ, Harrington SE, Enosh A, Halperin D, Tate CG, Ben-Tal N. 2006. Quasi-symmetry in the cryo-EM structure of EmrE provides the key to modeling its transmembrane domain. Journal of Molecular Biology 364: 54–67.
20. Fraipont C, Alexeeva S, Wolf B, van der Ploeg R, Schloesser M, den Blaauwen T, Nguyen-Disteche M. 2011. The integral membrane FtsW protein and peptidoglycan synthase PBP3 form a subcomplex in Escherichia coli. Microbiology 157:251–259.
21. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Research 43:D261–D269.
22. Gonen T, Cheng Y, Sliz P, Hiroaki Y, Fujiyoshi Y, Harrison SC, Walz T. 2005. Lipid-protein interactions in doublelayered two-dimensional AQP0 crystals. Nature 438:633–638.
23. Hayat S, Sander C, Elofsson A, Marks DS. 2014. Accurate prediction of transmembrane β-barrel proteins from sequences. bioRxiv:006577.

24. Hebert H, Jegerschold C. 2007. The structure of membrane associated proteins in eicosanoid and glutathione ¨ metabolism as determined by electron crystallography. Current Opinion in Structural Biology 17:396–404.
25. Hommais F, Krin E, Coppee JY, Lacroix C, Yeramian E, Danchin A, Bertin P. 2004. GadE (YhiE): a novel activator involved in the response to acid environment in Escherichia coli. Microbiology 150:61–72.
26. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. 2012. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149:1607–1621.
27. Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R. 2015. Amino acid coevolution reveals threedimensional structure and functional domains of insect odorant receptors. Nature Communications 6:6077.
28. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS. 2014. Sequence ¨ co-evolution gives 3D contacts and structures of protein complexes. eLife 3:e03430.
29. Jakobsson PJ, Morgenstern R, Mancini J, Ford-Hutchinson A, Persson B. 1999. Common structural features of MAPEG—a widespread superfamily of membrane associated proteins with highly divergent functions in eicosanoid and glutathione metabolism. Protein Science 8:689–692.
30. Jones DT, Buchan DW, Cozzetto D, Pontil M. 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28:184–190.
31. Kamisetty H, Ovchinnikov S, Baker D. 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proceedings of the National Academy of Sciences of USA 110: 15674–15679.
32. Kim DE, Dimaio F, Yu-Ruei Wang R, Song Y, Baker D. 2014. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. Proteins 82(Suppl 2):208–218.
33. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D. 2012. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proceedings of the National Academy of Sciences of USA 109:10873–10878.
34. Lauble H, Kennedy MC, Beinert H, Stout CD. 1994. Crystal structures of aconitase with trans-aconitate and nitrocitrate bound. Journal of Molecular Biology 237:437–451.
35. Lazaridis T, Karplus M. 1999. Effective energy function for proteins in solution. Proteins 35:133–152.
36. Lee C, Kang HJ, von Ballmoos C, Newstead S, Uzdavinys P, Dotson DL, Iwata S, Beckstein O, Cameron AD, Drew D. 2013. A two-domain elevator mechanism for sodium/proton antiport. Nature 501:573–577.
37. Manolaridis I, Kulkarni K, Dodd RB, Ogasawara S, Zhang Z, Bineva G, O'Reilly N, Hanrahan SJ, Thompson AJ, Cronin N, Iwata S, Barford D. 2013. Mechanism of farnesylated CAAX protein processing by the intramembrane protease Rce1. Nature 504:301–305.
38. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D structure computed from evolutionary sequence variation. PLOS ONE 6:e28766.
39. Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. Nature Biotechnology 30:1072–1080.
40. Matsumoto S, Shimada A, Nyirenda J, Igura M, Kawano Y, Kohda D. 2013. Crystal structures of an archaeal oligosaccharyltransferase provide insights into the catalytic cycle of N-linked protein glycosylation. Proceedings of the National Academy of Sciences of USA 110:17868–17873.
41. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of USA 108:E1293–E1301.
42. Morrison EA, DeKoster GT, Dutta S, Vafabakhsh R, Clarkson MW, Bahl A, Henzler-Wildman KA. 2012. Antiparallel EmrE exports drugs by exchanging between asymmetric structures. Nature 481:45–50.
43. Muth T, Garc´ıa-Mart´ın JA, Rausell A, Juan D, Valencia A, Pazos F. 2012. JDet: interactive calculation and visualization of function-related conservation patterns in multiple sequence alignments and structures. Bioinformatics 28:584–586.
44. Nugent T, Jones DT. 2012. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proceedings of the National Academy of Sciences of USA 109:E1540–E1547.
45. Ovchinnikov S, Kamisetty H, Baker D. 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife 3:e02030.
46. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D. 2015. Data from: Large scale determination of previously unsolved protein structures using evolutionary information. Dryad Digital Repository.

47. Pailler J, Aucher W, Pires M, Buddelmeijer N. 2012. Phosphatidylglycerol::prolipoprotein diacylglyceryl transferase (Lgt) of Escherichia coli has seven transmembrane segments, and its essential residues are embedded in the membrane. Journal of Bacteriology 194:2142–2151.
48. Pei J, Grishin NV. 2001. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712.
49. Pei J, Li W, Kinch LN, Grishin NV. 2014. Conserved evolutionary units in the heme-copper oxidase superfamily revealed by novel homologous protein families. Protein Science 23:1220–1234.
50. Pei J, Millay DP, Olson EN, Grishin NV. 2011a. CREST–a large and diverse superfamily of putative transmembrane hydrolases. Biology Direct 6:37.
51. Pei J, Mitchell DA, Dixon JE, Grishin NV. 2011b. Expansion of type II CAAX proteases reveals evolutionary origin of γ-secretase subunit APH-1. Journal of Molecular Biology 410:18–26.
52. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 77(Suppl 9):89–99.
53. Rapp M, Granseth E, Seppal ¨ a S, von Heijne G. 2006. Identification and evolution of dual-topology membrane ¨ proteins. Nature Structural & Molecular Biology 13:112–116.
54. Reaney SK, Begg C, Bungard SJ, Guest JR. 1993. Identification of the L-tartrate dehydratase genes (ttdA and ttdB) of Escherichia coli and evolutionary relationship with the class I fumarase genes. Journal of General Microbiology 139:1523–1530.
55. Remmert M, Biegert A, Hauser A, Soding J. 2012. HHblits: lightning-fast iterative protein sequence searching by ¨ HMM-HMM alignment. Nature Methods 9:173–175.
56. Rohl CA, Strauss CE, Misura KM, Baker D. 2004. Protein structure prediction using Rosetta. Methods in Enzymology 383:66–93. doi: 10.1016/S0076-6879(04)83004-0. Screpanti E, Hunte C. 2007. Discontinuous membrane helices in transport proteins and their correlation with function. Journal of Structural Biology 159:261–267.
57. Seemayer S, Gruber M, Soding J. 2014. CCMpred–fast and precise prediction of protein residue-residue contacts ¨ from correlated mutations. Bioinformatics 30:3128–3130.
58. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34:82–95.
59. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D. 2013. High-resolution comparative modeling with RosettaCM. Structure 21:1735–1742.
60. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348:1261359.
61. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288.
62. Sułkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. 2012. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences of USA 109:10340–10345.
63. Sa-Pessoa J, Paiva S, Ribas D, Silva IJ, Viegas SC, Arraiano CM, Casal M. 2013. SATP (YaaH), a succinate-acetate transporter protein in Escherichia coli. The Biochemical Journal 454:585–595.
64. Soding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960.
65. Theobald DL, Wuttke DS. 2006. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics 22:2171–2172.
66. Tian P, Boomsma W, Wang Y, Otzen DE, Jensen MH, Lindorff-Larsen K. 2015. Structure of a functional amyloid protein subunit computed using sequence variation. Journal of the American Chemical Society 137:22–25.
67. UniProt Consortium. 2014. Activities at the Universal protein resource (UniProt). Nucleic Acids Research 42: D191–D198.
68. Villar HO, Kauvar LM. 1994. Amino acid preferences at protein binding sites. FEBS Letters 349:125–130.
69. Wang G, Dunbrack RL. 2003. PISCES: a protein sequence culling server. Bioinformatics 19:1589–1591.

70. Wickles S, Singharoy A, Andreani J, Seemayer S, Bischoff L, Berninghausen O, Soeding J, Schulten K, van der Sluis EO, Beckmann R. 2014. A structural model of the active ribosome-bound membrane protein insertase YidC. eLife 3:e03035.
71. Xu Q, Dunbrack RL. 2012. Assignment of protein sequences to existing domain and family classification systems: Pfam and the PDB. Bioinformatics 28:2763–2772.
72. Yang Y, Faraggi E, Zhao H, Zhou Y. 2011. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27:2076–2082.
73. Yernool D, Boudker O, Jin Y, Gouaux E. 2004. Structure of a glutamate transporter homologue from Pyrococcus horikoshii. Nature 431:811–818.
74. Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. Proteins 57:702–710.
75. Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33:2302–2309.
76. Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8: 357–366.

Chapter 4. PROTEIN STRUCTURE DETERMINATION USING METAGENOME SEQUENCE DATA

A version of this chapter has been previously published as:

Ovchinnikov, Sergey, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. "Protein structure determination using metagenome sequence data." *Science* 355, no. 6322 (2017): 294-298.

4.1 ABSTRACT

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 triples 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 PDB. 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.

4.2 MAIN TEXT

There are 14849 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 NMR structure, and an additional 3984 for which reliable comparative models can be built based on homologues 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). There is 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 homologues 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*).

