Research Article

Integrating Multimeric Threading With High-throughput Experiments for Structural Interactome of *Escherichia coli*

Weikang Gong [1,2†], Aysam Guerler [1†], Chengxin Zhang [1], Elisa Warner [1], Chunhua Li [1,2*] and Yang Zhang [1,3*]

*1 - **Department of Computational Medicine and Bioinformatics,** University of Michigan, Ann Arbor, MI 48109, USA*
*2 - **Faculty of Environmental and Life Sciences,** Beijing University of Technology, Beijing 100124, China*
*3 - **Department of Biological Chemistry,** University of Michigan, Ann Arbor, MI, 48109, USA*

Correspondence to Chunhua Li and Yang Zhang: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. *chunhuali@bjut.edu.cn (C. Li), zhng@umich.edu (Y. Zhang)*
https://doi.org/10.1016/j.jmb.2021.166944
Edited by Michael Sternberg

Abstract

Genome-wide protein–protein interaction (PPI) determination remains a significant unsolved problem in structural biology. The difficulty is twofold since high-throughput experiments (HTEs) have often a relatively high false-positive rate in assigning PPIs, and PPI quaternary structures are more difficult to solve than tertiary structures using traditional structural biology techniques. We proposed a uniform pipeline, Threpp, to address both problems. Starting from a pair of monomer sequences, Threpp first threads both sequences through a complex structure library, where the alignment score is combined with HTE data using a naïve Bayesian classifier model to predict the likelihood of two chains to interact with each other. Next, quaternary complex structures of the identified PPIs are constructed by reassembling monomeric alignments with dimeric threading frameworks through interface-specific structural alignments. The pipeline was applied to the *Escherichia coli* genome and created 35,125 confident PPIs which is 4.5-fold higher than HTE alone. Graphic analyses of the PPI networks show a scale-free cluster size distribution, consistent with previous studies, which was found critical to the robustness of genome evolution and the centrality of functionally important proteins that are essential to *E. coli* survival. Furthermore, complex structure models were constructed for all predicted *E. coli* PPIs based on the quaternary threading alignments, where 6771 of them were found to have a high confidence score that corresponds to the correct fold of the complexes with a TM-score >0.5, and 39 showed a close consistency with the later released experimental structures with an average TM-score = 0.73. These results demonstrated the significant usefulness of threading-based homologous modeling in both genome-wide PPI network detection and complex structural construction.

Introduction

Most proteins conduct functions through interactions, either permanently or transiently, with other proteins. These interactions result in various protein–protein interaction (PPI) networks, or interactomes,[1] that are essential to accommodate many important cellular processes, ranging from transcriptional regulation to signal transduction and metabolic pathways. Experimental methods to elucidate these networks are, however, limited and many of them, including yeast-two hybrid and tandem-affinity purification, have high error rates up to 50%.[2] Furthermore, these high-throughput

experimental (HTE) methods only address the issue of what proteins interact, but cannot provide information as to where and how the proteins interact; this information is critical for understanding the biophysical mechanisms of the interaction networks and/or developing new therapies to regulate the networks.[3]

While structure biology through X-ray and NMR techniques could in principle provide the most accurate structural information of PPIs, these experiments are however often too expensive and labor intensive to be applied on a genomic scale. There are also many complexes that are currently difficult to solve due to technical difficulties in protein expression and crystallization. In *Escherichia coli*, the most studied bacterial organism of our time, for example, there are only 1559 out of the 4280 protein-coding genes (<36%) that have the structures experimentally solved.[4,5] The number of PPI complex structures is even less: as of PDB database in Feburary 2021, *E. coli* only have 707 PPI entries, which counts only for <7% of the ~10,000 putative PPIs in *E. coli*.[4,6] Homology modeling has been proved to be an effective approach to construct structure models by copying the frameworks from homologous PPI templates.[7] But until recently, the approach did not significantly contribute to the elucidation of PPI networks, due to the limited number of available homologous complex structures in the PDB.[7–9] Recent studies have shown that the structural library of PPI interfaces approaches to completion,[10] implicating that most of the complexes should have analogous interfaces in the PDB; this settles a promising base for the template-based structure modeling of a wide-range of interactions if advanced threading methods can be developed to recognize such analogies. There are also excellent efforts that tried to combine interaction data from different resources for large-scale PPI network identification;[11–13] many of the approaches however do not provide 3D structures of the complexes.

In this work, we proposed a new hybrid pipeline, Threpp, which extends the multiple-chain threading protocol[14] to address two central problems of protein interactomes (Figure 1). First, we developed a new Bayes classifier model to integrate high-throughput proteomic data with multimeric threading alignments to improve the accuracy and coverage of PPI recognitions. 3D structures of protein complexes were then constructed for all the predicted PPI pairs by threading the query sequences through a non-redundant complex structure library. Different from several existing homology-based methods that build complex structures by multiple-chain sequence comparison,[15,16] which requires separate complex library construction and often misses specific binding modes, Threpp deduces complex structure templates directly from monomer chain threading followed by oligomer-based mapping, which enables the multiple binding mode recognition through the entire PDB library. It is also different from the template-based docking[7,9,12] which associates monomer and dimer structures by pure structural similarity, while Threpp detects PPI frameworks and the monomer–dimer associations using profile-based threading alignments which often have a higher accuracy than pure structure comparisons. A similar Bayesian statistics approach was previously utilized by Lee et al. for functional gene linkage assignments, which allows integration of evidence from diverse sources for more accurate network construction.[17] Nevertheless, the functional linkages do not necessarily indicate physical protein–protein interaction, the latter of which is the focus of this study. In addition, an additional focus of Threpp is on the elucidation of the structural characteristics of these identified PPIs through multi-chain threading and monomeric structure recombination.

To examine the accuracy of Threpp, we carefully benchmarked the strength and weakness of the pipeline in PPI recognition on large-scale gold standard datasets. As a case study, the pipeline was applied to the *E. coli* genome to construct the structural networks of the species, with results revealing important functional implications of the modeled interactome. The Threpp algorithm, together with the structural models of all PPIs for the *E. coli* genome, are made freely downloadable to the community at https://zhanglab.ccmb.med.umich.edu/Threpp/.

Results

Benchmark test of Threpp on PPI assignments

To train and test Threpp for PPI recognitions, we collected a 'Gold Standard' (GS) set of PPIs in the *E. coli* that have definite positive and negative references as assigned by Hu et al.,[18] where the positive samples contain 763 experimentally-established physical interactions obtained from DIP,[19] BIND[20] and INTACT[21] databases, and the negative set consists of 134,632 putatively non-interacting protein pairs compiled from the protein pairs belonging to different cellular compartments (see Table S1 in Supplementary Information, SI). Here, membrane proteins were excluded due to the close physical proximity (and potential physical interaction) with both cytoplasmic and periplasmic proteins.

