PROTEIN NETWORKS

Protein interaction networks revealed by proteome coevolution

Qian Cong[1,2], Ivan Anishchenko[1,2], Sergey Ovchinnikov[3], David Baker[1,2,4]*

Residue-residue coevolution has been observed across a number of protein-protein interfaces, but the extent of residue coevolution between protein families on the whole-proteome scale has not been systematically studied. We investigate coevolution between 5.4 million pairs of proteins in *Escherichia coli* and between 3.9 millions pairs in *Mycobacterium tuberculosis*. We find strong coevolution for binary complexes involved in metabolism and weaker coevolution for larger complexes playing roles in genetic information processing. We take advantage of this coevolution, in combination with structure modeling, to predict protein-protein interactions (PPIs) with an accuracy that benchmark studies suggest is considerably higher than that of proteome-wide two-hybrid and mass spectrometry screens. We identify hundreds of previously uncharacterized PPIs in *E. coli* and *M. tuberculosis* that both add components to known protein complexes and networks and establish the existence of new ones.

Coevolution-based prediction of interacting residues from aligned protein sequences has enabled considerable progress in predicting the structures of monomeric proteins (*1*–*3*) and complexes of proteins known to interact (*4*–*7*). However, using coevolution to identify previously uncharacterized protein-protein interactions (PPIs) over the whole proteome, which requires fishing out the 0.1% of pairs (*8*) that interact among the vast majority of noninteracting pairs, remains a formidable task (*9*, *10*). Determining the extent of coevolution between residues in two different protein families requires pairing each protein in one family with its partner in the other (*7*, *11*). This pairing is not straightforward if either family includes multiple paralogues with different functions and partners in a species; previous studies have thus been largely limited to proteins encoded on the same operon (*4*–*7*). These difficulties, and the complexity of working with millions of protein pairs, have prevented coevolution-based approaches from being used to systematically identify PPIs on the whole-proteome scale.

To systematically investigate coevolution in the *E. coli* proteome (Fig. 1B), we began by using the "reciprocal best hit" criterion (*12*, *13*) to identify, when possible, putative orthologs for each of the 4262 *E. coli* proteins in each of the 40,607 representative bacterial proteomes. We aligned these orthologs (*14*, *15*) and constructed paired alignments for all $4262 \times (4262 - 1) \div 2 = 9{,}080{,}191$ protein pairs. The alignments for 5,433,039 pairs contain sufficient sequence information ($Nf90 >= 16$, see Fig. 1 legend) to assess coevolution (Fig. 1A).

Coevolution detection methods that eliminate transitivity using global statistical models, which consider all residue pairs simultaneously (*16*–*18*), are too slow for datasets of this size. Instead, we used the residue-residue mutual information $\left[\mathrm{MI};\sum_{aa1,aa2} P(aa1,aa2)\cdot log\left(\frac{P(aa1,\,aa2)}{P(aa1)\cdot P(aa2)}\right)\right]$ as an initial screen (*19*); this is a local statistical model because each residue pair is considered independently. We used the maximum value of the MI over all residue pairs as a metric for protein-protein coevolution (rather than an average or sum over the most strongly coevolving residue pairs) to reduce the impact of lack of independence due to transitivity.

We hypothesized that the most strongly coevolving protein pairs would likely physically interact. We assessed this by constructing a "gold-standard" PPI set from *E. coli* protein complexes (*20*) in the Protein Data Bank (PDB) (table S1) and a negative control set consisting of protein pairs drawn from two different complexes (table S2), with no experimental data (*21*–*23*) indicating interactions between them (we cannot be sure that all such pairs do not interact, but the fraction is likely to be small). Selection of coevolving pairs with high MI increases the frequency of physically interacting pairs compared with negative control pairs (Fig. 1C). Considerably better discrimination (Fig. 1C, green versus red curves) of the positive from the negative control could be achieved by down-weighting proteins that appear to coevolve with many others through an average product correction (APC) (*19*).