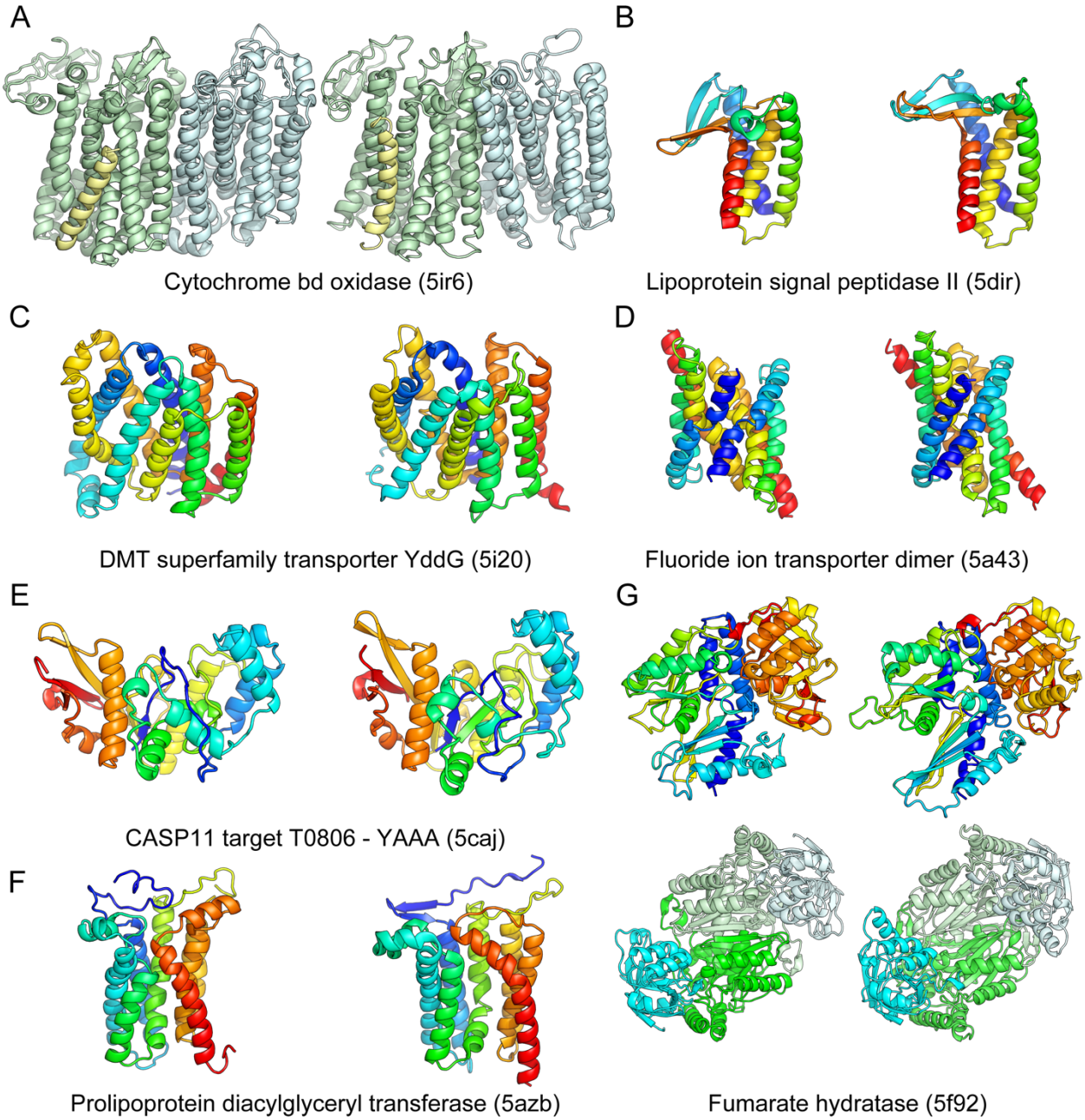


Figure 4.1. Comparison of Rosetta models (left) to subsequently published crystal structures (right).
The models accurately recapitulate the structural details of A) the Cytochrome bd oxidase (TMalign score 0.88) B) the Lipoprotein signal peptidase II (TMalign score 0.70) C) the DMT superfamily transporter YddG (TMalign score 0.70) D) the Fluoride ion transporter dimer (TMalign score 0.69) E) the CASP11 target T0806 F) Prolipoprotein diacylglyceryl transferase (TMalign score 0.69) and G) Fumarate hydratase (TMalign score 0.80 for monomer (top) and 0.76 for dimer (bottom)).

The increase in the number of known amino acid sequences has enabled the accurate prediction of residue-residue contacts 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. While coevolution based structure modeling has been used to generate models for individual proteins with fold-level accuracy (TMscore (23) > 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, providing an opportunity to assess this medium scale prediction effort. Recently solved structures of the Lipoprotein signal peptidase II (*24*), Prolipoprotein diacylglyceryl transferase (*25*), the fluoride ion transporter (*26*), cytochrome bd oxidase (*27*), DMT superfamily transporter YddG (*28*), and fumerate hydratase (*29*) are all very close to computational models published and publicly released well before the structures were solved (Figure 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 3 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 based on 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 *E. 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 (Figure 1A; TMalign 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 quite complex protein structure in the CASP11 structure prediction experiment (*31*) (Figure 1E). In all of the cases shown in Figure 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 quite accurate models (in all 6 cases, the TMalign score is greater than 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 supplemental text).

Structure models with the accuracy of those in Figure 1 would have broad utility for framing biological hypotheses about function and interpreting mutational data, as well as 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 as well as 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 Supplementary info) that employs 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 PDB (Protein Data Bank), while 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 since 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 (Figure 2A solid line; Figure S1; see Supplementary info). 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 (Nf, the 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 (Figure 2A, Figure S1) over a broad range of families.

How many protein families with currently unknown structure have Nf values in the range where accurate models can be built? The models in Figure 1 were all generated for families with Nf > 64; accuracy falls off for lower values of Nf (Figure 2A). As shown in Figure 2B, less than 8% of families have Nf values of 64 or better. Modeling the remaining 92% of families of

unknown structure at reasonable accuracy is not currently possible using the sequence information in the UniRef100 database (*35*).

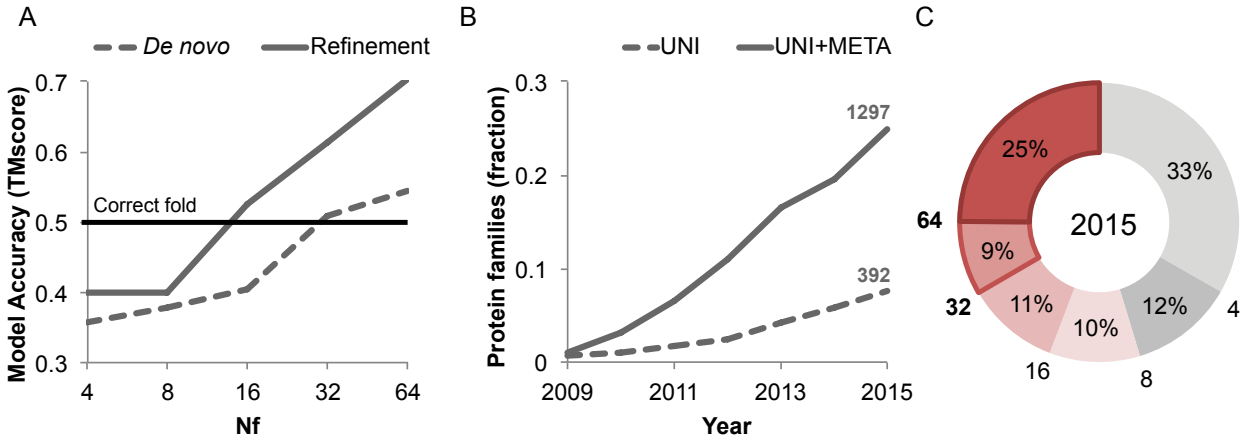


Figure 4.2. Metagenome data greatly increased fraction of structures which can be accurately modeled.
A) Dependence of coevolution guided Rosetta structure prediction accuracy on the effective number of sequences Nf (a function of both sequence number and diversity; see Methods definition) in the protein family. For each of 27 proteins of known structure, the multiple sequence alignment was subsampled and residue-residue contacts predicted using GREMLIN. Rosetta structure prediction calculations were then used to generate ~20,000 models, and a single model was selected based on the Rosetta energy and the fit to the coevolution constraints; the average TMscore 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 (TMscore > 0.5) are obtained for Nf ≥ 16, and models with accuracy typical of comparative modeling, for Nf of 64. B) Fraction of protein families of unknown structure with at least 64 Nf. Dashed line: including only sequences in UniRef100 database; solid line: including sequences in UniRef100 database together with metagenome sequence data from JGI (*37*). C) Distribution of Nf values for 5211 PFAM families with currently unknown structure, after the addition of metagenomic sequences; 25% of the protein-families have Nf > 64, 34% have Nf > 32 and 45% have Nf > 16.

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 (Figure 2B, solid versus dashed line). 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 Nf ≥ 64, the fraction increases from 0.08 to 0.25, and at Nf ≥ 32 (where fold level accuracy can be achieved (Figure 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 benchmark of 81 PFAMs with recently solved structures and Nf ≥ 64 (Figure S1E-F, 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 (Figure S1F and S2). For 42 families, the predictions converged with most of the predicted contacts satisfied (see Supplementary information for convergence criteria) and of these, 25 had a TMscore > 0.7 and 13 a TMscore > 0.6 (in 3 of the 4 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 (Figure S13)).

We generated coevolution based contact predictions using GREMLIN (*4*, *12*) for the 1297 protein families with Nf ≥ 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 greater than 0.65 for 614 of the 1024 domains according to the benchmarks. A list of the PFAM families covered by these models is in Table S3; the models are available at <https://gremlin2.bakerlab.org/meta.php>, 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 IMG 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 if the models belong to known protein folds, we carried out structure-structure comparisons against the 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 Xho1 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 (TMalign 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 Figure 3. They include the key developmental regulator Chordin, CobS a key enzyme in cobalbumin synthesis, a metalloendopeptidase, and mercury and iron transporters; six are transmembrane proteins, four have new folds and several have quite complex topologies. These and the remaining 590 structure models not shown in Figure 3 should provide a basis for understanding molecular function, mechanism and 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 five of the 614 families were published and are very similar to the corresponding models (TMalign score ≥ 0.7; See Figure 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 five years ago. This progress required integration of advances in quite 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 Figure 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 is for prokaryotes, but as fungal and other simple eukaryote genome sequencing projects ramp up the approach should become applicable to eukaryote specific protein families.

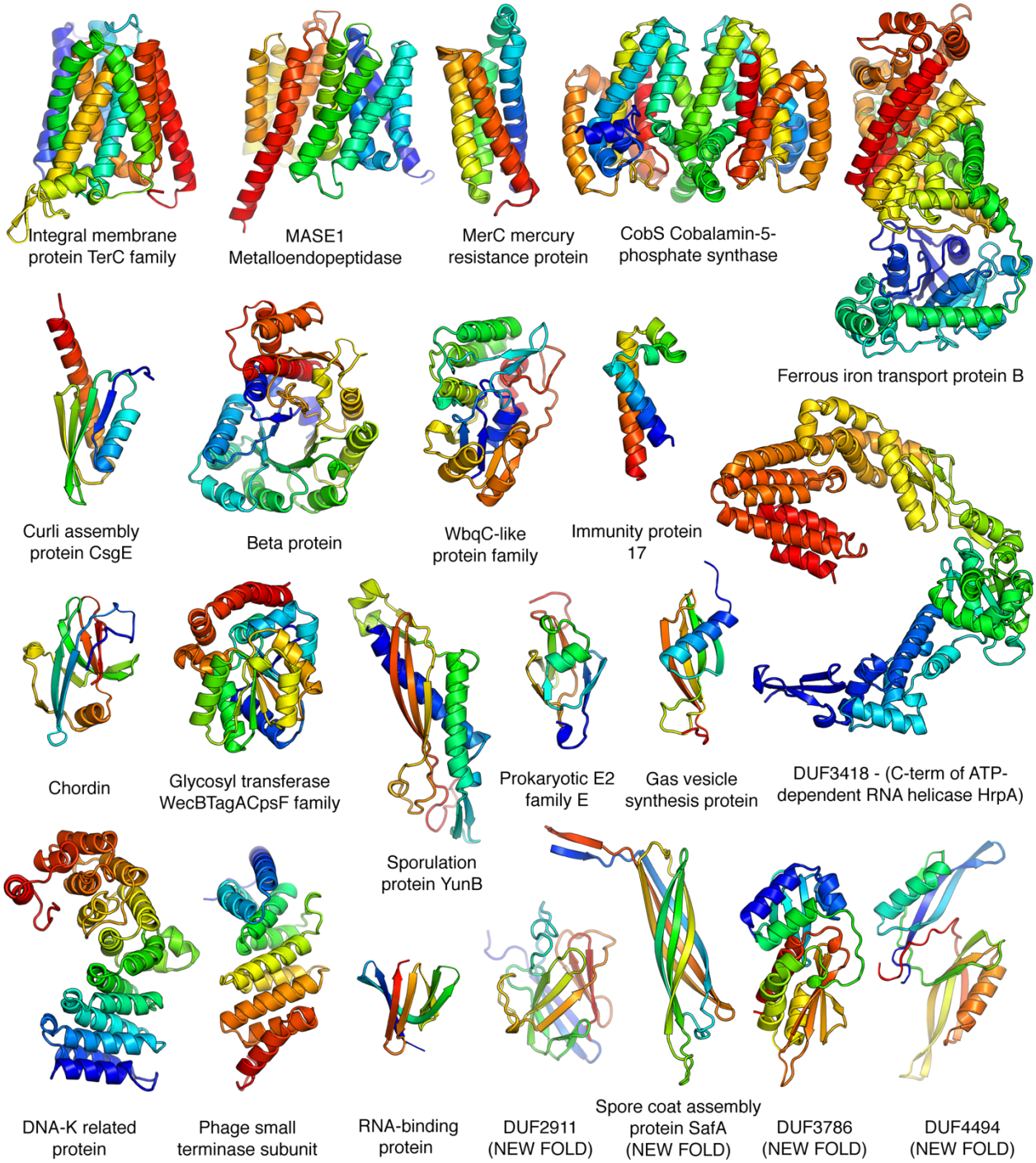


Figure 4.3. Representative structure models for selected PFAM families.
Membrane proteins are on the top row; new folds on the bottom right. The multi-domain models of the iron transporter and RNA helicase and the dimeric model of CobS, an enzyme in vitamin B synthesis, are guided by both intra- and inter-chain coevolution restraints.

