COMPUTATIONAL BIOLOGY

Mapping global protein contacts

Genome-scale coevolution experiments expand bacterial protein interactomes

By **Sandor Vajda** *and* **Andrew Emili**

The divergence of orthologous protein sequences across evolutionary lineages can be used to pinpoint possible contacts at specific amino acids by exploiting the tendency for compensating mutations to coevolve at interacting positions within proteins (*1*). This coevolutionary approach has galvanized the vast improvement in protein-structure prediction over the past two decades (*2*). It has also been used to locate contact points between pairs of interacting proteins (*3*), which can serve as distance restraints for high-quality models of multiprotein complexes by structural docking (*4*). On page xx of this issue, Cong *et al.* (*5*) use this method to explore potential interactions among all *Escherichia coli* and *Mycobacterium tuberculosis* proteins and thus enhance knowledge of bacterial protein interaction networks (interactomes).

Because macromolecular complexes drive most biological processes, elucidating the underlying networks of physically interacting proteins is key to understanding the molecular machinery of a cell. Protein-protein interactions have been investigated traditionally by labor-intensive experimental methods, applied separately to a small set of potentially interacting protein targets. Starting with microbes, scientists began to construct larger networks by performing large-scale analyses based on yeast two-hybrid (Y2H) screens (*6*) or affinity purification of protein complexes coupled to mass spectrometry identification (APMS) (*7*). However, such high-throughput approaches can miss certain interactions or yield spurious ones, and are limited to organisms in which molecular techniques such as gene manipulation are possible (*8*). To improve the coverage and reliability of microbial interaction networks, researchers have tried to incorporate knowledge of functional

Departments of Biomedical Engineering, Biochemistry, and Biology, Boston University, Boston, MA 02215, USA. Email: aemili@bu.edu

GRAPHIC: A. KITTERMAN/*SCIENCE*

relatedness, such as the closeness of bacterial genes (membership in operons) or similarities in phylogenetic profiles, but such integrative scoring approaches can lead to bias (*9*).

Given the need to enhance the scope and quality of protein interaction networks, the use of coevolutionary information on a genome scale represents a game-changing addition; it is also a tremendous challenge. Indeed, it is difficult to find true evolutionary covariation between residues for a single protein because one must minimize the effect of transitive (false-positive) correlations; these can be observed, for example, when two amino acid residues contact the same third residue but do not contact each other. Transitive correlation can be removed by global statistical approaches involving either direct coupling analysis (*10*), pseudo-likelihood optimization (*11*), or machine learning (*12*).

Cong *et al.* used GREMLIN, a pseudo-likelihood method they offer as a public server (*13*). As with most evolutionary tools, GREMLIN starts with multiple sequence alignment of all available orthologs for a target protein. The method is based on constructing a parametric model that generates the observed sequences with the highest probability. However, estimating the parameters of an exact model is computationally intractable; hence, GREMLIN optimizes a pseudo-likelihood function to reduce unlikely couplings early in the search and requires sequences of orthologs across a large number of diverse organisms. Cong *et al.* adapted their approach to one used to determine interprotein contacts, which requires large joint alignments of homologs of protein pairs known to interact (*4*). Cong *et al.* is their investigation of coevolution on a whole-proteome scale. Orthologs mapped across more than 40,000 known bacterial genomes. To understand the challenge, consider that *E. coli* lab strains express ~4262 proteins, with more than 9 million potentially interacting protein pairs. To lessen the complexity, Cong *et al.* performed a prescreen based on a less computationally demanding analysis of residue-residue correlations. Still, close to a million potentially interacting pairs remained to be scrutinized with both direct coupling analysis (*10*) and GREMLIN to rank higher-likelihood candidate protein pairs based on the number of predicted residue couplings. The authors selected the top 21,816 protein pairs from the ranked list and built protein-protein complexes by computational modeling, using the predicted contacts as distance restraints for docking.

Restricting considerations to complexes that displayed the predicted contacts within the putative protein-protein interface reduced the number of protein pairs to 804. This is a substantial gain relative to existing experimental datasets (*14*), but falls far short of the putative bacterial interactome. Underrepresentation was especially pronounced for less widely conserved assemblies and multiprotein complexes. As with previous data-filtering strategies, Cong *et al.*'s multistep algorithm eliminated potential interactions that would have been considered meaningful in later steps of the search. To reduce the number of false negatives, the authors introduced, into the later stages of their analysis, protein pairs reported in previous experimental studies or known to be expressed by the same operon, thus arriving at 1618 pairs.