We selected the top 961,929 pairs (to the left of the black vertical line in Fig. 1C) for further analysis using the global methods direct coupling analysis (DCA) (*6*) and generative regularized models of proteins (GREMLIN) (*16*). These methods improved discrimination of the gold-standard set with DCA followed by GREMLIN performing better than DCA alone (Fig. 1D, purple versus red curves); still better discrimination was achieved by again penalizing residues and proteins that coevolve with many others (Fig. 1D, green versus purple curves). Choosing a threshold balancing sensitivity and specificity, we selected the top 21,818 pairs (to the left of the black vertical line in Fig. 1D). As a final screen, three-dimensional models for proteins in each pair were docked together (*24*), guided by distance constraints between coevolving residue pairs, and we selected 804 protein pairs that exhibited the strongest coevolution across the docked interface (Fig. 1E and materials and methods M7.2). With each increasingly stringent screen step, the number of negative control pairs is reduced greatly, whereas the recovery of the gold-standard pairs decreases only slightly (table S3). We investigated whether deep learning methods (*25*) trained on contacts in monomeric proteins could provide better discrimination but found they did not improve performance (fig. S1; contacts are overpredicted as the prior probability of residue-residue interactions between proteins is far lower than within proteins); such methods likely will require direct training on PPIs to be useful for this purpose.

We compared the accuracy of the coevolution-based interaction predictions with those of interactions inferred from high-throughput yeast two hybrid (Y2H) (*8*) and affinity purification-mass spectrometry (APMS) (*26*, *27*) over several benchmark sets. One benchmark is from the Y2H study, one is from one of the two APMS (*26*) studies (the other APMS study did not contain a benchmark), and two additional benchmarks were derived from x-ray and cryo–electron microscopy complexes of *E. coli* proteins in the PDB and from gold-standard complexes in Ecocyc (*20*) (tables S4 to S7). We evaluate the performance of each method on each benchmark using precision (TP/P, where TP is the number of true interactions that are correctly predicted, and P is the number of predicted interactions) and recall (TP/T, where T is the number of true interactions in a benchmark). Our coevolution screen outperforms the experimental methods (Fig. 1F) in both precision and recall except for a worse recall on the Y2H benchmark—this includes more transient interactions that are not well conserved among many species and hence are harder to detect by using coevolution methods. The interacting partners predicted by coevolution, like those in structurally confirmed complexes, have more closely related functions (fig. S2) than those identified in the large-scale experiments. The fast coevolution detection methods used in the early, higher-throughput steps in our protocol may miss interactions that can be recognized by the slower but more sensitive methods in later steps. Therefore, we input protein pairs reported to interact in experimental studies or on the same operon directly into the GREMLIN and docking screens, resulting in 814 additional pairs (1618 in total, table S8) that pass our coevolution and docking thresholds (coevolution+ protocol).

We observed strong coevolution (predicted interacting probability > 0.7) across the interfaces

[1]Department of Biochemistry, University of Washington, Seattle, WA 98105, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA 98105, USA. [3]John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA 02138, USA. [4]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA.
*Corresponding author. Email: dabaker@uw.edu

of 40% of the gold-standard PPI set but little coevolution (interacting probability < 0.2) for 20% of interfaces in this set. To understand why there was such a wide range of coevolution across known protein-protein interfaces, we compared the properties of strongly coevolving protein complexes with those showing little coevolution (table S9). Overall, coevolution across interfaces in binary complexes (two components) was stronger than that in larger complexes (Fig. 2A); in large assemblies with interfaces between several protein pairs, any single interface may be less critical for complex formation, reducing the extent of coevolution. Coevolution was also lower in complexes that contain nucleic acids (Fig. 2B): the protein–nucleic acid interactions may contribute to the stability of the complex along with the PPIs. Less coevolution was also