4.3 MATERIALS AND METHODS

4.3.1 *Metagenome sequences*

The Integrated Microbial genomes (IMG) database (*37*) is a publicly available, comprehensive resource consisting of assembled and annotated metagenomes, the genomic information of which has been integrated successfully with isolate genomes from all three domains of life. The metagenomes present in IMG are a combination of Joint Genome Institute (JGI)-sequenced projects and data deposited by users. An initial dataset of over 2 billion proteins, predicted on assembled contigs, from ~5000 metagenomes in IMG served as the initial dataset. Both partial and full length proteins were included in this set in order to maximize the search space.

The *hmmsearch* tool from the HMMER package version 3.1b1 (*41*) was used to search the dataset of proteins described above, with each PFAM HMM (*1*) as the query, using the trusted cutoffs for each HMM. The results were then filtered to only retain those protein hits, that covered at least 75% of the PFAM query. The protein sequence of the retained gene hits were then extracted from the IMG database and used in the further analysis. These we will refer to as "metagenomic PFAM sequences" in remainder of the methods section.

4.3.2 *Nf (effective number of sequences) calculation*

The Hamming distance is computed between all of the sequences in a given MSA (multiple sequence alignment). Each sequence is given a weight of 1/(number of sequences > 80% identity). These weights are used within GREMLIN (*4*, *12*) to downweight redundant sequences, and the sum of these weights divided by square-root of length of the MSA is used for the Nf calculation. 80% identity threshold was chosen because it results in both the best accuracy for predicted contacts and the best correlation to accuracy for the Nf calculation (Figure S5).

To test the effects of re-weighting, the dataset from (ref *21*; Supplementary file 3) was used. The alignments were filtered to remove sequences that do not cover at least 75% of the query sequences. Positions that had more than 50% gaps from previous filtering (HHfilter version 2.0.15; -id 90 -cov 75)(*2*) were removed. For this test, only identical sequences were removed. Alignments with less than 100 unique sequences were excluded.

4.3.3 *Contact prediction*

In some cases, the domain boundaries defined in PFAM are not structurally realistic, this is evident when there are strong contacts that stretch beyond the PFAM boundaries, or when the N and C-terminal regions are split into two domains, yet there are extensive contacts between the two. In both cases, modeling a representative sequence alleviates this issue. For each PFAM with at least 64 Nf, *hmmsearch* (from HMMER version 3.1b1; default parameters) (*41*) was used to identify a representative sequence from either the SWISS-PROT or reference genome database. Starting with this sequence, HHblits (from HHsuite version 2.0.15; -n 8 -e 1E-20 -maxfilt ∞ -neffmax 20 -nodiff -realign_max ∞) (*42*) was run against the clustered UniProt database from 2015_06 to generate an initial alignment. After HHfilter (-id 90 -cov 75) this initial alignment was used to construct a hidden Markov model (HMM) using *hmmbuild* (from HMMER version 3.1b1) (*41*). A conservative bit-score value of 27 (this is typically the cutoff used for PFAM definitions) was used to search against a master database containing both the UniRef100 and metagenomic PFAM sequences. Bit-score instead of E-value was used, because it is independent of database size. The UniRef100 contains all the UniProtKB records plus the UniPrac records not covered by the UniProtKB (*35*). The output was again filtered using HHfilter (-id 90 -cov 75). If more than 2 fold fewer sequences were recovered compared to the PFAM alignment, the representative sequence was trimmed to the PFAM boundaries and the alignment was recreated (for the 81 protein benchmark described below, such cases were discarded and hence there was no trimming to PFAM boundaries). Following the filtering, positions that have more than 50% gaps are removed. GREMLIN (v2.01) with default parameters was used for contact prediction (*12*). CCMPRED (v0.1), a parallel implementation of GREMLIN (*43*), was used for the subsampled alignments and reweighting test. For CCMPRED, the default maximum number of iterations was modified to 100 to ensure convergence.

4.3.4 *Contact map alignment*

For the purposes of identifying structural homologs and characterizing protein families, we developed a contact map alignment method called *map_align*. *map_align* uses an iterative double dynamic programming algorithm similar to the one first described by Taylor (*33*) for

protein structure comparison, with some modifications for our purposes to produce an alignment that optimizes structural overlap.

The *map_align* algorithm comprises two dynamic programming steps. In the first step, a score is computed for each row (corresponding to a specific residue) of the first contact map with each row for the second contact map. This score is the sum of Gaussian functions: $\exp(-x^2/(2y^2))$, where x is the difference in sequence separation of aligned contacts and y (or standard deviation) is a function of the smaller of the two sequence separations, as described below. Dynamic programming is used to find the alignment of the contacts for the two rows being matched which maximizes the sum of these Gaussian functions. These optimized sums of scores are then entered in a second matrix, and the optimal contact alignment is then found by dynamic programming, using the Smith–Waterman algorithm (*44*). At this point, however, the scores for individual row-row comparisons are overestimates since in the first step the alignments for each pair are independent. We then update the second step similarity matrix based on the current alignment, and carry out the second step dynamic programming again (although the score is quite different, this updating strategy is similar to that used in Taylor's method (*33*) for comparing two structures based on distance matrices). This process is repeated 20 times (by 20th iteration, the alignment convergences and no more contacts are made). The initial estimate of the similarity matrix is critical in getting at least part of the alignment correct as this serves as a nucleation point for aligning the rest of the contacts. To maximize the chance of success, we try a number of variations in the first step. The standard deviation in the Gaussian function is made either a constant, linear and quadratic function of the lower of the two sequence separations, and a range of scaling parameters are tested. For each choice of functional form and scaling parameter, we carry out the full iterative optimization described above, and choose that alignment which best matches the two contact maps (maximizes the number of aligned contacts) assigning lower weight to low sequence separation contacts (weight of 0.50 for sequence separation ≤ 4, 0.75 for 5 and 1.00 for ≥ 6; the weight is based on the lower of the two sequence separations). The pseudocode of the aligorithm is provided at the end of this section.

Figure S7 compares our approach against A_purva (*45*), GR-align (*46*), Al-Eigen (*47*) and MSVNS (*48*). A_purva is an "exact" branch and bound algorithm which if run long enough will find the maximum overlap of contacts. The problem is that it can take a long time to run for a protein pair, which is not practical for database search, it is useful to obtain the best possible

alignment, once ranking has been achieved with approximate methods. We use the results of A_purva to establish the baseline for the Skolnick dataset of 40 SCOP domains(*45*). GR-align and AI-Eigen use a Needleman-Wunsch algorithm (*49*) where the cost for matching two residues is based on "graphlet" degree similarity and weighted eigenvalues respectively. MSVNS is a stochastic local search method that aims to find good solutions by adding and removing pairs of residues using different strategies. *Map_align* is able to find better contact map alignments on average because of the iterative updating of the similarity matrix.

The MRFalign (*50*) program elegantly compares two sequence families evaluating whether the co-evolution patterns are similar, and achieves very sensitive remote homologue detection. For our model building purposes, we are searching for structural fragments which fit contacts independent of evolutionary relatedness, and hence direct comparison of co-evolution derived predicted contact map to contact maps from known structure is most useful.

The source code for the algorithm can be downloaded from GitHub:
http://github.com/sokrypton/map_align.

The Pseudocode for map_align algorithm for aligning two input contact maps (map_a and map_b) is provided below.

```
for sep_x (0,1,2)
	for sep_y (1,2,4,8,16,32)
		ini_mtx = initialize_matrix(sep_x, sep_y)
		for gap_e (-0.2,-0.1,-0.01,-0.001)
			mtx = ini_mtx
			alignment = get_alignment(mtx,-1,gap_e)
			score = SWalign(mtx,-1,-0.01)/2
			if (score > best_score)
				best_alignment = alignment
				best_score = score
print(best_alignment)

function initialize_matrix(sep_x,sep_y)
	for ai (columns in map_a)
		for bi (columns in map_b)
			for aj (values in column ai)
				for bj (values in column bi)
					sa = ai-aj
					sb = bi-bj
					if (sa>0 and sb>0 or sa<0 and sb<0)
						s_dif = ||sa|-|sb||
						s_min = min(|sa|,|sb|)
						s_std = sep_y*(1+pow(s_min-2,sep_x))
						w = sep_weight(s_min)*gaussian(0,s_std,s_dif)
						M[aj][bj] = map_a[ai][aj] * map_b[bi][bj] * w
					else
						M[aj][bj] = -1
			mtx[ai][bi] = SWalign(M,0,0)
	return mtx

function get_alignment(mtx,gap_open,gap_extention)
	for i (0..20) //iterate
		alignment = SWalign(mtx,gap_open,gap_extention)
		for ai (columns in map_a)
			for bi (columns in map_b)
				sco = 0
				for aj (values in column ai)
					bj = alignment[aj]
					sa = ai-aj
					sb = bi-bj
					if (sa>0 and sb>0 or sa<0 and sb<0)
						w = sep_weight(min(|sa|,|sb|))
						sco += map_a[ai][aj] * map_b[bi][bj] * w
				mtx[ai][bi] = i/(i+1) * mtx[ai][bi] + sco/(i+1)
	return alignment

function sep_weight(sequence_seperation)
	if (sequence_seperation <= 4) return 0.50
	else if (sequence_seperation == 5) return 0.75
	else return 1.00

function SWalign(matrix,gap_open,gap_extention)
	return local alignment
```

The algorithm tries different *sep* (sequence separation difference) and gap extension (gap_e) penalties and reports the best alignment. For sep_x we try a constant, linear and quadratic function with different scaling factors (sep_y). The alignment that maximizes the number of contacts (while minimizing the number of gaps) is reported at the end.

4.3.5 *Structure prediction*

Each contact map was examined and trimmed at the N and C terminus to remove regions not constrained by strong non-local contacts. If the contact map was larger than 300 residues, it was split into multiple overlapping domains of 300 residues or less. 66 out of 921 protein families modeled were parsed into 2 or more overlapping domains (20 of the 612 converged models are from these parsed domains). 10,000 *de novo* models were generated using the standard Rosetta AbInitio protocol using sigmoid restraints as described before (*21*). An additional 10,000 models were generated also including bounded restraints (*51*) which improve convergence in many cases because large restraint violations are given large penalties. The bounded restraints were only used during the coarse-grain sampling and disabled during the full atom refinement and minimization. See end of this section for flags and parameters used for the AbInitio protocol.

