中国人民大学数学科学研究院

IMS. RUC

集成多聚体穿线和高通量实验的大肠杆菌蛋白质 - 蛋白质相互作用网络及结构的预测

- 巩卫康

- 时间：2021 年 6 月 23 日 上午 10:30-11:30
- 线下：数学科学研究院 316 会议室
- 线上：腾讯会议
  会议号：594 961 960

报告内容：

全基因组蛋白质 - 蛋白质相互作用（PPI）探测是结构生物学中一个重要但未解决的科学问题。由于高通量实验（HTE）在探测 PPI 时通常具有相对较高的假阳性率，同时 PPI 的四级结构预测比用传统结构生物学技术解析三级结构更难预测。我们提出一个计算方法 Threpp 用来解决这两个问题。Threpp 从一对单体序列开始，通过复合物结构库将两个序列联系起来，其中使用朴素贝叶斯分类器模型将比对分数与 HTE 数据相结合，用以预测两条链相互作用的可能性。紧接着，通过界面特定的结构对齐，将单体对齐于二聚体模板重新组装来预测复合物的结构。该方法应用于大肠杆菌（E. coli）基因组并预测了 35,125 对可信的 PPI，比单独的 HTE 高 4.5 倍。PPI 网络分析显示无标度属性，发现基因组进化的鲁棒性和对大肠杆菌生存至关重要的功能蛋白质。Threpp 构建了基于四级结构穿线对齐预测得到的所有（E. coli）PPI 的复合物结构，其中 6771 个预测的复合物结构具有正确的折叠（TM-score > 0.5）；尤其是 39 个预测的复合物结构和最近实验结晶出来的结构非常接近（平均 TM-score = 0.73）。这些结果证明基于多聚体穿线的同源建模在全基因组 PPI 网络探测和复合物结构预测中有重要意义。

报告人简介：

巩卫康，北京工业大学环境与生命学部在读博士，导师为李春华教授，主要研究方向为生物物理、计算生物学及生物信息学。目前在新加坡南洋理工大学做交流访学。

Integrating multimeric threading with high-throughput experiments for structural interactome of *Escherichia coli*