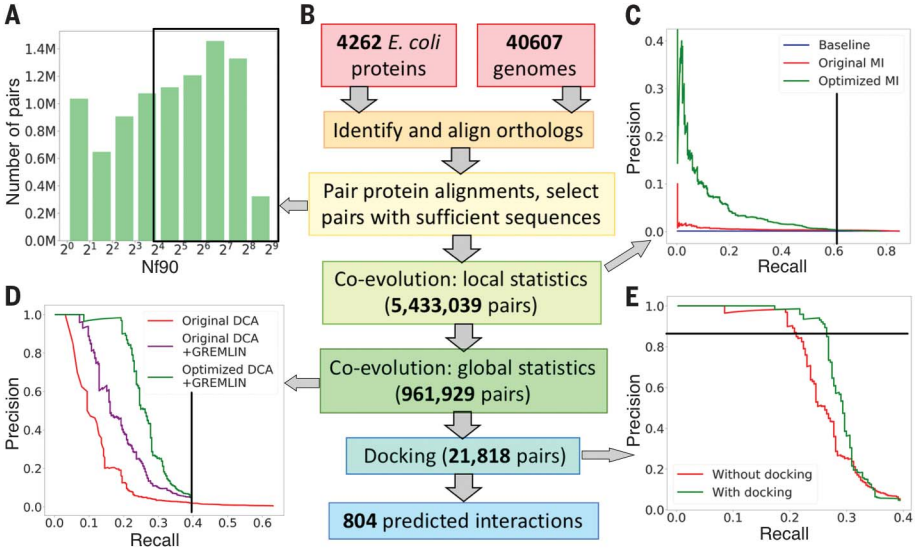

Fig. 1. PPI identification by using coevolution. (**A**) Distribution of *E. coli* protein family sizes. $Nf90 = N90 / \sqrt{L}$, where L is the number of aligned positions in the alignment, and $N90$ is the number of sequences in the alignment, filtered at 90% sequence identity. The black box indicates selected protein pairs. (**B**) Screening pipeline. (**C**) Protein pairs were ranked by MI, and lines sweep MI threshold values from high (left) to low (right). The number (P) of pairs above a MI threshold, the number (T) of gold-standard pairs, and their overlap (TP) are used to calculate Precision (TP/P, *y*-axis) and Recall (TP/T, *x*-axis). Baseline (blue) represents random ranking of pairs. Improved performance (green) is achieved by using an average product correction (APC). (**D**) Enhanced recovery of gold-standard pairs by using global statistical methods (DCA and GREMLIN). Green curve includes APC-like procedures to penalize false-positive hubs. (**E**) Further increase in precision through protein-protein docking calculations. Pairs were ranked by the sum of the probability of contacts made in the best-fitting docked complex. (**F**) Performance of experimental and coevolution screens on diverse benchmarks.

F

	Y2H benchmark (277)				APMS benchmark (1245)				ECOCYC benchmark (916)				PDB benchmark (868)			
	TP	Pre	Rec	F-score	TP	Pre	Rec	F-score	TP	Pre	Rec	F-score	TP	Pre	Rec	F-score
Y2H	59	2.8%	21.3%	4.9%	22	1.0%	1.8%	1.3%	54	2.6%	5.9%	3.6%	45	2.1%	5.2%	3.0%
APMS 1	77	1.3%	27.8%	2.5%	26	0.4%	2.1%	0.7%	96	1.6%	10.5%	2.8%	80	1.3%	9.2%	2.3%
APMS 2	33	0.3%	11.9%	0.5%	243	1.9%	19.5%	3.5%	177	1.4%	19.3%	2.6%	64	0.5%	7.4%	0.9%
Co-evolution	30	3.7%	10.8%	5.6%	175	21.8%	14.1%	17.1%	198	24.6%	21.6%	23.0%	104	12.9%	12.0%	12.4%
Co-evolution+	60	3.7%	21.7%	6.3%	296	18.3%	23.8%	20.7%	301	18.6%	32.9%	23.8%	171	10.6%	19.7%	13.8%

The size of each benchmark is shown in parentheses. Cells are colored by performance: green for the best and red for the worst. Coevolution+, increased coverage by supplementing input to GREMLIN and docking screens with pairs missed in initial stages but identified in previous experimental studies (materials and methods M6.1); F-score, harmonic mean of precision and recall; Pre, precision; Rec, recall; TP, true positives.