In addition to the *de novo* structure prediction calculations which start from an extended chain, models were generated by recombining portions of structures identified by the *map_align* contact map matching protocol described above. The search was performed against a subset of PDB(s) with a maximum mutual sequence identity of 30% (*52*). The predicted contacts were aligned against PDB contact maps (see section above). The PDB contact maps were defined using a 5 Å distance cutoff between any pair of heavy atoms for residues with sequence separation of 3 or higher. The top 20 hits were input into the Rosetta hybrid protocol (*32*): the input models are first superimposed and then split into secondary structure elements which are recombined. In addition to recombination, fragment insertion is allowed at all positions, allowing sampling of structures not seen in any of the input templates. 4,000 models are produced at this stage. See end of this section for flags and parameters used for the Rosetta hybrid protocol.

In all structure prediction calculations, structures were ranked based on their Rosetta energies and their fit to the contact restraints. A simple linear combination of the two metrics was used with a scale factor found previously (*21*) to give the two roughly equivalent dynamic range.

30 cluster centers of the whole structural pool, containing 500 top-scoring *de novo* structure prediction models and 100 top-scoring *map_align* models, are selected as initial structures for further refinement using iterative hybridization. For clustering, starting from the lowest energy conformation the next lowest energy conformation is added to the pool if it is not

close to any of pool structures added (TMscore < 0.4), and this is stopped when the pool size becomes 30. If the pool size is smaller than 30 after looking at all conformations, we repeat this step again starting from the lowest unadded conformation with more generous TMscore cut until the pool size reaches to 30.

At each iteration of refinement, Rosetta hybridization protocol is applied 60 times on different combinations of 5 randomly selected models from the structural pool. Once 60 new models are generated, structural pool is updated for the next iteration; new structures are scanned in ascending order of their combined score of Rosetta energy and coevolution restraint scores, and the structure "A" in the original pool is replaced by this new structure "B" if i) it has higher (= unfavorable) energy *and* ii) "A" and "B" are structurally similar or "B" is structurally different from any of the original structures (*53*). Structural similarity criteria linearly changes from 0.4 to 0.7 TMscore from first to 18th iteration, ands keeps unchanged until the end. Therefore size of structural pool (30 structures) is maintained throughout the refinement stage. This is repeated for 30 iterations generating 1,800 structures. For final model selection, structural averaging (*54*) is performed on this entire structural pool, followed by model relaxation to idealize local geometry of the model.

Total computational time for structure calculation of 200-residue protein takes approximately 13,000 core hours: generating 20,000 *de novo* models, 4,000 map_align models, and running structural refinement take 10,000, 2,000, and 1,000 core hours, respectively. This process can be highly parallelized; each of *de novo* and *map_align* models are modeled in parallel using Rosetta@home, and refinement is carried out using 64 cores in parallel.

AbInitio protocol parameters and flags:

```
AbinitioRelax # name of rosetta app
-abinitio::increase_cycles 10
-abinitio::fastrelax
-abinitio::rg_reweight 0.5
-abinitio::rsd_wt_helix 0.5
-abinitio::rsd_wt_loop 0.5
-constraints:cst_weight 3
-constraints:cst_file SIG_BND_cst # sigmoid + bounded constraints
-constraints:cst_fa_weight 3
-constraints:cst_fa_file SIG_cst # only sigmoid constraints for full atom refinement mode
-in::file::fasta t000_.fasta
-frag3 t000_.200.3mers.gz
-fragA t000_.200.9mers.gz
-fragB t000_.200.3mers.gz
-nstruct 10000
```

Rosetta hybrid protocol parameters and flags:

```
rosetta_scripts # name of rosetta app
-frag_weight_aligned 0.1
-beta # this flag enables the latest rosetta score function
-in:file:fasta t000_.fasta
-parser:protocol hyb.xml # rosetta script (see below)
-relax:minimize_bond_angles
-relax:jump_move true
-relax::dualspace
-default_max_cycles 200
-relax:min_type lbfgs_armijo_nonmonotone
-hybridize:stage1_probability 1.0
-hybridize:stage1_4_cycles 400
-nstruct 4000

<ROSETTASCRIPTS>
      <TASKOPERATIONS></TASKOPERATIONS>
      <SCOREFXNS>
            <ScoreFunction name="stage1" weights="stage1.wts" symmetric="0">
                  <Reweight scoretype="atom_pair_constraint" weight="3"/>
            </ScoreFunction>
            <ScoreFunction name="stage2" weights="stage2.wts" symmetric="0">
                  <Reweight scoretype="atom_pair_constraint" weight="3"/>
            </ScoreFunction>
            <ScoreFunction name="fullatom" weights="beta_cart.wts" symmetric=0>
                  <Reweight scoretype="atom_pair_constraint" weight="3"/>
            </ScoreFunction>
      </SCOREFXNS>
      <FILTERS></FILTERS>
      <MOVERS>
            <Hybridize name="hybridize" stage1_scorefxn="stage1"
            stage2_scorefxn="stage2" fa_cst_file="SIG_cst" fa_scorefxn="fullatom"
            batch="1" stage1_increase_cycles="2" stage2_increase_cycles="1"
            linmin_only="0" skip_long_min="1">
                  <Fragments three_mers="t000_.200.3mers.gz"
                  nine_mers="t000_.200.9mers.gz"/>
                  <Template pdb="X.pdb" weight="X" cst_file="SIG_BND_cst"/>
                  <Template pdb="Y.pdb" weight="Y" cst_file="SIG_BND_cst"/>
                  <Template pdb="Z.pdb" weight="Z" cst_file="SIG_BND_cst"/>
            </Hybridize>
      </MOVERS>
      <APPLY_TO_POSE></APPLY_TO_POSE>
      <PROTOCOLS>
            <Add mover="hybridize"/>
      </PROTOCOLS>
      <OUTPUT scorefxn="fullatom"/>
</ROSETTASCRIPTS>
```

4.3.6 *Benchmark with UniProt only for sweeping over Nf values*

The dataset was chosen using PDB20 database from PISCES (resolution limited to 3.0 or better, length >= 50, date 03Mar2015). Multiple sequence alignments were generated for each using HHblits (version 2.0.15; -n 8 -e 1E-20 -maxfilt ∞ -neffmax 20 -nodiff -realign_max ∞), and HHfilter (-id 90 -cov 75) using a clustered UniProt database from 2015_06. No metagenomic sequences were used for this benchmark. Alignments with at least 64 Nf and no homologs in the PDB prior to 2011 were selected. Homology detection was carried out using *blastpgp* (version 2.2.26; -t 1 -e 0.05) (*55*) against a database of PDB sequences from 2011, using a checkpoint file generated by *csblast* (*56*) from the multiple sequence alignment. The benchmark was further filter by removing targets that had missing internal density, reducing the size from 37 to 25 targets. In addition to these proteins, two targets from CASP11 were included: T0806 and T0824. These were the only targets with enough sequences during the CASP11 experiment, and were highlights of the BAKER group human efforts in CASP11; we wanted to test how robust our protocol was in automating their accurate prediction. See Table S1 for a list of these targets and ranges modeled. To avoid bias in the modeling, structures deposited in PDB before 2011 were used for both fragment picking and template selection using *map_align*.

For the subsampled MSAs (multiple sequences alignments), sequences were shuffled and added one at a time until the desired Nf was achieved. Since we are subsampling from a large MSAs, these are likely to result in MSAs that are more diverse on average than a natural MSA of the same Nf. To compensate for higher diversity and still have the same Nf, the subsampled MSAs will need to contain lower number of sequences than natural MSAs. To test if these compositional differences of the natural and subsampled MSAs affect accuracy, we binned our PDB30 set from the "Nf (effective number of sequences) calculation" section (see above) into different Nf categories and compared the distribution of accuracies of the predicted contacts (Figure S6). We find the accuracies to be on average very similar (Figure S6A), even though there does indeed exists a compositional differences (Figure S6B-C).

4.3.7 *Additional benchmark with metagenomic sequences for testing the modeling protocol*

For the additional benchmark set, a more stringent threshold for selecting protein families with no homology in the PDB, no restriction to source and quality of experimental density, and

combination of UniRef100 and metagenomic sequences were used. Each PFAM hmm (provided by HHsuite) was used to search against two different PDB HMM databases; one from 01Jan2012 and one from 13Aug16. HHsearch (default options with -glob flag) was used for the HMM-HMM alignment. For 496 PFAMs with no hits in 2012 (E-value of the top PDB hit > 1) and a strong hit in 2016 (E-value of the top hit < 1E-10), *hmmsearch* (default options with -T 27 flag) was used to collect sequences from the master database containing both the UniRef100 and metagenomic PFAM sequences. For each of 126 PFAMs with more than 64 Nf, the untrimmed UniProt sequence corresponding to the top PDB hit from 13Aug16 was used as the representative sequence. Of the 126, only 70 pfams had that at least 64 Nf for the full length UniProt sequence and the remainder were discarded (there was no trimming to PFAM boundaries to avoid incorporating any native structural bias in PFAM). For three cases where the UniProt sequences were over 400 residues in length (Q96MU8, A9JTH8 and P24043), the "Family & Domains" information from UniProt was used to trim the sequences to domain boundaries. To confirm that these boundary definitions were not influenced by structure, we used hmmscan (from the hmmer package) to scan against the 2012 hmm boundaries (Pfam 26, before structures for these families was released) and the 2016 hmm boundaries (Pfam 30). The boundaries were identical. Contact prediction and model generation were carried out exactly as described above for the families with unknown structures, unbiased by knowledge of the correct structure. In all, 86 domains were modeled (11 of the targets were split based on the contact map into multiple overlapping domains). Five of the domains were excluded from analysis, as they contained no density in the target PDB. For a list of the remaining targets see Table S5.

4.3.8 *Convergence Criteria*

As shown in Figure S2, there is a strong correlation between model accuracy and the extent of convergence of the structure prediction calculations. Based on the results shown in this figure, we used as a measure of convergence *max(con_DN, con_MP, iDN_iMP)*, in which DN and MP refer to *de novo* and *map_align*, respectively, and *iDN_iMP* is the consistency between independent runs of the iterative hybrid protocol on the DN and MP models. Criteria of 0.65 is used for the selection of 614 among 1024 families. The first two measures are the average pairwise TMscore (*23*) between the top-10 models produced using each method (Figure S2 A and B) while the third measure is the TMscore between the top scoring model of each method

(Figure S2 C). This convergence criteria is based on the result from Figure S2 that final models -- derived by iterative hybridization of the DN and MP models -- are likely to be accurate when a) either the DN or MP method converges on its own or b) the models produced by the two independent methods are similar to each other.

4.3.9 *SCOPe classification and new fold detection*

A filtered subset of SCOPe domains (version 2.06) (*40*) was downloaded from Astral with sequence identity of 40%. TMalign was used to superimpose each model to each SCOPe domain and compute fold similarity. If the the best hit (best TMalign score scaled by length of model) is greater than 0.5 TMalign score, it is used to classify the model, otherwise the model is labelled as new fold.

For the 137 models with no hits in SCOPe, an all vs. all analysis was performed. The TMscore >= 0.5 scaled by the average of two length was used for clustering. 129 connected components were found. This includes one cluster with 3 members and six clusters with 2 members.

4.3.10 *Evaluation of recently solved structures*

The solved structures are often close homologs of the predicted structures (rather than the identical protein), making the TMscore calculation (that requires identical sequence) difficult. For evaluation, we instead used TMalign, which is a sequence independent structure comparison method. Though, before running TMalign, HHsearch (from HHsuite package; version 2.0.15) is used to identify regions that are common to both predicted model and PDB homolog. Regions that are not aligned are removed from both model and PDB homolog. These trimmed structures are used in Figure 1 and Figure S3, and used in TMalign score calculation.