PPI recognition by individual threading and HTE methods. Table 1 and Figure 2 presents the true positive rate (TPR) and false positive rate (FPR) of PPI assignments for the test proteins by Threpp based only on the Z-score of dimeric threading alignments, Z_{com} (named as 'Threpp_threading', see Methods), where the detail of the data is listed in Table S2. Here, $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$, with the standard true

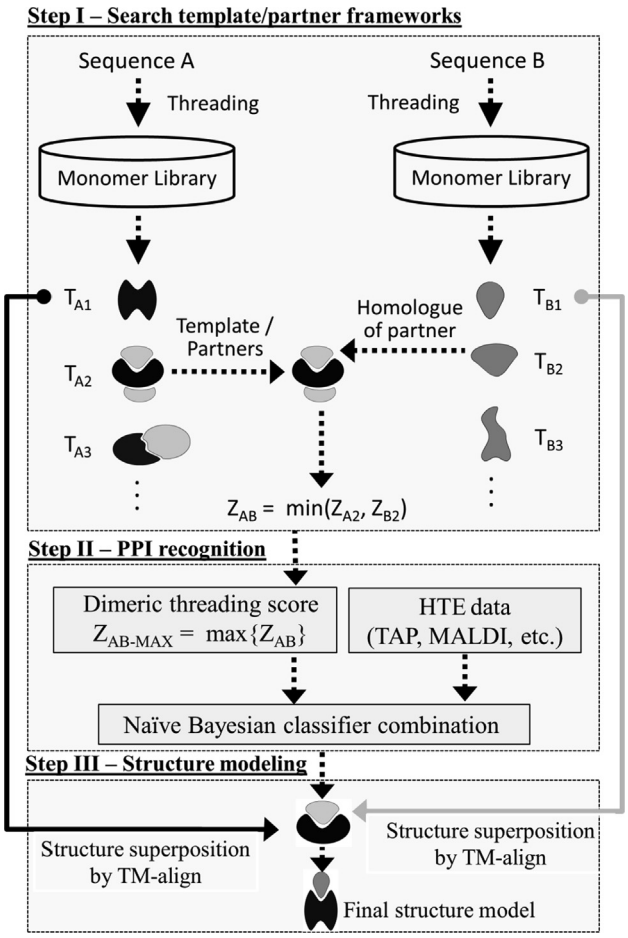

Figure 1. Flowchart of Threpp for PPI recognition and structure construction. The pipeline consists of three steps of threading-based PPI framework identification, Bayesian classifier PPI recognition, and PPI complex structure construction by monomer/dimer template recombination.

Table 1 Summary of PPI recognition by different methods.

	Num of Preys	Num of Baits	Num of Detected interactions	MCC	TPR	FPR
Individual datasets from high-throughput experiments and threading						
Tandem affinity purification (Butland set)	1000	530	6067[a]	0.54	36.2%	0.05%
MALDI-TOF (Arifuzzaman set)	4339	4339	11,478[a]	0.41	32.4%	0.16%
Tandem affinity purification (Hu set)	4225	4225	5993[a]	0.35	24.4%	0.13%
Yeast-two hybrid (Rajagopala set)	3606	3305	2191[a]	0.27	9.6%	0.02%
Threpp_threading	4280	4280	28,263	0.41	26.2%	0.08%
Bayes combinations						
Classifier without Threpp_threading	3459[b]	3459[b]	7872	0.58	42.4%	0.07%
Threpp	4280	4280	35,125	0.64	59.1%	0.14%

[a] With the repeated PPIs (e.g., A-B and B-A) removed from the 4 HTE datasets respectively, the numbers of PPIs become 6067, 11478, 5993 and 2191 from the original ones 6234, 11511, 5993 and 2234.
[b] The number of preys/ baits for the classifier without Threpp_threading is calculated by the union set of preys/baits from the HTE datasets used to train the classifier.

positive (*TP*), true negative (*TN*), false positive (*FP*) and false negative (*FN*) calculated by comparing the PPI predictions with the GS assignments. As a comparison, we also list the results from four sets of HTEs, including two tandem-affinity purification (TAP) sets ('Butland set'[22] and 'Hu set'[18]), the 'Arifuzzaman set' derived through matrix-assisted laser desorption/ionization time-of-fight (MALDI-TOF) mass spectrometry,[23] and the 'Rajagopala set' obtained by yeast two-hybrid (Y2H) screening.[24] While the TPRs of HTE studies can be limited by the prey/bait proteins

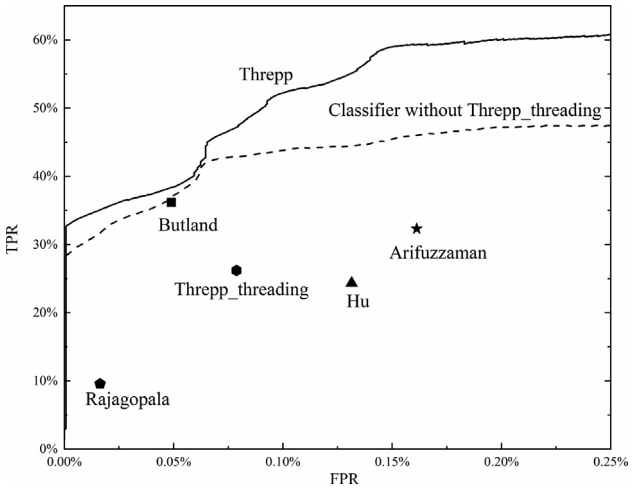

Figure 2. True positive (TPR) and false positive (FPR) rates of PPI recognition by different approaches. The 4 HTE datasets include two tandem-affinity purification (TAP) sets ('Butland'[22] and 'Hu'[18] sets), the 'Arifuzzaman' set derived through matrix-assisted laser desorption/ionization time-of-fight (MALDI-TOF) mass spectrometry,[23] and the 'Rajagopala' set obtained by yeast two-hybrid (Y2H) screening.[24] The predictions of Bayesian classifiers combining different sources of interaction evidences, with and without Threpp_threading, are shown in solid and dashed lines, respectively.

involved in the PPI assayed, it is not the major factor to decide the performance here. In fact, Table 1 shows that the 'Butland set' that only has 1000 preys and 530 baits produces a 4-times greater TPR than the 'Rajagopala set' with a much larger number of preys (3606) and baits (3305). Accordingly, the MCC of the former is time times higher than that of the latter. This is partly due to the 'Butland set' having a higher portion of essential proteins (19.8%), which participate in more PPIs than non-essential proteins.

Table 1 (upper panel) also summarizes the Matthew's correlation coefficient (MCC) by the individual methods, where $MCC = (TP \times TN - FP \times FN)/\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}$ represents a balanced metric of precision and recall of the PPI predictions. While the MCC of Threpp_threading (0.41) is lower than that of the 'Butland set' (0.54), it is comparable or slightly higher than other HTE results, including the 'Arifuzzaman set' (0.41), the 'Hu set' (0.35), and the 'Rajagopala set' (0.27).

Bayesian classifier models increase PPI recognition accuracy of individual methods. In the lower panel of Table 1, we list the results of combined data from different methods by Threpp. First, we used the Bayesian classifier to combine the data from 4 HTE datasets, which results in a higher MCC (=0.58) than all individual datasets. As shown in Figure 2 (the dashed curve), both TPR and FPR increase with the decrease of the threshold, but the curve is above all individual experimental datasets, demonstrating the effectiveness of the Bayesian classifier model in selecting correct PPIs. Nevertheless, the MCC difference between the Bayesian model (0.58) and the best HTE data from the 'Butland set' (0.54) is modest.