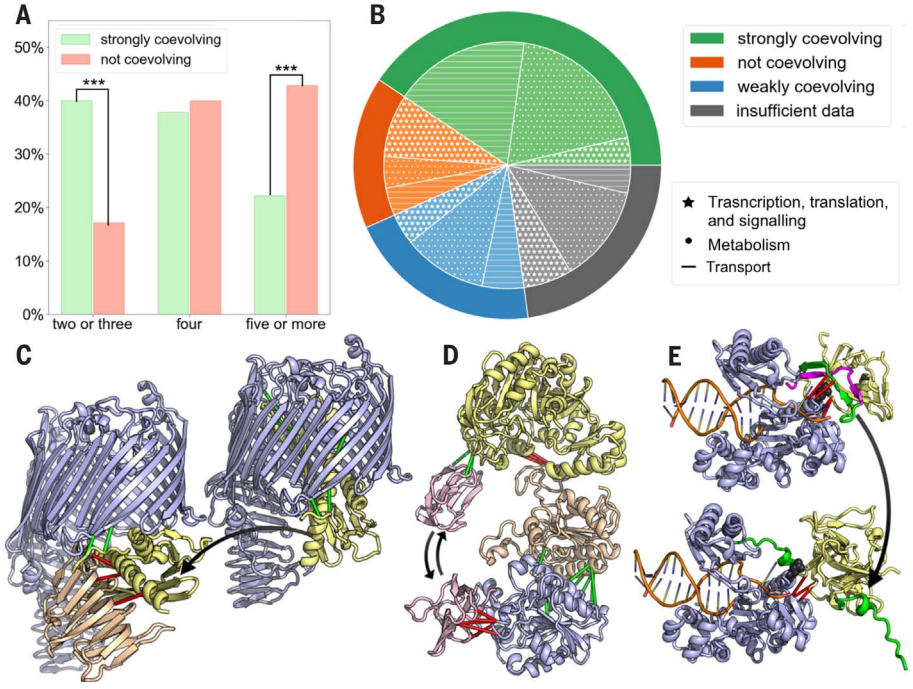

Fig. 2. Coevolution in known protein complexes. The extent of coevolution is higher in complexes with fewer subunits (**A**) and varies with the function of the complexes (**B**). (**C** to **E**) Obligate and transient interactions revealed by coevolution provide insights into function. Bars connecting coevolving residues are in green if an experimental structure containing the interface has been determined and in red if not. Black arrows indicate inferred movements of proteins. (C) LPS transporter consisting of periplasmic LPS-binding protein LptA (orange), LptE (yellow), and outer-membrane β-barrel LptD (light blue). (D) Acetyl-CoA carboxylase complex consisting of biotin carboxylase (AccC, yellow), biotin carboxyl carrier (AccB, pink), and carboxyltransferase subunits (AccA and AccD, light blue and orange). (E) Self-inhibitory mechanism of DNA polymerase V (umuD2C). The magenta β-strand in umuD (yellow, top) is cleaved upon activation by RecA. The remaining umuD′ dimerizes (bottom) and causes the green β strand blocking the active site (black spheres below the magenta strand) in umuC (light blue) to move away and release inhibition of polymerase activity.

observed for small and low-affinity interfaces (fig. S3) and for interfaces exhibiting more variation between different species (fig. S4). On the basis of these observations, we expect that the large set of strongly coevolving protein pairs we have identified includes most of the higher-affinity binary complexes in the core prokaryotic proteome (we are not able to observe coevolution between proteins present in small numbers of species) but likely is quite incomplete in the coverage for larger protein and protein–nucleic acid assemblies.

It is instructive to consider three examples in which the interfaces revealed by coevolution are incompatible with a single static complex, suggesting dynamic interactions important for function. We observed coevolution between the periplasmic lipopolysaccharide (LPS)–binding subunits LptA and LptE and between the N-terminal tail of LptE and the outer-membrane β-barrel LptD in the LPS transporter. LptE sits inside LptD in the cocrystal structure (PDB ID: 4RHB), where it cannot bind LptA (Fig. 2C). The coevolution data suggest a handoff mechanism: LptE dips into the periplasmic space to accept LPS from LptA, maintaining interaction with LptD through the N terminus, and then delivers LPS to the extracellular space by transitioning to the conformation seen in the crystal structure.

In addition to the experimentally determined interfaces between the biotin carboxylase subunit AccC and the biotin carboxyl carrier AccB (PDB ID: 4HR7) and between the carboxyltransferase subunits AccA and AccD (PDB ID: 2F9Y) in the acetyl–coenzyme A (CoA) carboxylase complex, we observed coevolution between AccC and AccD and between AccB and AccA (Fig. 2D). These interactions suggest a dynamic model in which AccB shuttles biotin carboxyl from AccC (which produces it) to AccA-AccD complex, which then transfers the carboxyl group from biotin to the substrate acetyl. For polymerase V (UmuD2C), coevolution data predict a model of the complex between UmuD and UmuC that can explain the self-inhibition mechanism (Fig. 2E). In the predicted UmuD-UmuC complex, the active site of UmuC is obstructed by a segment of UmuD that strongly coevolves with residues around the active site. After cleavage of the 24 N-terminal amino acids of UmuD to generate UmuD′, which then dimerizes (*28*), the active site–blocking segment changes conformation and relieves the inhibition. These examples show that coevolution can suggest transient interactions difficult to capture by using conventional structural biology methods.