4.4 SUPPLEMENTARY TEXT

4.4.1 *Limitation of modeling*

While all of our benchmark tests and comparisons of recently determined crystal structures to previously generated models suggest that the structures presented in this paper are likely to be

quite accurate over their entire length (TMscore > 0.7), there are several systematic limitations users of these models should be aware of. For membrane proteins the lipid bilayer is not modeled explicitly, and intra and extracellular domains can dip into the membrane region (for example Figure S8A). Ligands and co-factors are also not modeled explicitly and models can collapse to obscure binding sites (Figure S8A-C; if the co-factors are known they can be easily incorporated during modeling). Finally, short segments can have the incorrect local secondary structure in the context of an overall correct topology (Figure 1E). A potential additional source of error is confounding intra- and inter-domain contacts in homo-oligomers, but our benchmark set calculations suggest that the convergence based selection criterion eliminates almost all such cases except for intertwined homo-oligomers where it is possible to make all the predicted contacts within the monomer (an example of such case is 5AN6 (Figure S2F)).

4.4.2 *Models for groups of functionally related proteins*

As noted in the main text, it is a challenge to fully present the large numbers of new structures described in this paper. A number of the structures fall into groups with related functions or classifications. These include four cobalamin biosynthesis (Figure S9), two citrate lyase (Figure S10), ten sporulation (Figure S11), and eight immunity proteins (Figure S12).

Since we prepared our Report, additional work has been published in which coevolution contacts were used to make models (*58-63*).

4.5 ACKNOWLEDGMENTS

We would like to thank Pietro Di Lena, Noel Malod-Dognin and Rumen Andonov for providing the source code for their software (Al-eigen and a_purva) and for their discussion and advice on contact map alignment. 3-D structures of 614 PFAMs modeled in the study is available at https://gremlin2.bakerlab.org/meta.php. We also thank Rosetta@home and Charity engine participants for donating their computer time. The work performed by NV, GAP and NCK, was supported by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, under Contract No. DE-AC02-05CH11231. Research reported in this publication was supported by NIGMS of the National Institutes of Health under award number

R01GM092802. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

4.6 REFERENCES

1. R. D. Finn *et al.*, The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–85 (2016).
2. J. Söding, Protein homology detection by HMM–HMM comparison. *Bioinformatics*. **21**, 951–960 (2005).
3. 1. G. T. Montelione, The Protein Structure Initiative: achievements and visions for the future. *F1000 Biol. Rep.* **4**, 7 (2012).
4. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* **110**, 15674-15679 (2013).
5. D. S. Marks *et al.*, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
6. F. Morcos *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* **108**, E1293-1301 (2011).
7. T. A. Hopf *et al.*, Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607-1621 (2012).
8. T. Nugent, D. T. Jones, Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proc Natl Acad Sci U S A* **109**, E1540-1547 (2012).
9. D. T. Jones, D. W. Buchan, D. Cozzetto, M. Pontil, PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184-190 (2012).
10. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nat Biotechnol* **30**, 1072-1080 (2012).
11. J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, J. N. Onuchic, Genomics-aided structure prediction. *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, Learning generative models for protein fold families. *Proteins* **79**, 1061-1078 (2011).
13. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E Stat Nonlin Soft Matter Phys* **87**, 012707 (2013).
14. S. Wickles *et al.*, A structural model of the active ribosome-bound membrane protein insertase YidC. *Elife* **3**, e03035 (2014).
15. P. Tian *et al.*, Structure of a functional amyloid protein subunit computed using sequence variation. *J Am Chem Soc* **137**, 22-25 (2015).
16. S. Hayat, C. Sander, A. Elofsson, D. S. Marks, Accurate prediction of transmembrane β-barrel proteins from sequences. *bioRxiv*, 006577 (2014).
17. T. A. Hopf *et al.*, Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun* **6**, 6077 (2015).
18. L. A. Abriata, An homology-and coevolution-consistent structural model of bacterial copper-tolerance protein CopM supports function as a “metal sponge”and suggests regions for …. *bioRxiv* (2015).
19. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
20. T. A. Hopf *et al.*, Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, (2014).
21. S. Ovchnnikov *et al.*, Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife.* **4**, e09248 (2015).
22. S. Antala, S. Ovchinnikov, H. Kamisetty, D. Baker, R. E. Dempski, Computation and Functional Studies Provide a Model for the Structure of the Zinc Transporter hZIP4. *J. Biol. Chem.* **290**, 17796–17805 (2015).
23. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710 (2004).
24. L. Vogeley *et al.*, Structural basis of lipoprotein signal peptidase II action and inhibition by the antibiotic globomycin. *Science*. **351**, 876–880 (2016).
25. G. Mao *et al.*, Crystal structure of E. coli lipoprotein diacylglyceryl transferase. *Nat. Commun.* **7**, 10198 (2016).
26. R. B. Stockbridge *et al.*, Crystal structures of a double-barrelled fluoride ion channel. *Nature*. **525**, 548–551 (2015).

27. S. Safarian *et al.*, Structure of a bd oxidase indicates similar mechanisms for membrane-integrated oxygen reductases. *Science*. **352**, 583–586 (2016).
28. H. Tsuchiya *et al.*, Structural basis for amino acid export by DMT superfamily transporter YddG. *Nature*. **534**, 417–420 (2016).
29. P. R. Feliciano, C. L. Drennan, M. C. Nonato, Crystal structure of an Fe-S cluster-containing fumarate hydratase enzyme from Leishmania major reveals a unique protein fold. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9804–9809 (2016).
30. F. DiMaio, et al. Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat Methods* **10** (11):1102–1104 (2013).
31. S. Ovchinnikov, D. E. Kim, R. Wang, Improved de novo structure prediction in CASP11 by incorporating Co-evolution information into rosetta. *Proteins: Struct. Funct. Bioinf.* (2015).
32. Y. Song *et al.*, High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735-1742 (2013).
33. W. R. Taylor, Protein structure comparison using iterated double dynamic programming. *Protein Sci.* **8**, 654–665 (1999).
34. K. T. Simons *et al.*, Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82-95 (1999).
35. B. E. Suzek *et al.*, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*. **31**, 926–932 (2015).
36. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. "A bioinformatician's guide to metagenomics." Microbiol Mol Biol Rev. (2008) 72(4):557-78
37. V. M. Markowitz *et al.*, IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* **42**, D568–73 (2014).
38. N. Rego, D. Koes, 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*. **31**, 1322–1324 (2015).
39. M. Bostock, V. Ogievetsky, J. Heer, D #x0B3; Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **17**, 2301–2309 (2011).
40. A. Andreeva *et al.*, Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* **36**, D419–25 (2008).
41. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**, 205-211 (2009).
42. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175 (2012).
43. S. Seemayer, M. Gruber, J. Söding, CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*. **30**, 3128–3130 (2014).
44. T. F. Smith, M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
45. N. Malod-Dognin,, Nicola Yanev, and Rumen Andonov. "Comparing protein 3d structures using a_purva." PhD diss., INRIA, 2010.
46. N. Malod-Dognin, N. Pržulj, GR-Align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. *Bioinformatics*. 30, 1259–1265 (2014).
47. P. Di Lena, P. Fariselli, L. Margara, M. Vassura, R. Casadio, Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics*. 26, 2250–2258 (2010).
48. D. A. Pelta, J. R. González, M. Moreno Vega, A simple and fast heuristic for protein structure comparison. *BMC Bioinformatics*. 9, 161 (2008).
49. S. B. Needleman, C. D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
50. J. Ma, S. Wang, Z. Wang, J. Xu, MRFalign: protein homology detection through alignment of Markov random fields. *PLoS Comput. Biol.* **10**, e1003500 (2014).
51. D. E. Kim, F. Dimaio, R. Yu-Ruei Wang, Y. Song, D. Baker, One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins*. **82 Suppl 2**, 208–218 (2014).
52. G. Wang and R. L. Dunbrack, Jr. PISCES: a protein sequence culling server. Bioinformatics, 19,1589-1591 (2003).
53. J. Lee, H.A. Scheraga, and S. Rackovsky, New optimization method for conformational energy calculations on polypeptides: Conformational space annealing, *J Comput. Chem.* 18, no. 9, 1222-1232 (1997).
54. H. Park, F. DiMaio, D. Baker, The origin of consistent protein structure refinement from structural averaging, *Structure* **23**, 1123-1128 (2015).
55. A. A. Schäffer *et al.*, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994–3005 (2001).

56. Angermüller, A. Biegert, J. Söding, Discriminative modelling of context-specific amino acid substitution probabilities. *Bioinformatics*. **28**, 3240–3247 (2012).
57. Team, R. Core. R: A language and environment for statistical computing. (2013).
58. M. M. Kassem, Y. Wang, W. Boomsma, K. Lindorff-Larsen, Structure of the Bacterial Cytoskeleton Protein Bactofilin by NMR Chemical Shifts and Sequence Variation. *Biophys J* **110**, 2342-2348 (2016).
59. A. Kedrov *et al.*, Structural Dynamics of the YidC:Ribosome Complex during Membrane Protein Biogenesis. *Cell Rep* **17**, 2943-2954 (2016).
60. D. Lloyd Evans, S. V. Joshi, Elucidating modes of activation and herbicide resistance by sequence assembly and molecular modelling of the Acetolactate synthase complex in sugarcane. *J Theor Biol* **407**, 184-197 (2016).
61. D. G. Schep, J. Zhao, J. L. Rubinstein, Models for the a subunits of the Thermus thermophilus V/A-ATPase and Saccharomyces cerevisiae V-ATPase enzymes by cryo-EM and evolutionary covariance. *Proc Natl Acad Sci U S A* **113**, 3245-3250 (2016).
62. M. J. Skwark, M. Michel, D. M. Hurtado, M. Ekeberg, A. Elofsson, Accurate contact predictions for thousands of protein families using PconsC3. *bioRxiv*, 079673 (2016).
63. W. R. Taylor, T. R. Matthews-Palmer, M. Beeby, Molecular Models for the Core Components of the Flagellar Type-III Secretion Complex. *PLoS One***11**, e0164047 (2016).