After combining the HTE with the Threpp_threading models, the MCC is increased to 0.64, which is 18.5% higher than the best individual dataset from the 'Butland set'. This difference in MCC corresponds to a p-value = 1.1 E−59 in the Student's t-test, indicating that the difference is statistically significant. The result suggests that although the accuracy of the threading score is not high on its own, the modeling data is highly complementary to the HTE evidences, where a naïve Bayesian combination of the computer-based and experimental data can thus result in a highly significant improvement of both the recall and the precision of the PPI predictions.

Integrating threading model with HTE data for E. coli network detection

The *Escherichia coli* genome contains in total 4280 protein-coding genes.[25] As an application, we used Threpp to evaluate all the 9,157,060 putative pairs by the Bayesian combination of dimeric threading and HTE datasets.[18,22–24] The 4 HTE datasets are listed in Table S3 where the PPI with at least one protein not belonging to the gold standard sets and the 4280 proteins was removed, which does not affect the PPI prediction for the 4280 proteins. Despite the huge number of putative interactions, only 4280 × 2 monomer threading runs are needed with the interaction frameworks assigned by a pre-calculated homology look-up table for all templates, where the genome-scale

network calculation is fast with ~2 hours on a 2000-HPDL1000h core cluster.

The experiment yielded 35,125 confident PPIs (Table S4), which has a likelihood rate score above 1.87 by Threpp (see Eq. (3) in Methods). In case where the HTE data are not available, only Threpp threading scores are employed for the targets with a stringent complex framework Z-score cutoff of $Z_{com} \geq 25$. Our benchmark results on the GS datasets show that the PPI assignments with such likelihood score and Z_{com} cutoffs have an average accuracy of 0.996. Overall, these interactions are combined from 28,263 PPIs by Threpp_threading and 21,932 by the four HTE datasets, where there are only 1153 PPIs in the intersection set of the two and 13,917 were dropped off by Threpp due to insufficient likelihood rate score. These predicted interactions contain 451 out of the 763 PPIs in the GS dataset, which is significantly higher than the number of GS PPIs predicted by either Threpp_threading (200) or the four HTE approaches (346).

Here, if we ignore the threading alignments and only combine the HTE data, the number of PPIs detected by Threpp will be reduced to 7872 that have the similar level of likelihood score, which corresponds to only 22% of all PPIs identifiable by the full Threpp pipeline. These data demonstrate again a high complementarity of the technical approaches of the computer-based threading alignments to the HTE, and in particular the impact of consideration of threading-based approach on the hybrid PPI recognitions, although we believe both approaches tend to detect permanent PPIs.

PPI networks reveal dominant roles of essential proteins in E. coli

The 35,125 high-confidence PPI assignments detected by Threpp involve 3273 proteins. Based on these PPIs, we constructed a comprehensive *E. coli* protein interaction network (Figure 3(a)). In the plot, nodes represent individual proteins with edges being the interactions between proteins, where self-loops (corresponding to orphan proteins) and multiple edges (repeated PPI predictions) have been excluded.

Node degree distribution is scale free. Figure 3 (b) shows the degree distribution of the PPI networks for all 3273 involved proteins, which follows a power-law of $P(k) \propto k^{-1.33}$, where the degree (k) of a protein node equals to the number of edges that have this node as one of its endpoints. This network possesses two outstanding characteristics which are important to facilitate the biological functionality and evolution of the *E. coli* genome. First, there are dominantly more proteins in the genome with few interaction partners; this property of PPI networks helps enhance the robustness of the network against random mutations in the evolution, as the overall network is not influenced by the deletion and insert of individual proteins. On the other hand, the scale-free nature of the degree distribution indicates that a non-trivial number of proteins, which is significantly higher than what is expected from a normal distribution, have many interaction partners; this feature allows a substantial number of important proteins to serve as hubs of interactions and dominate the functional interaction networks. The scale-free property of *E. coli* PPI network is consistent with the previous finding by Rajagopala et al. in their Y2H experiment.[24]

Essential proteins interact with more partners than non-essential ones. In Figure 3(c), we present the degree distributions of PPI networks for two sets of essential and non-essential proteins separately, where the 303 essential proteins are taken from Baba *et al*. that were found unable to be deleted from the chromosome for the survival of *E. coli* through the large-scale gene-deletion assay, and the rest are considered as 'non-essential'.[26] While both protein sets follow a stringent power-law distribution (i.e., $P(k) \propto k^{-0.60}$ for essential and $P(k) \propto k^{-1.34}$ for non-essential proteins), the average connectivity (or degree k) per node is significantly higher for essential proteins (33.5) than for non-essential genes (18.5). In particular, the percentage of proteins with more than 34 interaction partners in the essential proteins (30%) is much larger than that in the non-essential proteins (15%), indicating that the essential proteins tend to serve as the interaction hub which has resulted in their significant functional importance for *E. coli* to survive. The significantly higher average connectivity per node for essential proteins compared to non-essential genes indicates that the hub character of the essential proteins is well predicted.

The scale-free property of the predicted interactome is not necessarily an indicator for the PPI prediction accuracy. Although it is difficult for a uniformly randomly generated network to follow a scale-free distribution,[27] it is still possible to for an incorrect network to become scale-free if its generation follows the Barabasi-Albert preferential attachment process.[28] Meanwhile, as an indirect verification for the accuracy of the predicted PPI, we found that positively predicted protein pairs by Threpp are more likely to share similar biological pathways than negative pairs. Specifically, in terms of Biological Process Gene Ontology (BP GO) term annotation, positive PPI pairs has a significantly higher average BP annotation similarity (average F-measure = 0.439) than negative pairs (average F-measure = 0.192), with p-values <1E−300 by both t-test and Wilcoxon rank sum test, as shown in Figure S2. This is consistently with the previous