As shown in Fig. 3A, *936* (*332* + *604*) of the strongly coevolving pairs have been reported previously. Because the false-positive rate is non-negligible (10 to 20% overall, see materials and methods M8), we searched for additional supporting data for each of the remaining 682 (*1618* – *936*) new predictions. We found homologous PDB templates consistent with coevolution data for 126 pairs, nearby genomic locations for an additional 143 pairs, and an additional 231 pairs with related functions and/or particularly strong coevolution across the docked interface (fig. S5 and table S10). The predicted PPIs include both previously unknown complexes and previously uncharacterized components of known complexes (table S11). The ribosome provides a notable illustration of the latter. We find considerable coevolution between core ribosomal proteins and other components, extending over the full surface of the assembly (Fig. 3B and table S15). Some of these PPIs have been structurally characterized (green bars) or inferred in high-throughput experiments (blue bars), whereas others to our knowledge have not been reported previously (magenta bars).

Our coevolution-guided docking models (Fig. 3, C to T) reveal interfaces that may provide new insights into biological processes: for example, how the type II toxin MqsR and antitoxin MqsA intermesh, enabling antibiotic-resistant cells to dominate in a bacteria community (Fig. 3C);

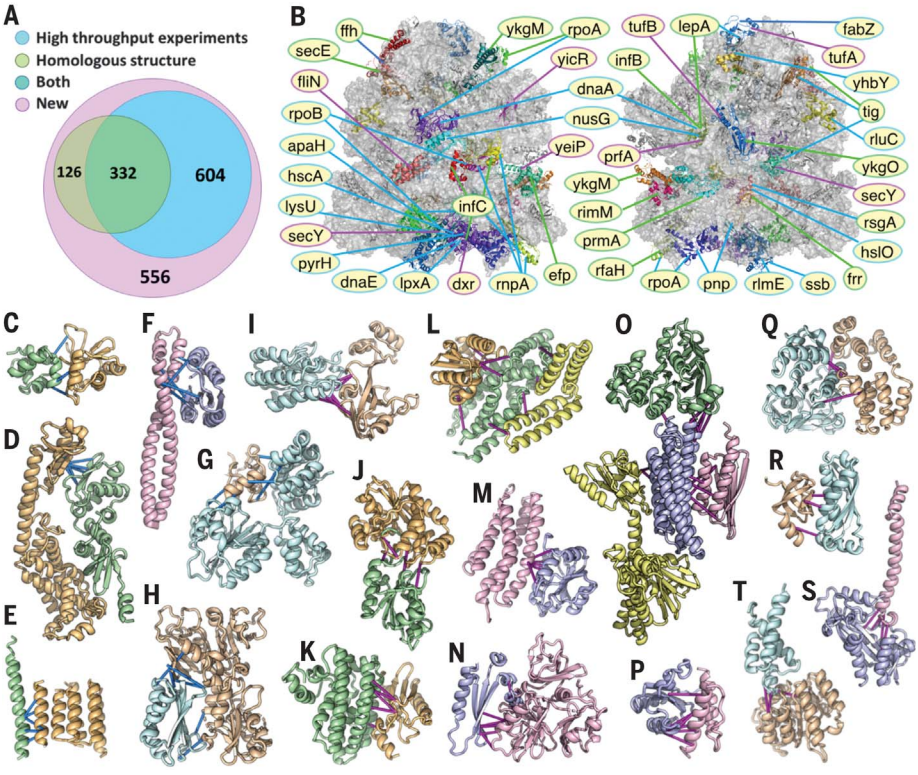

Fig. 3. Examples of new components of known complexes and newly identified complexes. (**A**) Fractions of coevolving complexes that are consistent with previous structural and experimental data. (**B**) Predicted interactions between nonribosomal proteins and core ribosomal proteins are indicated by bars color-coded as in (A) (full names are in table S15). (**C** and **D**) Previously unknown interfaces extending those in crystal structures. (**E** to **H**) Interactions supported by large-scale experiments. (**I** to **T**) Previously unidentified interactions. (C) Coevolution suggests that both the C- (shown) and N-terminal (not shown, in cocrystal) domains of antitoxin MqsA interact with toxin MqsR, possibly forming a higher-order complex. (D) DNA mismatch repair proteins MutS and MutL (C terminus). (E) Sec translocon accessory protein YajC and membrane protein insertase YidC. (F) Cell division protein FtsX and murein hydrolase activator EnvC. (G) DNA polymerase III subunit delta and ferredoxin YfhL. (H) Protein YciI and riboflavin biosynthesis protein RibD. (I) Thioesterase TesA and protein YbbP. (J) tRNA methyltransferase TrmD and tRNA sulfurtransferase ThiI. (K) 1,2-phenylacetyl-CoA epoxidase, subunits C and D. (L) RNA polymerase sigma factor FliA (green), flagellar biosynthetic protein FlhB (orange), and secretion chaperone FliS (yellow). (M) D-ribose pyranase RbsD and sigma D regulator Rsd. (N) Cell division topological specificity factor MinE and tRNA-modifying protein YgfZ. (O) Phosphate transporter ATPase PstB (green), phosphate transporter accessory protein PhoU (blue), phosphate regulon sensor protein PhoR (yellow), and ribosome hibernation promoting factor Hpf (pink). (P) Transcriptional factor BolA and ribosome modulation factor Rmf. (Q) LPS exporter ATPase LptB and protein YbbN. (R) Membrane protein quality-control factor QmcA and protein YbbJ. (S) Nucleoside triphosphatase RdgB and DNA utilization protein HofN. (T) Macrodomain Ter protein MatP and protein YjjV.