Department of Computational Medicine and Bioinformatics, University of Michigan

Faculty of Environmental and Life Sciences, Beijing University of Technology

Supervisors: Prof. Yang Zhang and Prof. Chunhua Li

Reporter: Weikang Gong

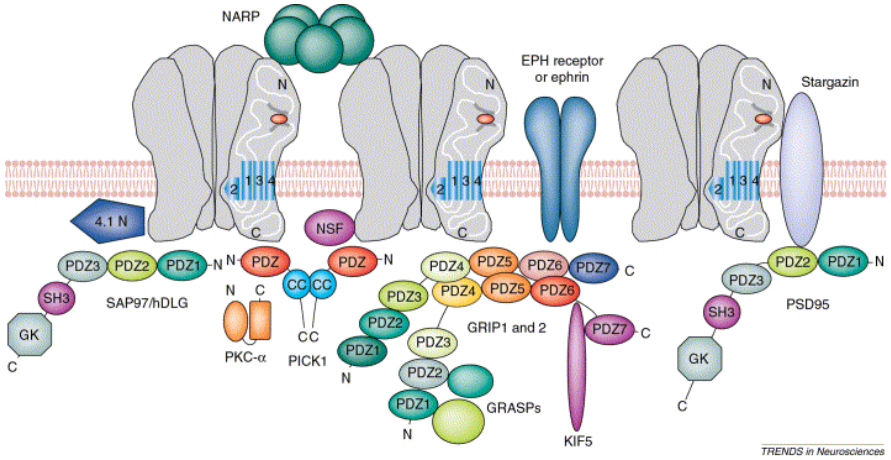
06-23-2021

Overview

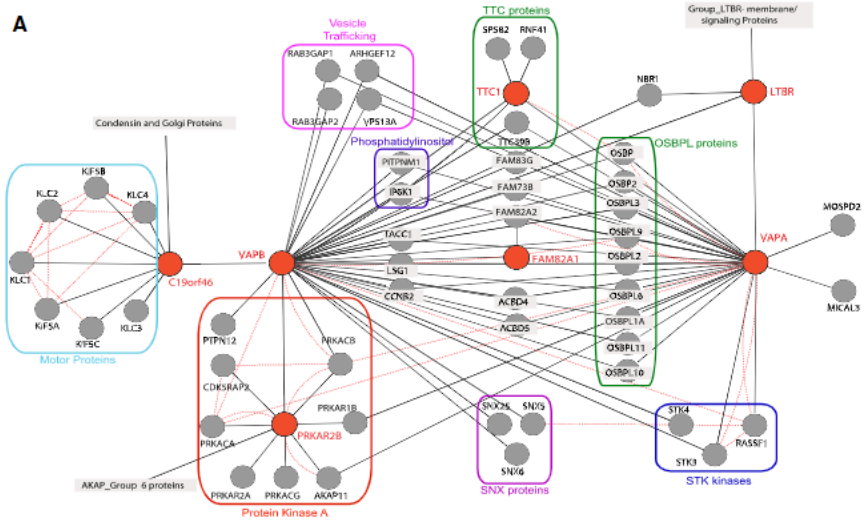
1. Introduction
2. Results
3. Conclusion
4. Server

Introduction-Significance of PPI Network Prediction on Genome Scale

- 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, that are essential to accommodate many important cellular processes, ranging from transcriptional regulation to signal transduction and metabolic pathways

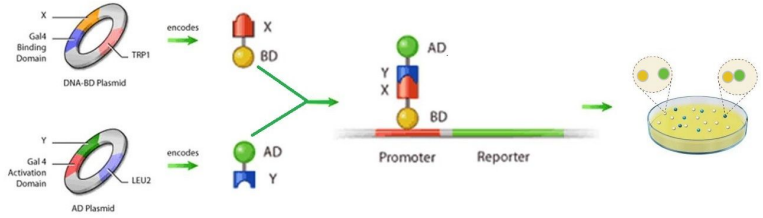


Ref: Huttlin E L., *et al*., ***Cell*** 2019, 162:425-440.

Introduction-High Throughput Experimental Methods for PPI Detection

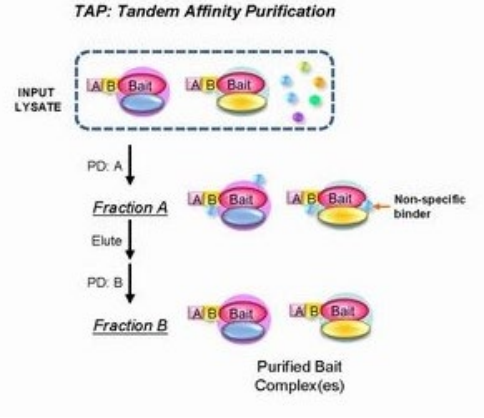
- Experimental methods to elucidate these networks are, however, limited and many of them, including yeast-two hybrid (Y2H) and tandem-affinity purification (TAP), have high error rates up to 50%

Yeast two-hybrid (Y2H)



- 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

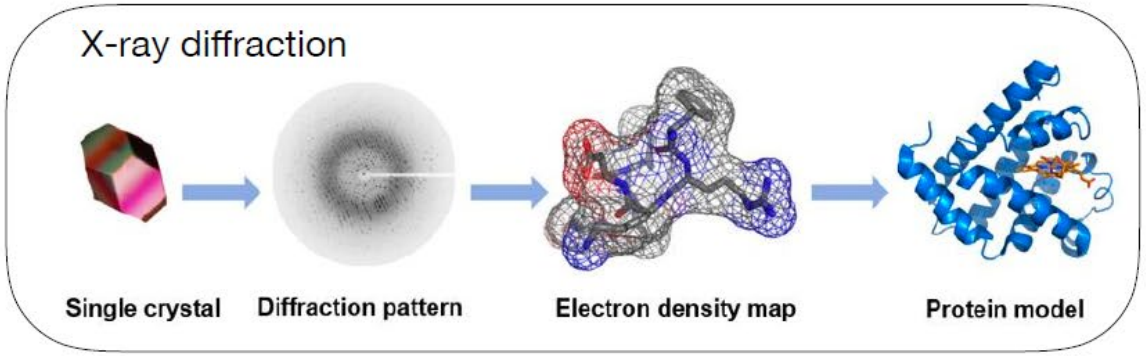
Tandem affinity purification (TAP)