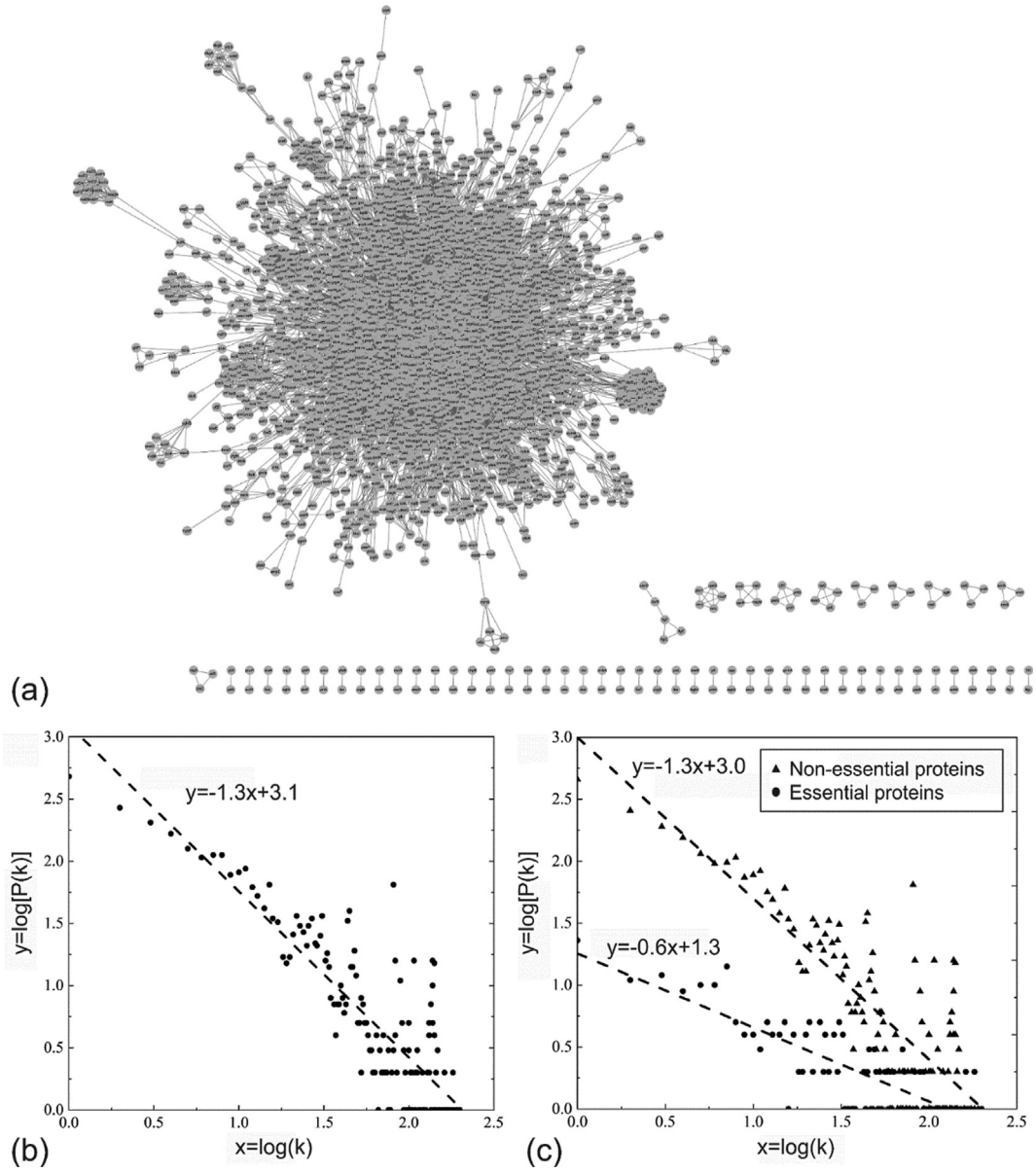

Figure 3. PPI networks and degree distributions of the *E. coli* genome. (a) PPI networks constructed from 35,125 high-confidence PPIs by Threpp, which involve 3,273 proteins. (b) Distribution of PPI node degree (k) that is defined as the number of edges cross the considered node in the network. (c) PPI node degree distribution for essential (circles) and non-essential (triangles) proteins. Lines in (b) and (c) are power law fit to.$P(k) \propto k^{-\gamma}$

observation that true PPI proteins pairs are more likely to share similar BP annotations than non-interacting pairs.[29]

Betweenness centrality. To examine the centrality of proteins in the PPI network, we define the betweenness centrality (BC) of a protein node v by[30]:

$$BC(v) = \sum_{s \neq t \neq v} \frac{\sigma(s,t|v)}{\sigma(s,t)} \quad (1)$$

where $\sigma(s,t)$ denotes the number of shortest paths from nodes s to t, and $\sigma(s,t|v)$ is the number of the shortest paths from s to t that cross through v. The sum in Eq. (1) runs through all node pairs in the network excluding the target node v. Here, although both BC and degree (k) defined above are related to the number of interaction partners for a given protein node, BC measures the number of the shortest paths passing through one node and reflects the information flow through the protein, i.e., a protein with a higher BC tends to control more functional flow of the PPI networks.

We also calculated the betweenness centrality (BC) distributions of PPI networks for all 3273 involved proteins, and for two sets of essential and non-essential proteins separately, with the results shown in Figure S1(a) and (b). For the former, it follows a power-law of $P(BC) \propto BC^{-1.8}$,

meaning that few proteins mediate many information communications in the PPI network and majority of network proteins mediate few. Additionally, while both protein sets follow a stringent power-law distribution (i.e., $P(BC) \propto BC^{-1.4}$ for essential and $P(BC) \propto BC^{-1.8}$ for non-essential proteins), the average BC per node is significantly higher for essential proteins (0.04) than for non-essential genes (0.02). In particular, the percentage of the proteins whose BC more than 0.04 in the essential proteins (21%) is much larger than that in the non-essential proteins (13%), indicating again that the essential proteins tend to serve as the information communication hubs exerting their significant functions for *E. coli* to survive.

In Table S5, we list the BC values for all protein nodes in *E. coli* that have at least one interaction partner in the Threpp predicted PPI networks. The top ten nodes with the highest BCs are presented in Table 2, which all correspond to the functionally important proteins involving complex cellular processes, including chaperone, elongation factor, transcriptional regulatory and ribosomal proteins.

As an illustration, we present in Figure 4 a local PPI network involving the DnaK protein, which has the highest BC score (=0.049). DnaK is known to serve as a chaperone to promote protein folding, interaction and translocation, both constitutively and in response to stress, by binding to unfolded polypeptide segments.[31] Here, RcsA, RcsB (with the 10th largest BC value) and RcsD are all involved in the Rcs phosphorelay pathway, a complex signal transduction system. Through this pathway, phosphate travels from the phosphotransfer protein RcsD to RcsB, which is essential to the regulation of a variety of cellular processes in the bacteria. In this example, the BC-based analysis helps to reveal the key role that the DnaK protein exerts in connecting the metabolic pathway (Rcs phosphorelay pathway) and cell process (cell division regulated by gene ftsA).[32] Additionally, through searching for the PPIs (in Figure 4) in PPI databases, we found that there are 181 out of 217 (83%) PPIs described in IntAct database (Table S8), indicating that Threpp is complementary to existing PPI databases. With the PPI network data provided by the Threpp modeling, the BC analysis can be extended to other systems for key protein and pathway identifications to facilitate various medical and pharmaceutical studies.

Table 2 The ten proteins with the highest betweenness centrality (BC) values.

ID	BC	Name of proteins
DnaK	0.049	Chaperone protein DnaK
TufA	0.037	Elongation factor Tu 1
RpsB	0.029	30S ribosomal protein S2
MetN	0.029	Methionine import ATP-binding protein MetN
LpdA	0.027	Dihydrolipoyl dehydrogenase
RplL	0.027	50S ribosomal protein L7/L12
TufB	0.027	Elongation factor Tu 2
RlmN	0.020	Dual-specificity RNA methyltransferase RlmN
RplV	0.018	50S ribosomal protein L22
RcsB	0.018	Transcriptional regulatory protein RcsB

Structural modeling of protein interactome in E. coli

For structural interactome, Threpp was used to create 3D structure models for all the predicted 35,125 PPIs (see http://zhanglab.ccmb.med.umich.edu/Threpp/Ecoli3D.zip), where 6771 are found to have a Threpp S-score >13 (see https://zhanglab.ccmb.med.umich.edu/Threpp/download/Ecoli3D.txt for S-score values of all the PPI complexes). Here, S-score is defined in Eq. (4) in Methods for estimating model quality of Threpp predictions. In a previous benchmark study,[14] it was shown that 78% of the dimer-threading models with a S-score >13 can have a TM-score >0.5 to their co-crystallized reference structures, indicating correct quaternary structure fold.[33] Below, we selected two complexes from a DMSO reductase (DmsAB) and a hetero-trimeric xanthine dehydrogenase (YagRST), as illustrative examples (S-score >13) to analyze in detail the Threpp models. Although these PPIs have been shown critical to the function of *E. coli*, the interactions were not reported by any of the four high-throughput experimental datasets. However, DmsA-DmsB interaction is described in DIP, IntAct and STRING databases, YagR-YagS in STRING, and the remaining interactions YagR-YagT and YagS-YagT in both IntAct and STRING.