how MutS cooperates with MutL to repair DNA mismatches (Fig. 3D); and how the expression, folding, and secretion of flagellar proteins is controlled by the dedicated sigma factor FliA and regulatory proteins such as FlhB (Fig. 3L). Other predicted interactions suggest functional roles for poorly characterized proteins. For example, we predict that uncharacterized proteins YfhL (Fig. 3G), YbbJ (Fig. 3R), and YjjV (Fig. 3T) are involved in regulation of DNA replication, membrane protein quality control, and chromosome segregation based on their coevolution with DNA polymerase III subunit delta, membrane protein quality-control factor QmcA, and macrodomain Ter protein MatP, respectively. Particularly interesting cases involve cross-talk between different pathways (Fig. 3, F and M to P). For example, some metabolic enzymes (Fig. 3M) and transporters (Fig. 3O) coevolve with transcriptional and translational regulators (ribosome hibernation factor and sigma D regulator) involved in the transition from growth to stationary phase; these may link the metabolic status of the cell with the transition to stationary phase.

Beyond the binary interactions considered above, we observe extended networks of mutually coevolving proteins (Fig. 4). Some of these involve proteins with similar biochemical functions—for example, networks of sequentially acting enzymes (fig. S6) and of adenosine triphosphate (ATP)–binding cassette (ABC) transporters (fig. S7). Given the number of components and range of functions, many of the networks are unlikely to form single complexes; rather, they may form multiple reconfiguring complexes involved in more complex functions (fig. S8). Some of the larger networks involve proteins mediating the same biological processes, such as transcriptional regulation (Fig. 4A), outer-membrane integrity maintenance (Fig. 4B), and flagella biosynthesis and assembly (Fig. 4F). Other networks connect different processes and perhaps enable bacteria to modulate DNA replication (Fig. 4D), or transcription and translation (Fig. 4C) on the basis of the internal and external environment—for example, under stress conditions (Fig. 4E).

Fig. 4. Examples of coevolving protein networks. Blue lines connect coevolving protein pairs, and green lines connect proteins interacting in experimentally determined structures. (**A**) Network of transcription elongation factors. (**B**) Outer-membrane integrity maintenance network. (**C**) Linkage between phosphate transport and regulation of transcription and ribosome activity. (**D**) Chaperones and tRNA modification enzymes are coupled to DNA replication initiation, perhaps decreasing it under stress conditions. (**E**) Stress response network. (**F**) Network connecting flagella components and regulators of their synthesis.