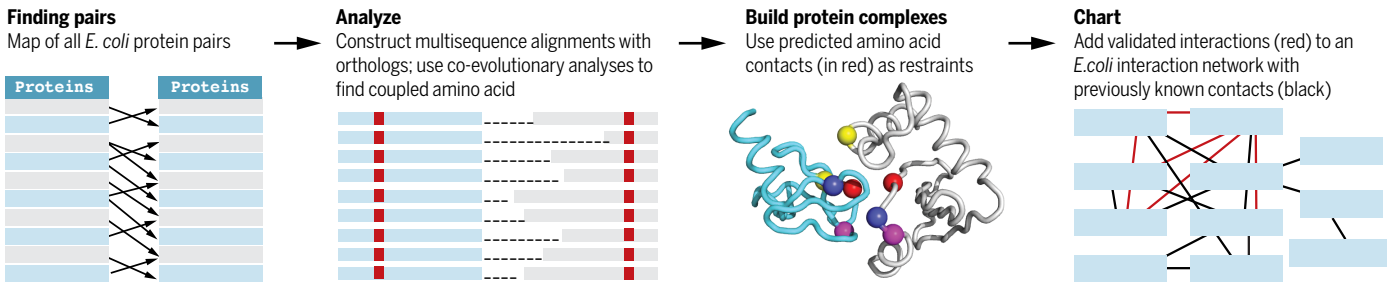
False-positive predictions can also be substantial. In fact, paralogs that do not interact can show strong coevolution. The authors attempted to minimize this uncertainty by eliminating proteins that show nonspecific coevolution with many other proteins, but also conducted independent validation experiments. They compared co-evolution-based interaction predictions with those inferred from high-throughput Y2H and APMS studies relative to structure-based benchmarks derived from x-ray and cryo–electron microscopy analyses of *E. coli* protein assemblies. The coevolution-based screen outperformed existing experimental methods. One potential caveat is that APMS studies report co-complexes, rather than binary interactions.

By exploiting coevolution, Cong *et al.* boosted interactome coverage while leveraging existing bacterial protein–interaction data. In this way, their approach is complementary to experimental ones. In *E. coli*, only 24.7% of the strongly coevolving 1618 pairs are new or unexpected. In fact, 936 interactions were reported previously, homologous templates in the Protein Data Bank (PDB) exist for a further 126 pairs, and 156 pairs have genomic associations. These results increase confidence in the coevolution approach.

The results for *M. tuberculosis* are more informative. Of the 667 predicted pairs, only 203 (30.4%) are supported by homologous templates in the PDB or were reported previously. Thus, the most notable advances of Cong *et al.* may be the potential to map the binary protein interfaces and global interaction networks of bacterial pathogens and study how the core protein interactomes have evolved across microbial species. Adapting this approach to eukaryotic networks represents a formidable future challenge. ■

From protein pairs to complex interaction networks

With the method shown below, co-evolutionary analyses of amino-acid contacts in *E. coli* protein pairs can unearth new interprotein contacts that serve to expand protein interaction networks. The multiple sequence alignment shown in the "Analyze" step is for a hypothetical *E. coli* protein and its orthologs (blue rectangles) and a potentially interacting partner from the same organisms (gray rectangles). Red boxes indicate interprotein contacts predicted by GREMLIN.



GRAPHIC: N. CARY

REFERENCES AND NOTES

1. D. S. Marks *et al.*, *Nat. Biotechnol.* **30**, 1072 (2012).
2. R. F. Service, *Science* (2018). 10.1126/science.aaw2747
3. O. Lichtarge *et al.*, *J. Mol. Biol.* **257**, 342 (1996).
4. S. Ovchinnikov *et al.*, *eLife* **3**, e02030 (2014).
5. Q. Cong *et al.*, *Science* **365**, xxx (2019).
6. B. Schwikowski *et al.*, *Nat. Biotechnol.* **18**, 1257 (2000).
7. A. Kumar, M. Snyder, *Nature* **415**, 123 (2002).
8. J. S. Bader *et al.*, *Nat. Biotechnol.* **22**, 78 (2004).
9. L. J. Lu, *et al.*, *Genome Res.* **15**, 945 (2005).
10. F. Morcos *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, E1293 (2011).
11. H. Kamisetty *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15674 (2013).
12. S. Wang *et al.*, *PLOS Comput. Biol.* **13**, e1005324 (2017).
13. GREMLIN, http://gremlin.bakerlab.org
14. M. Babu *et al.*, *Nat. Biotechnol.* **36**, 103 (2018).

ACKNOWLEDGMENTS

A.E. acknowledges financial start-up support from Boston University; S.V. was supported by grants NIH R35 GM118078 and NSF DBI 1759472.

10.1126/science.aay1440