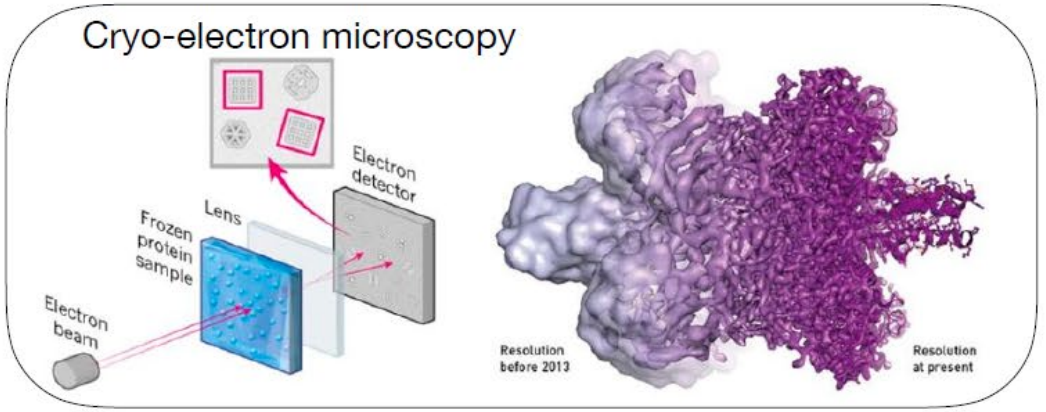


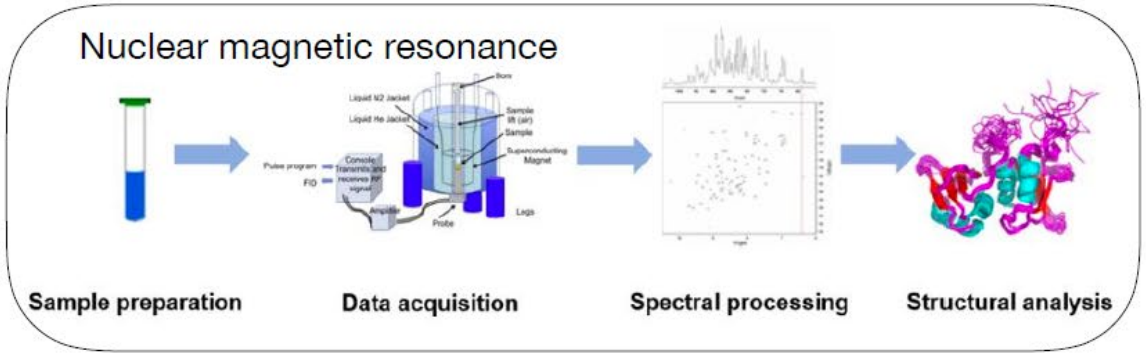
Ref: Montañez G., *et al*., ***Current Bioinformatics*** 2015, 8:339-346.
Archakov AI., *et al*., ***Proteomics*** 2003, 3:380-391.