Dimethyl sulfoxide reductase complex (DmsAB). *E. coli* is well known to withstand anaerobic conditions through the utilization of correlated reductases in anaerobic media, while DmsAB is a critical dimethyl-sufoxide reductase complex that supports the bacterial growth in anaerobic media via electron transport. Although no structure has been solved for any of the protein components, there are several experimental evidences that can be used as indirect validations of the Threpp structure modeling. For example, DmsA and DmsB are known to contain one (FS0) and four [4Fe-4S] clusters (FS1 to FS4) respectively for electron shuttling,[34] and DmsB is anchored on the membrane via residues Pro80, Ser81, Cys102 and Tyr104, where these resides are also used for mediating the downstream electron transferals.[35]

Figure 5(a) shows a cartoon representation of the Threpp model for the DmsAB complex, which has a high-confidence S-score of 52.3. The monomeric structure models for DmsA and DmsB were derived from the templates of PDB ID 1EU1 (chain A) and 2VPZ (chain B), while the orientation of the monomers was modeled using 2IVF (chains A and B) that was recognized by

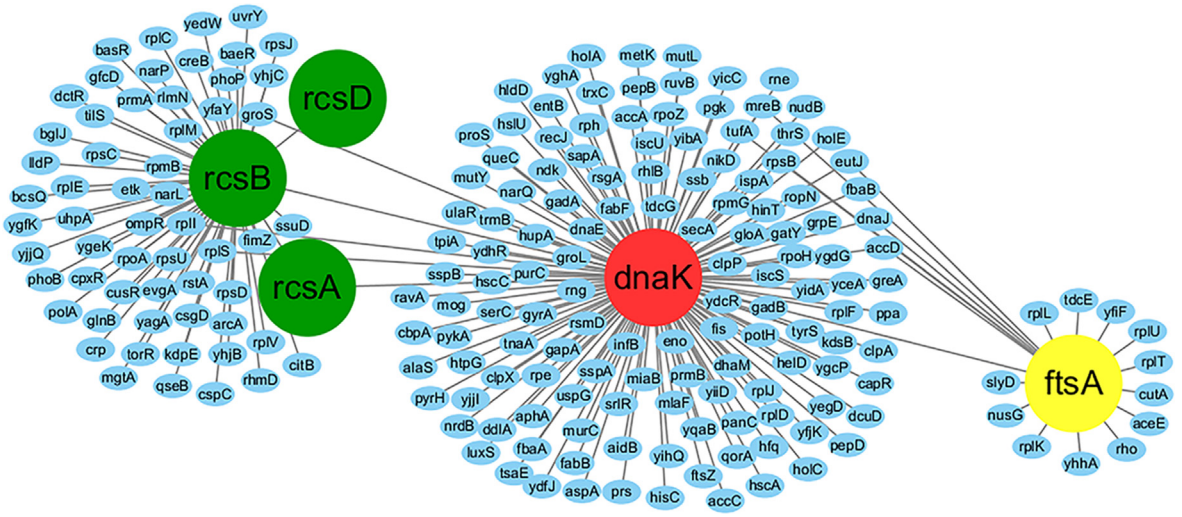

Figure 4. A local PPI network involving the chaperone protein DnaK (red) that mediates the Rcs phosphorelay signaling pathway. The RcsA, RcsB and RcsD proteins (in green) positively regulate the expression of the cell division gene ftsA (yellow) through the interactions with the DnaK protein.

Threpp_threading as dimeric framework (see Table S6). Although the monomer and complex templates have been identified separately, the TM-score and RMSD of the predicted model from the dimer framework are 0.89 and 3.54 Å, respectively, showing a high consistency of the monomer threading and dimeric framework. The framework protein, 2IVF, is a member of the DMSO reductase family and serve as an Ethylbenzene Dehydrogenaes from *Aromatoleum aromaticum*. As highlighted in Figure 5(a), the complex model for DmsAB also contains well-shaped five [4Fe-4S] clusters, demonstrating the close consistency with the insights from the biochemical experiments.[34,35]

Trimeric iron-sulfur complex (YagRST). YagRST is a molybdenum-containing iron-sulfur enzyme located in the periplasm of *E. coli*, which functions in cell maintenance by detoxifying aromatic aldehydes to avoid cell damage.[36] Structurally, YagRST is a heterotrimer complex consisting of a large 78.1 kDa molybdenum-containing subunit (YagR), a medium 33.9 kDa FAD-containing subunit (YagS), and a small 21.0 kDa 2Fe2S-containing subunit (YagT). Built on the threading alignments, Threpp first created monomeric structure models for YagR, YagS and YagT using templates with PDB ID 1RM6 (chain A), 1RM6 (chain B) and 3SR6 (chain A), respectively. Accordingly, three framework templates were collected for constructing the quaternary structural models, including PDB ID 1FIQ (chains C and A, with a high S-score of 142.8), 3HRD (chains C and D, S-score = 98.0), and 1RM6 (chains A and B, S-score = 110.9) (Table S6). Functionally, all the three framework templates are related to molybdenum activities, where the 1FIQ is a mammalian xanthine oxidoreductase which parallels yagTSR in its capabilities as an aldehyde oxidoreductase; the 3HRD is characterized as nicotinate dehydrogenase and consists of similar subunits to YagRST, i.e., two larger molybdopterin subunits, one medium FAD-subunit, and a small FeS subunit; finally, the 1RM6 is another member of the xanthine oxidase family from *Thaura aromatica*. This enzyme differs however in its enzymatic role, demonstrating affinities towards phenolic compounds rather than aldehydes.[34]

The complex structure of YagRST was solved by Correia et al. with a PDB ID: 5G5G, after the structural modeling was performed; this experimental structure can therefore be used as a blind test of the Threpp models.[37] In Figure 5(b), we present a superimposition of Threpp-predicted model (in Cα-trace) and the X-ray structure (cartoon) of the YagRST complex, which has a TM-score = 0.90 and interface RMSD = 2.01 Å. Here, an interface RMSD was calculated on the Cα pairs with an inter-chain distance <5 Å, where the Threpp model covers 96.7% of interface residues. This result shows that a close similarity can be achieved between the Threpp model and the native in both global and interface structures.