4.7 SUPPLEMENTAL FIGURES

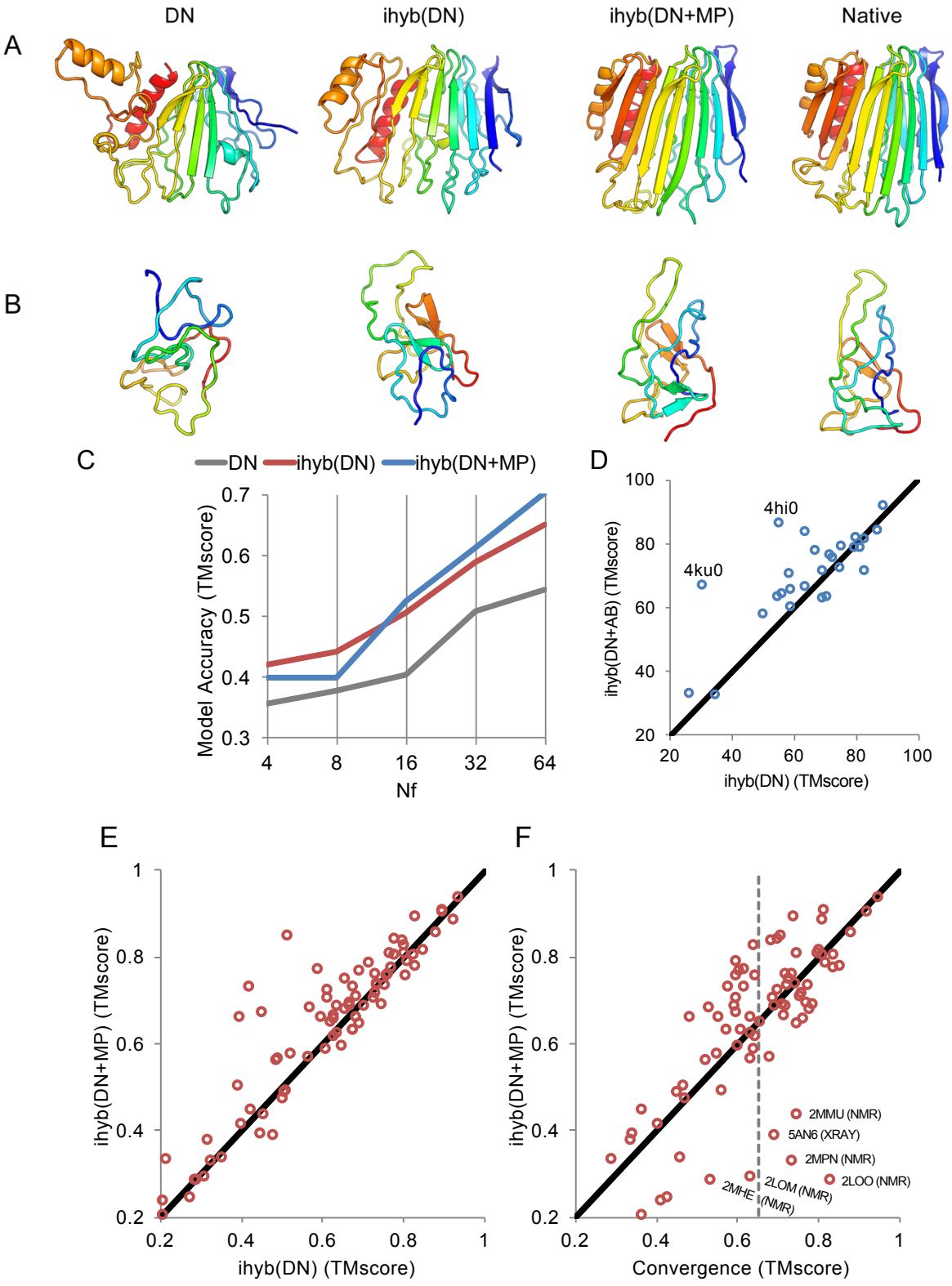


Figure S1. Adding partial threads detected by *map_align* (MP) improves accuracy of the final model more frequently than reducing it for Nf > 16. A) Models for 4hi0 at different stages of the protocol. B) Models for 4ku0 at different stages of the protocol. C) Average model accuracy with and without *map_align* partial threads for different Nf bins over 27 protein benchmark. D) Model accuracy with and without *map_align* partial threads for each protein in the 27 protein benchmark at Nf=64. For 5 targets, the TMscore increases by more than 10 with partial threads, and for only 1 target does it decrease by more than 10 with partial threads. Abbreviations in the figure: DN, *de novo*; ihyb(DN), iterative hybridization of *de novo* models; ihyb(DN+MP), iterative hybridization of *de novo* models with *map_align* partial threads. E) Same comparison as in D for 81 protein benchmark with Nf > 64. For 9 targets, the TMscore increases by more than 10 with partial threads, and for no targets, does it decrease by more than 10 with partial threads. To test the significance of these improvements (C-E) we performed Wilcoxon Signed-Rank Test using R (57), for more details see SI Table S6. F) Correlation between convergence criteria and model quality. With the exception of small transmembrane proteins (~100 length) solved by NMR and an 5AN6 (intertwined dimer), the convergence cutoff of 0.65 (grey dotted line) is a good predictor of accuracy (average TMscore ~ 0.7). For details of the outliers see Supplementary Figure S13.

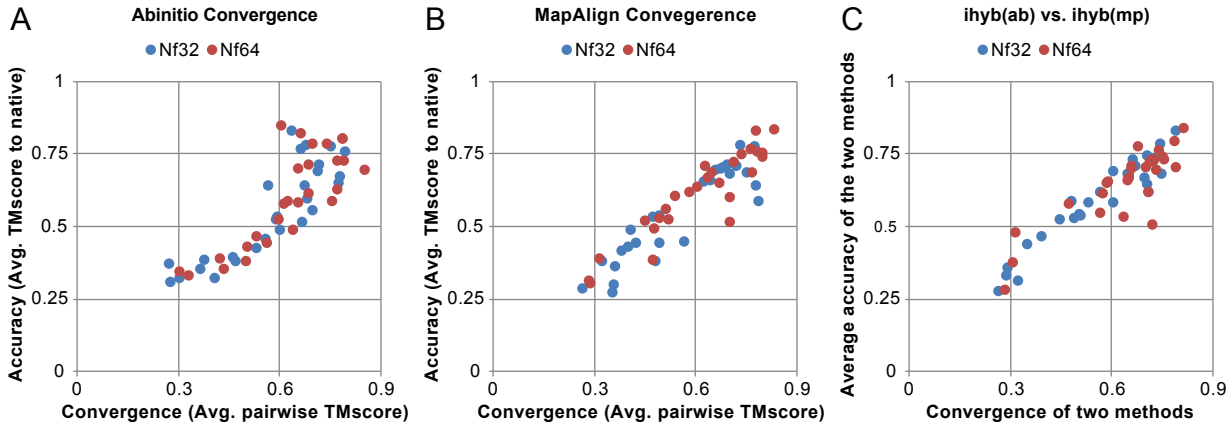


Figure S2. Correlation between convergence (as measured by average pairwise TMscore) and model accuracy over benchmark set. The extent of convergence correlates with model accuracy for (A) the top ranked (using the linear combination of Rosetta energy and contact score described in the Methods) ten out of 10,000 *de novo* models and (B) the top ranked 10 of 4,000 *map_align* models after a single round of the Rosetta hybridization protocol. C) Following multiple rounds of iterative hybrid refinement of the *de novo* models and the map-align models independently, the similarity between the top model from each protocol also correlates with model accuracy. The convergence criterion we use in this paper is that the maximum of the three metrics is greater than 0.65; this corresponds to an average TMscore to native of 0.7 over the benchmark set.

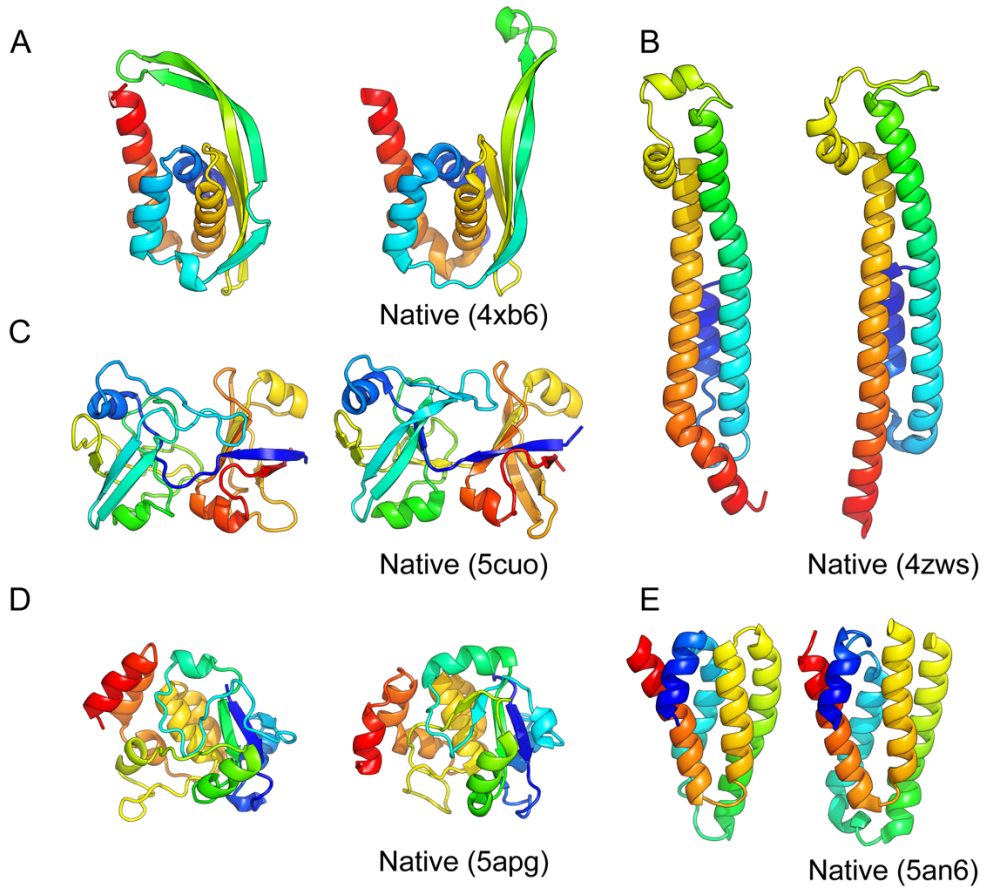


Figure S3. Structures solved while the manuscript was in preparation. For TMalign score and convergence criterion values see Table S4.

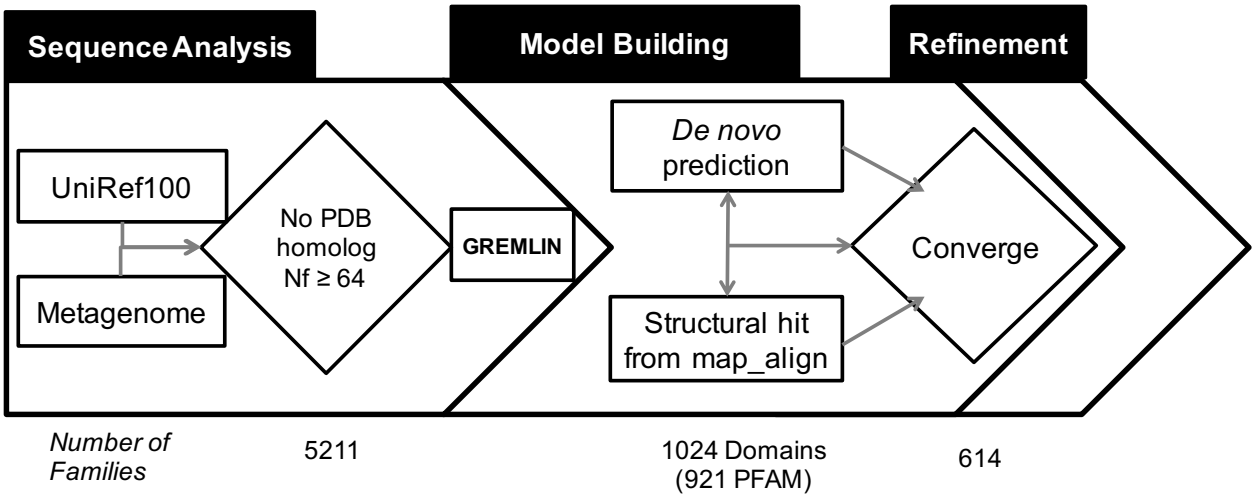


Figure S4. Flowchart of method. Details of the steps are explained in Supplementary text.