We investigated the applicability of coevolution-based interaction prediction to *Mycobacterium tuberculosis*, a human pathogen evolutionarily distant from *E. coli*: only 41% of *M. tuberculosis* proteins have clear *E. coli* homologs (BLAST e-value < 10^{-5}). Using the protocol developed for *E. coli*, we inferred 911 PPIs in *M. tuberculosis* with an expected precision of 83%. Of the predicted PPIs, 662 do not have *E. coli* orthologs, and 593 do not have *E. coli* homologs (BLAST e-value < 10^{-5}). The majority (95%, fig. S9) of these predicted PPIs have not been previously described, because of the limited experimental characterization of *M. tuberculosis* proteins (for *E. coli*, 42% have not been characterized). Forty percent of the predicted interacting partners are functionally related according to the STRING (functional protein networks) database (*23*) (Fig. 5); in comparison, the agreement between pairs identified in the bacterial two-hybrid screen (*29*) and functional networks in STRING is almost as low as that of randomly selected pairs and much lower than Y2H or APMS studies on *E. coli* (fig. S2), reflecting the challenge of carrying out experimental screens on nonmodel organisms (poor protein expression, etc.). In contrast, our coevolution screen is likely as accurate for *M. tuberculosis* as it is for *E. coli*. We provide a full list of the predicted PPIs in table S12. Among these predictions, 293 link poorly characterized proteins to partners with well-annotated function (table S13), and 70 involve proteins that were suggested to contribute to the virulence of *M. tuberculosis* (table S14) (*30*). We expect this list to be useful for therapeutic target identification and deciphering the hard-to-study biology of *M. tuberculosis*.

The large number of previously unpredicted binary interactions, networks, and protein complex structures described here is notable because no new experiments were required; these results instead leverage ongoing genome sequencing efforts. Despite the inevitable errors in such large-scale studies, our benchmarking suggests a considerably higher accuracy than previous large-scale experimental screens and, hence, a sound starting point for detailed experimental testing by biochemistry and mutagenesis. It will

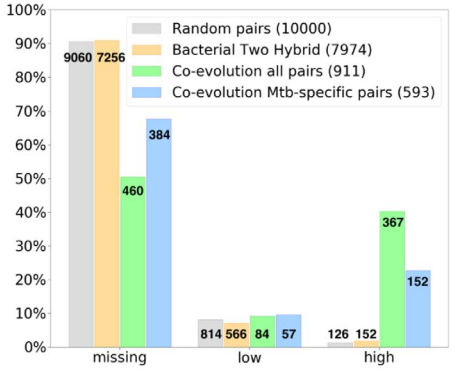

Fig. 5. Functional relatedness of predicted interacting partners in *M. tuberculosis*. Functional relatedness was assessed by using the *M. tuberculosis* functional network in the STRING database (high: STRING combined score ≥ 0.4, missing: not in the STRING database). A considerable fraction (40%) of coevolution-based predictions (green) involve partners that are predicted to be functionally related, whereas a much lower fraction (orange, 2.0%) of PPIs identified in a previous experimental screen involves functionally related partners, almost as low (gray, 1.3%) as randomly selected pairs. The 384 predicted PPIs involving partners lacking homologs in *E. coli* and without STRING annotations (blue bar on left) are likely to be of most interest to *M. tuberculosis* researchers. Mtb, *M. tuberculosis*.

be particularly interesting to follow up on interactions that shed light on the function of previously uncharacterized proteins and to investigate the suggested couplings between different biological processes such as metabolism and translation regulation. The coevolution screen can be carried out on organisms like *M. tuberculosis*, for which experimental PPI screens are difficult or intractable. By carrying out the analysis on many organisms, it should be possible to follow the evolutionary dynamics of PPI networks. The use of coevolution to unravel interaction networks in the core eukaryotic proteome will likely require improved decoupling of coevolutionary and phylogenetic contributions to residue-residue covariation and more genome sequence data on less complex eukaryotes spanning a wide evolutionary distance. Deep learning methods, which have greatly improved the contact prediction in individual proteins by considering the full spectrum of amino acid substitutions (*3*, *11*, *25*) rather than just the overall magnitude of the covariation, have the potential to more accurately predict contacts across interfaces with fewer sequences and further facilitate the identification of PPIs in eukaryotes.