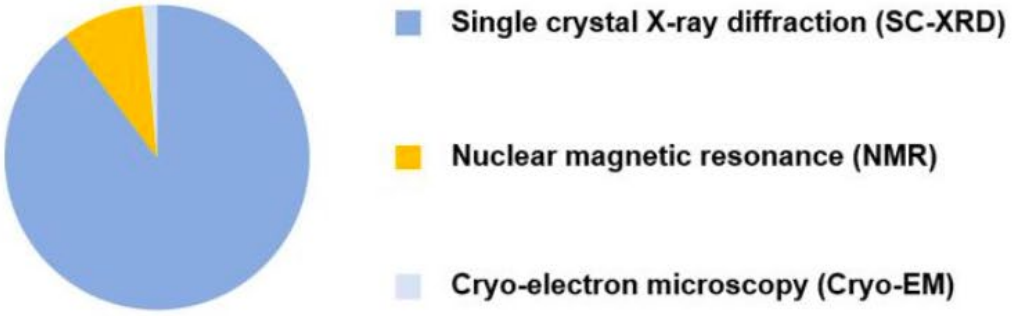
Introduction-Experimental Techniques

- 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







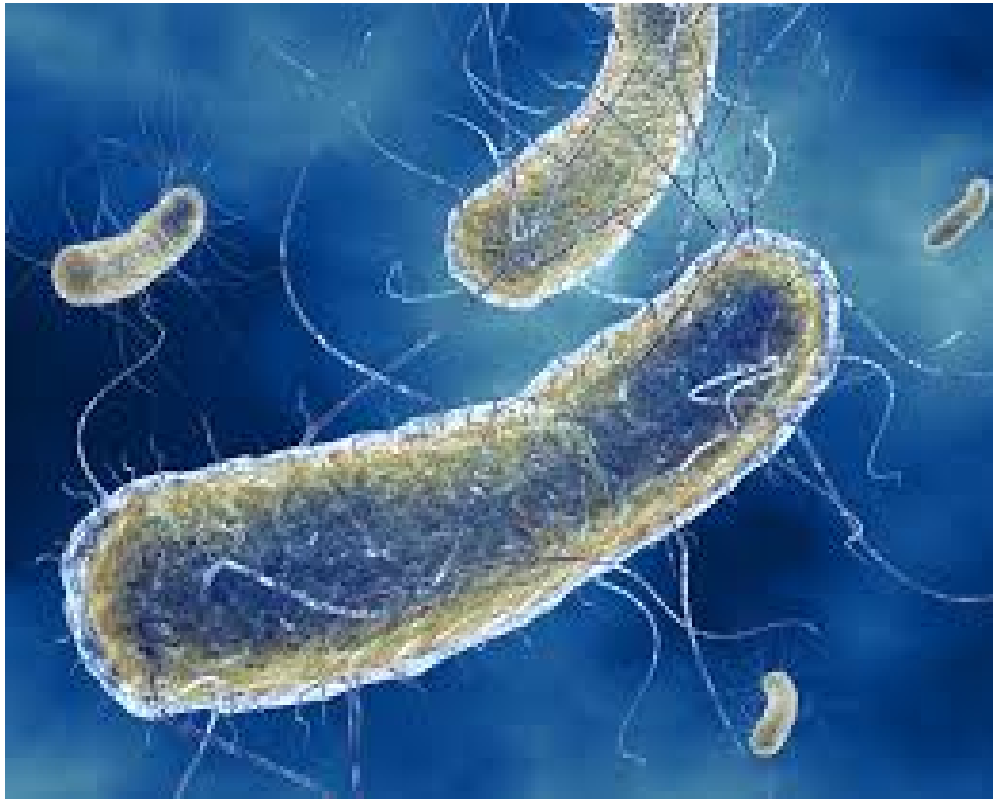


- There are also many complexes that are currently difficult to solve due to technical difficulties in protein expression and crystallization

Introduction-*Escherichia coli* (*E. coli*)

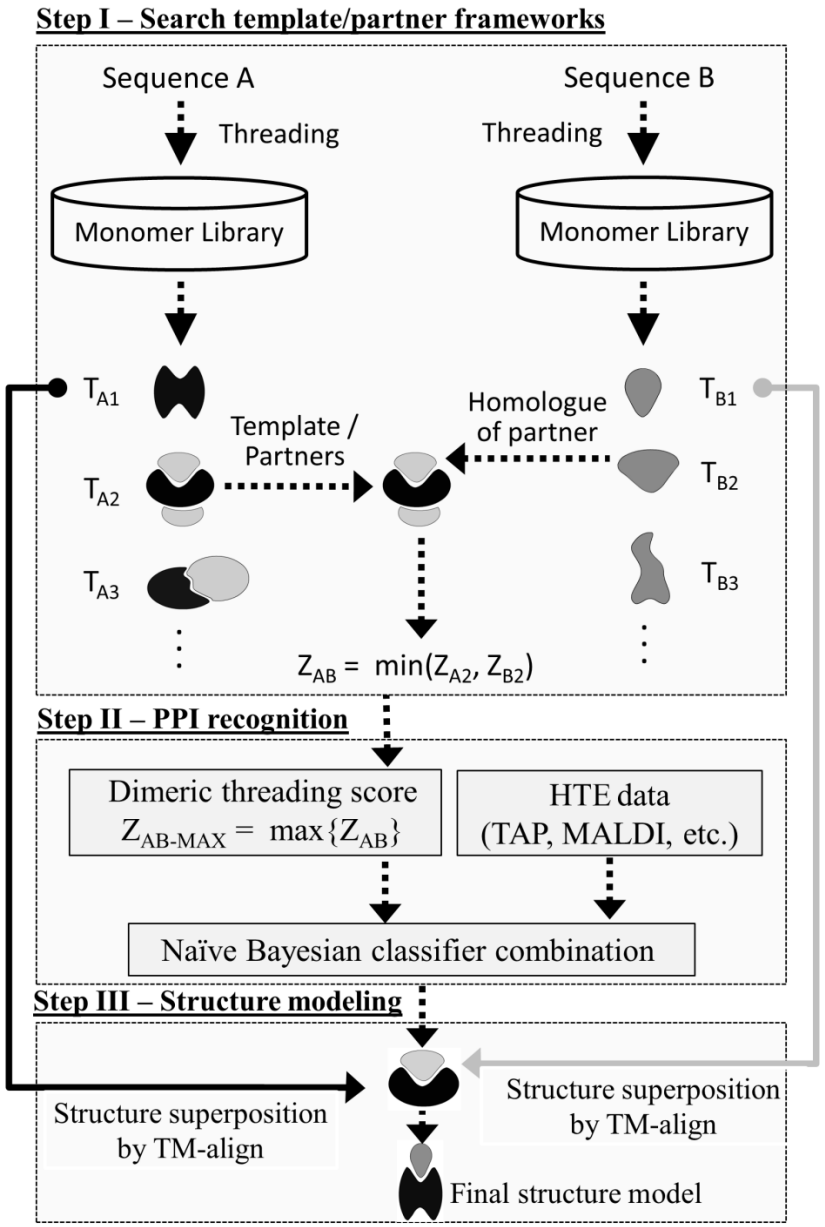
- In *Escherichia coli*, the most studied bacterial organism of our time, for example, there are only 1,559 out of the 4,280 protein-coding genes (<36%) that have the structures experimentally solved