Comparison of Threpp models on 39 solved PPI complexes. In fact, there are in total 39 out of the 35,125 protein–protein complexes whose structures have been experimentally solved in PDB since 2016, which is the time when our PPI structure library was constructed, on which the Threpp structural modeling was based. Compared to these experimental structures, the average TM-score of the Threpp models is 0.73, where the

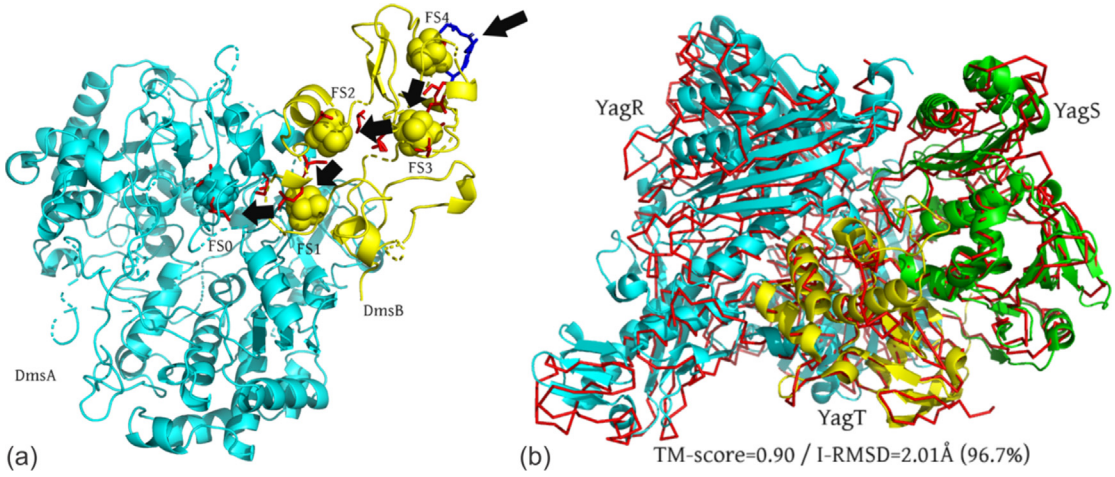

Figure 5. Illustrative examples of quaternary structure models by Threpp. (a) DmsA (cyan) and DmsB (yellow) complex, where the predicted [4Fe-4S] clusters (FS0 to FS4) are highlighted as spheres with arrows indicating the direction of electron transportation. The binding sites of the [4Fe-4S] clusters determined by biochemistry experiments are shown as red sticks, where the membrane anchor residues Pro80, Ser81, Cys102 and Tyr104 in DmsB are shown as blue sticks. (b) Trimeric iron-sulfur complex YagRST, where the Threpp model (red lines) is superposed on the X-ray structure (cartoon) that was solved after the modeling was completed. The monomer chains of YagR, YagS and YagT are shown in blue, green and yellow, respectively.

average sequence identity between the target and complex template in our modeling is 48%, and the average sequence identity between the each chain in the complex and the respective template is around 34% (see Table S7 for detailed list of the 39 proteins and https://zhanglab.ccmb.med.umich.edu/Threpp/download/solved_structures.zip for PDB format of the structural models). These results further demonstrate the effectiveness of Threpp to the quaternary structure prediction. Additionally, we checked the modeled complexes with TM-score $\leq$0.5 (5AEE_AB, 6EI9_AB, 6GFL_AB, 5DUD_BD, 5NJ9_BD, 4UHT_AB, 6AGL_AB, 5VM2_AB, 5ZE6_AB, 5DUD_AC, 5CNX_AB and 5XU7_AB, with all TM-score $\geq$0.41), and found that all the monomers are well modeled with TM-score >0.5, whereas the high quality complex templates are not found by Threpp in our database, as shown in Figure S3 for a case study of 6GFL_AB. These data indicate that modeling the inter-chain orientations is more challenging than modeling individual chains, likely due to the incompleteness of experimental complex structures.[38] Further, among these 39 complexes, 31 and 8 are homo- and hetero-dimers. Threpp achieves average TM-scores of 0.71 and 0.81 for homo- and hetero-dimers, showing that the threading process can handle the modeling of both types of complexes.

Conclusion

We developed a new pipeline, Threpp, for recognizing and structure modeling of protein–protein interactions in organisms. Starting from a pair of monomer sequences, dimeric threading was extended to scan both sequences against a complex structural library collected from the PDB. The alignment score of the dimeric threading was then combined with the high-throughput experimental data through a naïve Bayesian classifier model to predict the likelihood of the target sequences to interact with each other, where the quaternary structure models of the identified PPIs were built by reassembling the monomeric alignments with the quaternary template structural frameworks.

The pipeline was tested on a large set of protein pairs containing 763 experimentally established PPIs and 134,632 non-interacting protein pairs compiled from different cellular compartments. It was shown that although the threading-based assignment does not create a higher accuracy of PPI recognitions than the best high-throughput experiments, the combination of them can result in a significantly higher PPI recognition rate with the Matthews correlation coefficient 18.5% higher than the best dataset from the HTE; this increase is mainly attributed to the complementarity of the threading-based approach to the HTE results.

As an application, Threpp was extended to scan all sequences in the *E. coli* genome and created 35,125 high-confidence PPI predictions, which is 4.5 times higher than that without using the threading-based component scores (7872). This significant data boost demonstrates the usefulness of complimentary computer-based PPI predictions in the interactome constructions against the high-throughput experiments. A detailed network analysis was performed on the Threpp PPI predictions, which indicates that the degree of the PPI networks strictly follows a

power-law distribution, consistent with previous studies.[24] This scale free feature is essential to the robustness of the PPI networks against evolution as the majority of proteins interact only with few partners, which makes the networks less sensitive to the deletion and insertion of local protein nodes. On the other hand, a substantial amount of functionally important proteins, which is significantly higher than that expected from a normal degree distribution, are found in direct interactions with nearly twice more proteins than the non-essential proteins. These proteins serve as the hubs of the PPI networks and play an essential role for *E. coli* survival.

To create structure models of the protein–protein interactions, Threpp reassembles the monomer models of each component obtained by single-chain threading approaches on the dimeric framework of the complex from the dimeric threading alignments. 6771 out of the putative 35,125 PPIs are found to have a high confidence score that corresponds to the correct fold of the complexes with a TM-score >0.5. As a case study, two examples from dimethyl sulfoxide reductase (DmsAB) and trimeric iron-sulfur (YagRST) complexes are examined in detail, where the predicted models are found highly consistent with the experimental data from previous functional studies. Overall, 39 complex structures were solved after the structure library was created, where 72% of them have a TM-score >0.5, resulting in an average TM-score 0.73 compared to the native (Table S7).

Historically, as a major technique of PPI assignments, the HTE has a relatively high false positive rate. Meanwhile, despite their accuracy, traditional structural biology techniques (X-ray and NMR) have more difficulties in determining the PPI complex structures than that encountered for monomer proteins. The incompleteness of experimental complex structures put a limit on the number of PPIs that can be detected by threading. These difficulties have frustrated the progress of the interactome studies compared to the success of structural genomics that focuses on the structure and function of monomer proteins. The results presented in this study demonstrate promising improvement on both aspects of interactome through a hybrid pipeline that take advantage of the technical complementarity between computational threading and traditional HTE datasets with analogy-based structure modeling. The results presented in this study demonstrates promising improvement on both aspects of interactome through a hybrid pipeline that combines computational threading and traditional HTE datasets with analogy-based structure modeling. Although the pipeline has been applied only to *E. coli* in this study, it can be readily extended to the study of other organisms. With continuous improvements of the threading techniques and the enlargement of PPI structure datasets through new techniques such as cryo-EM,[39] the Threpp pipeline, which has been made freely downloadable to the community, should find the increasing usefulness on the studies of other interactome systems.