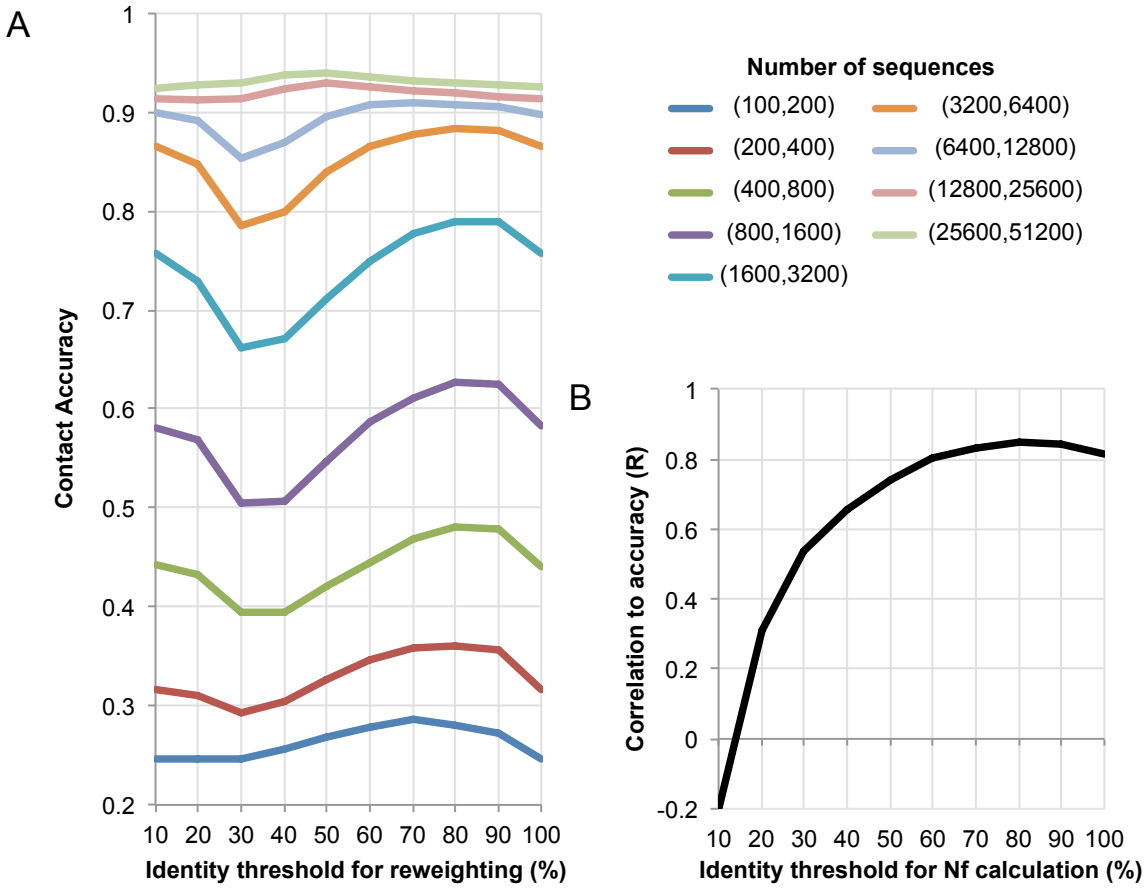


Figure S5. Accounting for sequence redundancy and diversity in the MSAs (multiple sequence alignments) results in both higher accuracy and better prediction of accuracy. For the PDB30 set (see methods), each sequence in MSA is given a weight of 1/(number sequences > X% identity), where X is search parameter to be determined. These weights are used in GREMLIN to down weight redundant sequences. A) For a broad range of protein families sizes (different lines) the best accuracy is achieved at X=80% sequence identity cutoff. The weights computed with identity threshold of 10% or 100% result in uniform weights for each sequence and hence result in identical accuracies. The accuracy is computed using the top 0.5L (L is the length of the sequence) contacts with sequence separation greater than or equal to 6. A contact is considered made when the smallest distance between any two heavy atoms is less than 8 Å. B) The sum of weights divided by square-root of length is used to calculate the Nf value. Here we try different identity thresholds for the Nf calculations and compute the correlation to accuracy computed with the default identity threshold (80%).

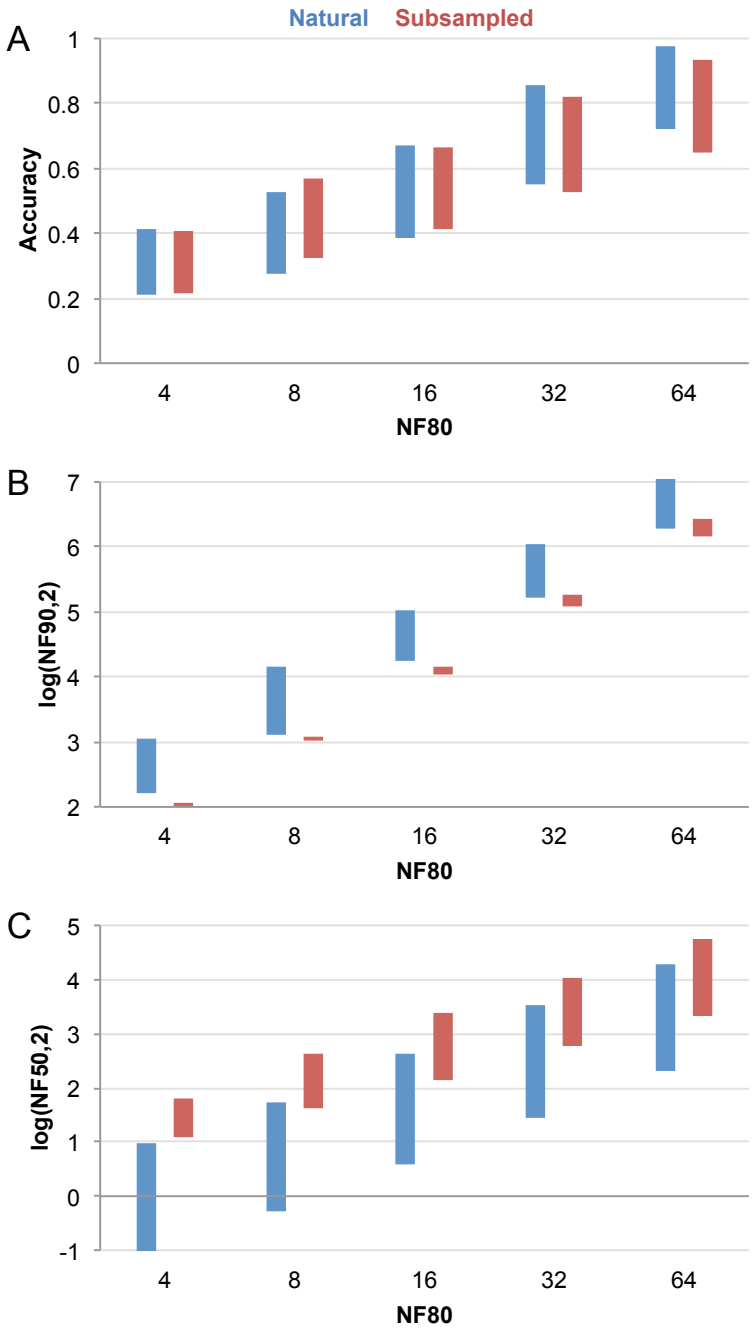


Figure S6. The distribution of accuracies and diversities of natural and subsampled MSA (multiple sequence alignments). The natural set comes from the PDB30 set (see Supplementary text) where the MSAs are binned based on the Nf calculation at 80% sequence identity threshold. The subsampled set comes from the 27 benchmark set. A) The contact accuracy of both the natural and subsampled MSAs is the same. B) There are typically less sequences in the subsampled MSAs, but this is compensated by C) higher diversity. The top and bottom of each bars indicates +/- standard deviation from the mean.

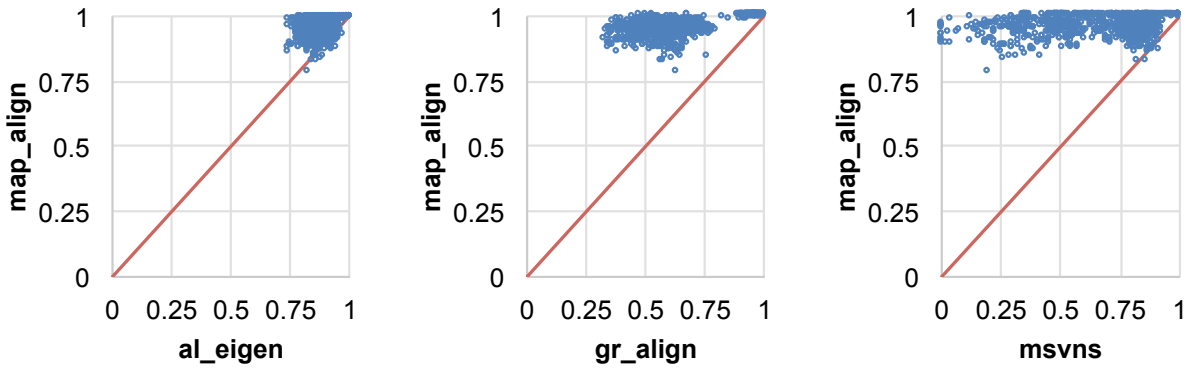


Figure S7. Comparison of *map_align* to previously published methods for detecting structures satisfying a given set of residue-residue contacts. For the Skolnick dataset of 780 SCOP protein pairs brought from Malod-Dognin *et al* (45), *map_align* recovers a larger fraction of contacts than the other methods. The x and y-axes are the fraction of max contacts recovered, as computed by Malod-Dognin *et al* by the exact method a_purva.

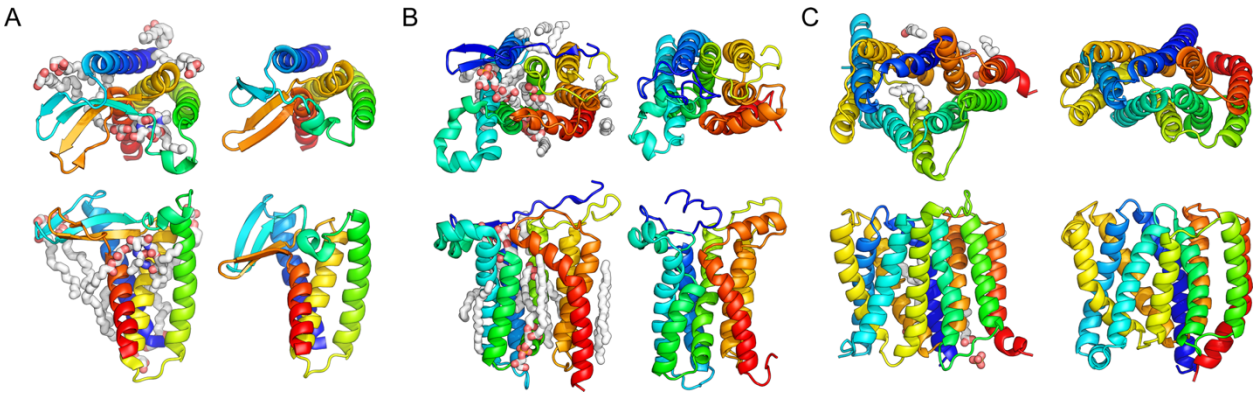


Figure S8. Inaccuracies in models due to missing ligands. A) Lipoprotein signal peptidase II; B) Prolipoprotein diacylglyceryl transferase; C) the DMT superfamily transporter YddG. Left, crystal structures; right models. Top view on top row, side view on bottom row. In all three cases, the models overlap with ligands in the crystal structures shown in spheres.