REFERENCES AND NOTES

1. D. S. Marks *et al.*, *PLOS ONE* **6**, e28766 (2011).
2. S. Ovchinnikov *et al.*, *Science* **355**, 294–298 (2017).
3. S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, *PLOS Comput. Biol.* **13**, e1005324 (2017).
4. T. A. Hopf *et al.*, *eLife* **3**, e03430 (2014).
5. S. Ovchinnikov, H. Kamisetty, D. Baker, *eLife* **3**, e02030 (2014).
6. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
7. A. F. Bitbol, R. S. Dwyer, L. J. Colwell, N. S. Wingreen, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 12180–12185 (2016).
8. S. V. Rajagopala *et al.*, *Nat. Biotechnol.* **32**, 285–290 (2014).
9. S. Hashemifar, B. Neyshabur, A. A. Khan, J. Xu, *Bioinformatics* **34**, i802–i810 (2018).
10. A. F. Bitbol, *PLOS Comput. Biol.* **14**, e1006401 (2018).
11. H. Zeng *et al.*, *Nucleic Acids Res.* **46** (W1), W432–W437 (2018).
12. D. P. Wall, H. B. Fraser, A. E. Hirsh, *Bioinformatics* **19**, 1710–1711 (2003).
13. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403–410 (1990).
14. S. R. Eddy, *PLOS Comput. Biol.* **7**, e1002195 (2011).
15. F. Sievers *et al.*, *Mol. Syst. Biol.* **7**, 539 (2011).
16. H. Kamisetty, S. Ovchinnikov, D. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674–15679 (2013).
17. D. S. Marks, T. A. Hopf, C. Sander, *Nat. Biotechnol.* **30**, 1072–1080 (2012).
18. F. Morcos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293–E1301 (2011).
19. S. D. Dunn, L. M. Wahl, G. B. Gloor, *Bioinformatics* **24**, 333–340 (2008).
20. I. M. Keseler *et al.*, *Nucleic Acids Res.* **39** (Database), D583–D590 (2011).
21. S. Orchard *et al.*, *Nucleic Acids Res.* **42** (D1), D358–D363 (2014).
22. I. Xenarios *et al.*, *Nucleic Acids Res.* **28**, 289–291 (2000).
23. D. Szklarczyk *et al.*, *Nucleic Acids Res.* **43** (D1), D447–D452 (2015).
24. I. A. Vakser, *Protein Eng.* **8**, 371–378 (1995).
25. D. T. Jones, S. M. Kandathil, *Bioinformatics* **34**, 3308–3315 (2018).
26. M. Babu *et al.*, *Nat. Biotechnol.* **36**, 103–112 (2018).
27. P. Hu *et al.*, *PLOS Biol.* **7**, e1000096 (2009).
28. Q. Jiang, K. Karata, R. Woodgate, M. M. Cox, M. F. Goodman, *Nature* **460**, 359–363 (2009).
29. Y. Wang *et al.*, *J. Proteome Res.* **9**, 6665–6677 (2010).
30. B. Liu, D. Zheng, Q. Jin, L. Chen, J. Yang, *Nucleic Acids Res.* **47** (D1), D687–D692 (2019).

ACKNOWLEDGMENTS

We thank N. V. Grishin, H. S. Malik, L. Stewart for inspiring discussions about this project, I. Vakser for sharing the latest version of GRAMM software, D. E. Kim for help in setting up the pipeline for protein structure prediction, and L. Goldschmidt for maintaining the computers used in this study. We also thank Rosetta@home and Charity Engine participants for donating their computer time. **Funding:** This project has been funded in part with Washington Research Foundation, National Institute of General Medical Sciences (grant no. R01-GM092802-07), National Institute of Allergy and Infectious Diseases (contract no. HHSN272201700059C), and Office of the Director of the National Institutes of Health (grant no. DP5OD026389). This research used resources of the National Energy Research Scientific Computing Center (contract no. DE-AC02-05CH11231). **Author contributions:** Q.C. and D.B. formulated the research goals and drafted the manuscript. Q.C. designed the methodology, implemented most of the code, collected most of the data, and performed the study. I.A. and S.O. participated in the data collection, methodology design, and code implementation. All authors participated in the final draft. **Competing interests:** The authors declare no competing interests. **Data and material availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/365/6449/185/suppl/DC1
Materials and Methods
Figs. S1 to S24
Tables S1 to S16
References (*31*–*52*)

15 January 2019; accepted 7 June 2019
10.1126/science.aaw6718