Methods

Threpp consists of three consecutive steps of multiple-chain threading, Bayesian classifier-based interaction prediction, and complex structure construction, where the flowchart is depicted in Figure 1.

Dimer-threading based PPI recognitions

The multi-chain threading procedure in Threpp is extended from a former version of SPRING that was designed to detect complex structure templates for protein pairs of known interactions.[14] Initially, one of the target sequences (e.g. Chain A) is threaded by HHsearch, a profile-profile sequence aligner assisted with secondary structure,[40] against the monomeric template library from the PDB, to create a set of putative templates (T_{Ai}, $i = 1, 2, \ldots$) each associated with a Z-score (Z_{Ai}). Here, the Z-score is defined as the difference between the raw alignment score and the mean in the unit of standard deviation, where a higher Z-score indicates a higher significance and usually corresponds to a better quality of the alignment. In parallel, the opposite chain (e.g. Chain B) is threaded separately by HHsearch through the PDB, yielding a set of templates (T_{Bi})with Z-score (Z_{Bi}). Then, all binding partners of the T_{Ai} are gathered from the oligomer entry that is associated with T_{Ai} in the PDB. If any of the binding partners of T_{Ai} is homologous to any of the high-ranking templates of Chain B (T_{Bi}), an interaction framework is established for the target complex from the oligomer associated with T_{Ai} (middle column in Figure 1).

The homology comparisons between the PDB templates are pre-calculated by an all-to-all PSI-BLAST scan where a homology is defined between two templates if the E-value <0.01. The Z-score of the framework is defined as the smaller of the two monomeric Z-scores. For heterodimer proteins, this threading process is repeated using Chain B as the starting probe to identify binding partners and the frameworks. The confidence of the target chain interactions by the threading alignments is evaluated by the highest Z-score of the complex, named Z_{com}, among all the templates identified by the procedure.

Bayes classifier for multiple evidence combination

To evaluate if the putative chains (A and B) interact, we combined the Z_{com} score with the interaction evidence from HTEs through a model of the naïve Bayes classifiers.[41] With the classifier,

Threpp_threading is encoded by a single binary feature which equals '1' if $Z_{com} \geq 25$ and otherwise '0'. Here, to determine the threshold of 25, we used the positive and negative gold standard sets (https://zhanglab.ccmb.med.umich.edu/Threpp/download/groundtruth.zip), which were randomly split into five subsets of equal size. For each subset, we identified the threshold which maximizes the Matthew's correlation coefficient. The resulting average threshold was subsequently rounded to the closest multiple of five, yielding a Z_{com}threshold value of 25. Repeating this experiment with different subsets did not significantly alter the identified threshold. Furthermore, using a threshold of 20 or 30 only marginally impacts the overall performance of the resulting classifier. In an upcoming study, we applied our method to the Yeast genome, which confirms the same threshold. Data from each of the experimental datasets is also represented by a binary feature which equals '1′ if the corresponding experiment indicates that the pair of proteins interacts and otherwise '0′. In the present study, we identified four experimentally derived PPI networks for *E. coli* from the literature,[18,22–24] although more HTE features can be combined similarly when available.

The classifier is parameterized using the positive and negative gold standard sets, which are randomly split into five subsets with equal size, where four of the five subsets are used to estimate the conditional probabilities for the positive (P) and negative (N) samples by

$$\begin{cases} p_P(f_i) = n_P(f_i)/n_P \\ p_N(f_i) = n_N(f_i)/n_N \end{cases} \quad (2)$$

where $n_{P(N)}(f_i)$ is the number of positive (negative) interacting cases for a given score of f_i of the ith feature ($i = 1, \cdots, 5$ represents the five features from Threpp_threading and HTE datasets), and n_P and n_N are the total numbers of positive and negative interacting cases in the training sample, respectively. The remaining subset of protein pairs are used for testing in Results, where the likelihood of interaction is evaluated by

$$L(f_1, \cdots, f_5) = \prod_{i=1}^{5} [p_P(f_i)/p_N)f_i)]. \quad (3)$$

We note that the likelihood ratio is derived solely from features that are available, indicating that the sample protein pairs are interacting. The remaining features are excluded (i.e. treated as missing evidence) since the unavailability of an experimental confirmation or threading alignments does not indicate whether a pair of proteins interacts or not.

Structure assembly of protein complexes

If the proteins are deemed to interact, the complex structures are constructed by structurally aligning the top-ranked monomer templates of Chain A and B to all putative interacting frameworks using TM-align.[42] The structural alignment is built on the subset of interface residues. The resulting models are evaluated by the Threpp score of

$$\text{S-score} = \min(Z_A, Z_B) + w_1 TM_{min} + w_2 E_{contact} \quad (4)$$

where $Z_{A(B)}$ is the Z-score of the monomer threading alignment by HHsearch for Chain A(B); TM_{min} is the smaller TM-score returned by TM-align when aligning the top-ranked monomer models of A and B to the interaction framework; $E_{contact}$ is a residue-specific, atomic contact potential derived from 3897 non-redundant structure interfaces from the PDB using the formula of RW.[43] The weight parameters w_1 and w_2 are set to 12.0 and 1.4 through a training set of protein complexes to maximize the modeling accuracy of the interface structures.

GO similarity analysis for predicted PPI

As a systematic analysis on whether interacting protein pairs are more likely to participate in the same or similar biological pathways, function annotations for *E. coli* proteins are extracted from UniProt Gene Ontology Annotation (UniProt-GOA version 2021-02-12). In total, there are 3394 *E. coli* proteins that with at least one BP GO annotation among all *E. coli* proteins analyzed in this study. These 3394 proteins result in 5,757,921 protein pairs, among which 27,979 are positive PPI pairs predicted by Threpp and the remaining 5,729,942 are negative PPI pairs. Since UniProt-GOA usually only includes child BP terms, we propagate all BP annotations towards the root of GO hierarchy to obtain the full set of BP annotations. In addition to BP GO terms, there are also Molecular Function (MF) and Cellular Component (CC) GO annotations for *E. coli*. However, CC is not analyzed here because the *E. coli* CC is too simple, i.e., the *E. coli* cell is not divided into cellular compartments by organelle membranes. This analysis does not consider MF either, because our previous study found that PPI correlates poorly with MF similarity.[29]

To quantify the similarity between the set of all BP terms, BP_A and BP_B for the Protein A and B, respectively, the F-measure, also known as F1-score is calculated according to the definition in Critical Assessment of Function Annotation (CAFA) challenge[44]:

$$F = \frac{2 \cdot |BP_A \cap BP_B|}{|BP_A| + |BP_B|} \quad (5)$$

where $|BP_A|$, $|BP_B|$ and $|BP_A \cap BP_B|$ are the number of terms in GO terms sets BP_A, BP_B and their intersections, respectively. F-measure ranges between 0 and 1, where 1 stands for a perfect overlap between two sets of GO terms.