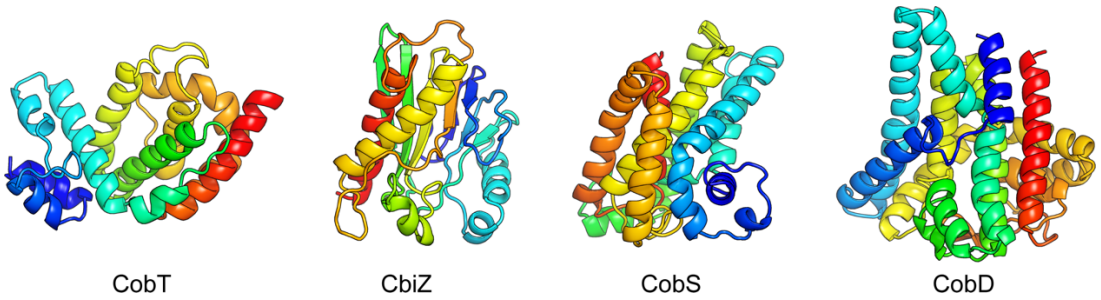


Figure S9. Models for four proteins in the Cobalamin biosynthesis pathway.

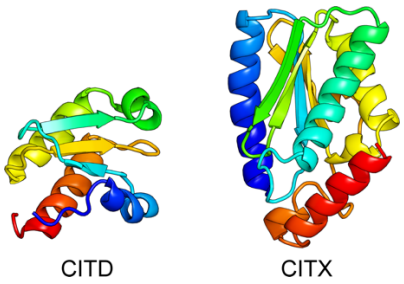
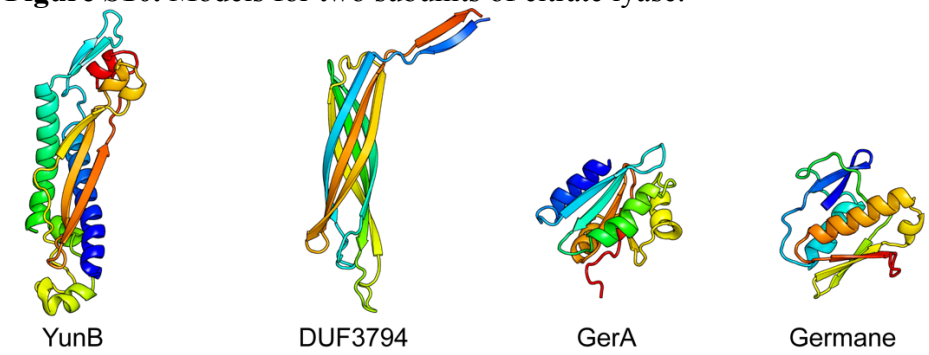


Figure S10. Models for two subunits of citrate lyase.



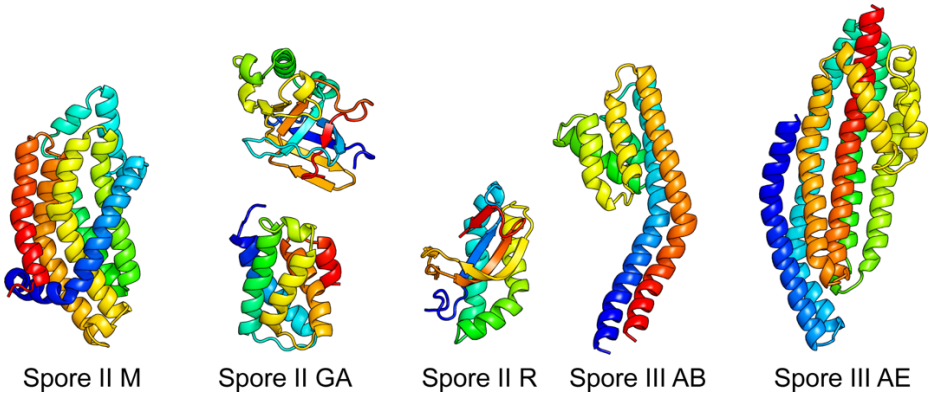
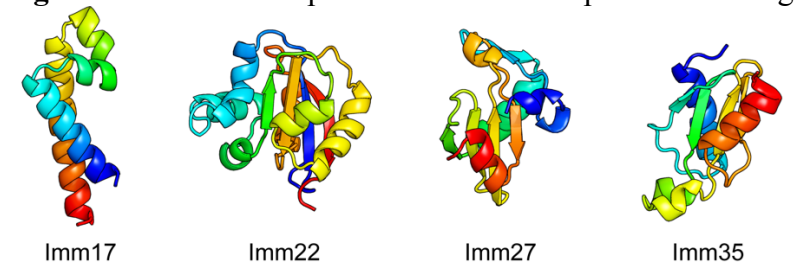


Figure S11. Models of proteins involved in sporulation and germination



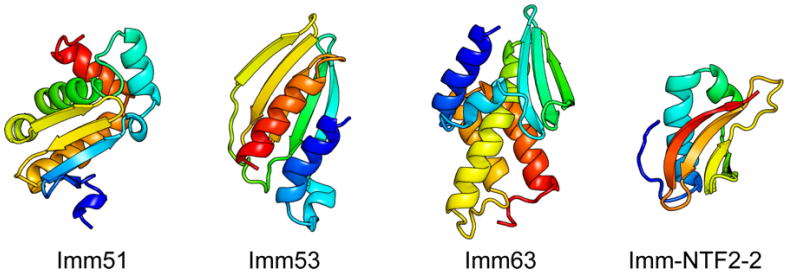


Figure S12. Models of Bacterial immunity proteins.

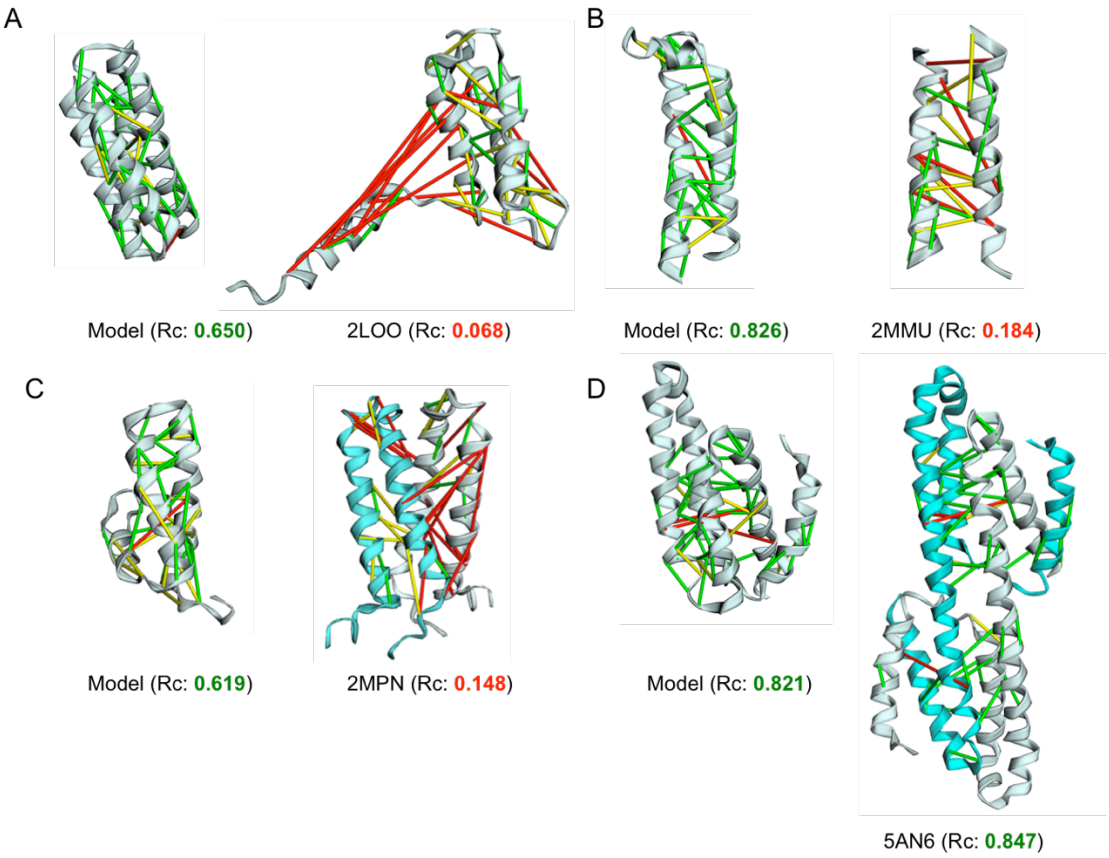


Figure S13. Cases from the additional benchmark set that converged but did not agree with the experimentally determined structure. The top 3L/4 contacts are shown as lines. Green (less than 5 Å), Yellow (between 5-10 Å) and red (greater than 10 Å). Rc is a ratio of contacts made divided by the expected number of contacts (See ref *21*). 3 of the top 4 outliers (A-C) are all small transmembrane proteins; these include A) 2LOO: human membrane protein TMEM14A, B) 2MMU: CrgA, a Cell Division Structural and Regulatory Protein, C) 2MPN: inner membrane protein YgaP. D) 5AN6: Csm2 is an soluble x-ray structure that is also an intertwined dimer. The same set of contacts that are formed at the dimer interface can also be made at the monomer level.

4.8 Supplementary Tables

Table S1. Benchmark set used to evaluate the performance of the method at different Nf (number of effective sequences). Nf values and the TMscore of the top scoring model are reported for each target. The average across all targets is reported in Fig. 2A (solid line) and Fig. S1C (blue line).

Table S2. Nf values for each year (2009-2015) for each of the protein families with currently no detectable homologs using HHsearch before and after addition of metagenomic sequences. The HHsearch log10(E-value) and probability of PDB hit being the same fold is reported . For modeling we selected protein families from this list that had an E-value ≥ 1 and had at least Nf of 64 in the year of 2015 after the addition of metagenomic sequences.

Table S3. List of domains that converged in the structure prediction calculations and the top TMalign hit against the SCOP domain database.

Table S4. List of targets solved while this manuscript has been in preparation and their agreement to crystal structure. Since the newly determined structures have somewhat different sequences than the family representative we chose to model, a sequence independent measure (TMalign score) was used to evaluate model accuracy. In the table we report the convergence criteria (CON) and fraction of contacts made (RC). The RC value (described (21)) is the ratio of the number of contacts made divided by the number of contacts expected given the number of sequences and gremlin score.

Table S5. Additional benchmark with metagenomic sequences for testing modeling protocol. TMscore of the top scoring models is reported.

Table S6. The p-value for each of the datasets using the Wilcoxon signed-rank test. To compute these we used the R function wilcox.test(A,B,paired=TRUE,alternative = c("less")), where A are the TMscores from ihyb(DN) and B are the TMscores from ihyb(DN+MP). Abbreviations in the table: DN, de novo; ihyb(DN), iterative hybridization of de novo models; ihyb(DN+MP), iterative hybridization of de novo models with map_align partial threads. The improvement is significant at cutoff of 0.05 for the new benchmark set (of 81 proteins) and the old benchmark set (of 27 proteins) when $Nf \geq 32$.

VITA

Sergey Ovchinnikov grew up in Portland, OR. He studied Micro/Molecular Biology and Chemistry at Portland State University, graduating in 2010 with a Bachelor of Science. In 2017, he earned a Doctor of Philosophy in Molecular and Cellular Biology at the University of Washington, under the advisement of Dr. David Baker.