Acknowledgments

We like to thank Dr Peter L Freddolino, Zi Liu, Eric W Bell and Yang Liu for technical assistances.

Authors contributions

Y.Z. conceived and designed the experiments; W. G., A.G. and C.L. performed the experiments and analyzed the data; C.Z. developed the webserver; E.W. helped prepare the figures; W.G., A.G., C.L. and Y.Z. wrote the manuscript. All authors have approved the manuscript for submission.

Funding

This work is supported in part by the National Institute of General Medical Sciences (GM136422 and S10OD026825 to Y.Z.), the National Institute of Allergy and Infectious Diseases (AI134678 to Y. Z.), and the National Science Foundation (IIS1901191, DBI2030790 and MTM2025426 to Y. Z.), and National Natural Science Foundation of China (31971180 and 11474013 to C.L.).

Availability of data and materials

All benchmark and *E. coli* PPI modeling data, together with the on-line server of the Threpp method, are made available at https://zhanglab.ccmb.med.umich.edu/Threpp/.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.166944.

Received 20 November 2020;
Accepted 9 March 2021;
Available online 16 March 2021

Keywords:
protein-protein interaction networks;
multiple-chain threading;
Escherichia coli genome;
structural interactome;
network centrality

† These authors contribute equally

References

1. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M. P., Szpyt, J., et al., (2015). The BioPlex network: A systematic exploration of the human interactome. *Cell*, **162**, 425–440.
2. Montañez, G., Cho, Y.R., (2015). Predicting false positives of protein-protein interaction data by semantic similarity measures. *Curr. Bioinform.*, **8**, 339–346.
3. Archakov, A.I., Govorun, V.M., Dubanov, A.V., Ivanov, Y. D., Veselovsky, A.V., Lewi, P., et al., (2003). Protein-protein interactions as a target for drugs in proteomics. *Proteomics*, **3**, 380–391.
4. Keseler, I.M., Mackie, A., Santos-Zavaleta, A., et al., (2017). The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.*, **45**, D543–D550.
5. UniProt Consortium, (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
6. Burley, S.K., Bhikadiya, C., Bi, C., et al., (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
7. Szilagyi, A., Zhang, Y., (2014). Template-based structure modeling of protein-protein interactions. *Curr. Opin. Struct. Biol.*, **24**, 10–23.
8. Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., et al., (2004). Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.
9. Kundrotas, P.J., Zhu, Z., Janin, J., Vakser, I.A., (2012). Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 9438–9441.
10. Gao, M., Skolnick, J., (2010). Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 22517–22522.
11. Zhang, Q.C., Petrey, D., Garzón, J.I., et al., (2012). PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res.*, **41**, D828–D833.
12. Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., et al., (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.
13. Cong, Q., Anishchenko, I., Ovchinnikov, S., et al., (2019). Protein interaction networks revealed by proteome coevolution. *Science*, **365**, 185–189.
14. Guerler, A., Govindarajoo, B., Zhang, Y., (2013). Mapping monomeric threading to protein-protein structure prediction. *J. Chem. Inf. Model.*, **53**, 717–725.
15. Lu, L., Lu, H., Skolnick, J., (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364.
16. Mukherjee, S., Zhang, Y., (2011). Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*, **19**, 955–966.
17. Lee, I., Date, S.V., Adai, A.T., et al., (2004). A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
18. Hu, P., Janga, S.C., Babu, M., Diaz-Mejia, J.J., Butland, G., Yang, W., et al., (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol.*, **7**, e96.
19. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D., (2000). DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.

20. Bader, G.D., Betel, D., Hogue, C.W., (2003). BIND: the biomolecular interaction network database. *Nucleic Acids Res.*, **31**, 248–250.
21. Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al., (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
22. Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., et al., (2005). Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*, **433**, 531–537.
23. Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., et al., (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.*, **16**, 686–691.
24. Rajagopala, S.V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., et al., (2014). The binary protein-protein interaction landscape of Escherichia coli. *Nature Biotechnol.*, **32**, 285–290.
25. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K.D., et al., (2018). GenBank. *Nucleic Acids Res.*, **46**, D41–D47.
26. Baba, T., Ara, T., Hasegawa, M., et al., (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, 2006–10008.
27. Zhang, C., Zheng, W., Cheng, M., et al., (2021). Functions of essential genes and a scale-free protein interaction network revealed by structure-based function and interaction prediction for a minimal genome. *J. Proteome Res.*, **20**, 1178–1189.
28. Barabási, A.L., Albert, R., (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.
29. Zhang, C., Zheng, W., Freddolino, P.L., et al., (2018). MetaGO: Predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. *J. Mol. Biol.*, **430**, 2256–2265.
30. Freeman, L.C., (1977). A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–41.
31. Zhu, X., Zhao, X., Burkholder, W.F., Gragerov, A., Ogata, C.M., Gottesman, M.E., et al., (1996). Structural analysis of substrate binding by the molecular chaperone DnaK. *Science*, **272**, 1606–1614.
32. Carballes, F., Bertrand, C., Bouche, J.P., Cam, K., (1999). Regulation of Escherichia coli cell division genes ftsA and ftsZ by the two-component system rcsC-rcsB. *Mol. Microbiol.*, **34**, 442–450.
33. Xu, J., Zhang, Y., (2010). How significant is a protein structure similarity with TM-score = 0.5?. *Bioinformatics*, **26**, 889–895.
34. Rothery, R.A., Workun, G.J., Weiner, J.H., (2008). The prokaryotic complex iron-sulfur molybdoenzyme family. *BBA*, **1778**, 1897–1929.
35. Cheng, V.W., Rothery, R.A., Bertero, M.G., Strynadka, N. C., Weiner, J.H., (2005). Investigation of the environment surrounding iron-sulfur cluster 4 of Escherichia coli dimethylsulfoxide reductase. *Biochemistry-Us*, **44**, 8068–8077.
36. Neumann, M., Mittelstadt, G., Iobbi-Nivol, C., Saggu, M., Lendzian, F., Hildebrandt, P., et al., (2009). A periplasmic aldehyde oxidoreductase represents the first molybdopterin cytosine dinucleotide cofactor containing molybdo-flavoenzyme from *Escherichia coli*. *FEBS J.*, **276**, 2762–2774.
37. Correia, M.A., Otrelo-Cardoso, A.R., Schwuchow, V., Sigfridsson Clauss, K.G., Haumann, M., Romao, M.J., et al., (2016). The Escherichia coli periplasmic aldehyde oxidoreductase is an exceptional member of the xanthine oxidase family of molybdoenzymes. *ACS Chem. Biol.*, **11**, 2923–2935.
38. Garma, L., Mukherjee, S., Mitra, P., et al., (2012). How many protein-protein interactions types exist in nature?. *PLoS ONE*, **7**, e38913.
39. Cheng, Y., (2015). Single-particle cryo-EM at crystallographic resolution. *Cell*, **161**, 450–457.
40. Wu, S., Zhang, Y., (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins*, **72**, 547–556.
41. Domingos, P., Pazzani, M., (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, **29**, 103–130.
42. Zhang, Y., Skolnick, J., (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
43. Zhang, J., Zhang, Y., (2010). A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS ONE*, **5**, e15386.
44. Zhou, N., Jiang, Y., Bergquist, T.R., et al., (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, **20**, 1–23.