- The number of PPI complex structures is even less: as of PDB database in June 2021, *E. coli* only have 717 PPI entries, which counts only for <7% of the ~10,000 putative PPIs in *E. coli*



Ref: Keseler I M., *et al*., ***Nucleic acids research*** 2017, 45:D543-D550.
UniProt Consortium., ***Nucleic acids research*** 2019, 47:D506-D515.
Burley S K., et al., ***Nucleic acids research*** 2021, 49:D437-D451.

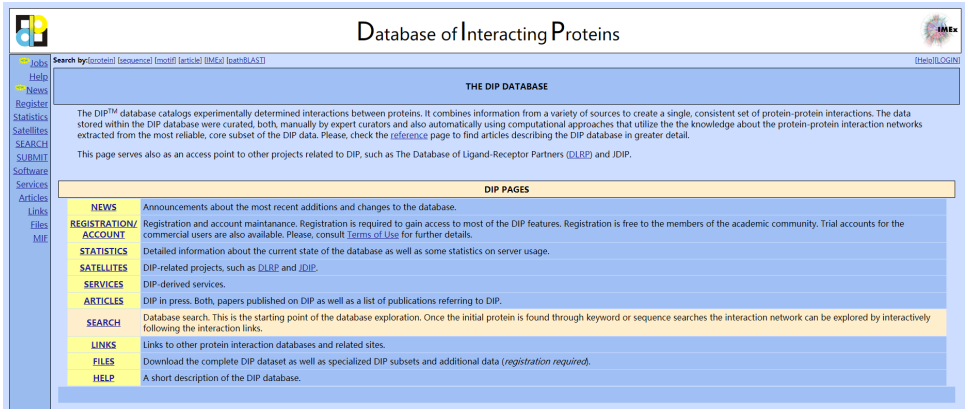
Introduction-Threpp

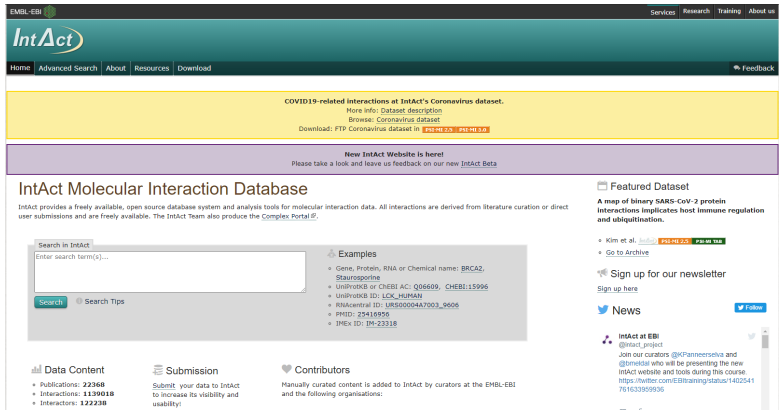


Results-Benchmark Test of Threpp on PPI Assignments

- To train and test the pipeline, we collected a 'Gold Standard' (GS) set of interactions: positive samples 763 obtained from DIP, BIND and INTACT databases

- The negative samples 134,632 compiled from protein pairs belonging to different cellular compartments





Ref: Hu, P., *et al*., ***Plos biology*** 2009, 7: e1000096.
Xenarios, Ioannis., *et al*., ***Nucleic acids research*** 2000, 28: 289-291.
Bader, Gary D., *et al*., ***Nucleic acids research*** 2003, 31: 248-250.
Kerrien, Samuel., *et al*., ***Nucleic acids research*** 2012, 40: D841-D846.

Results-PPI Recognition by Individual Threading and HTE Methods

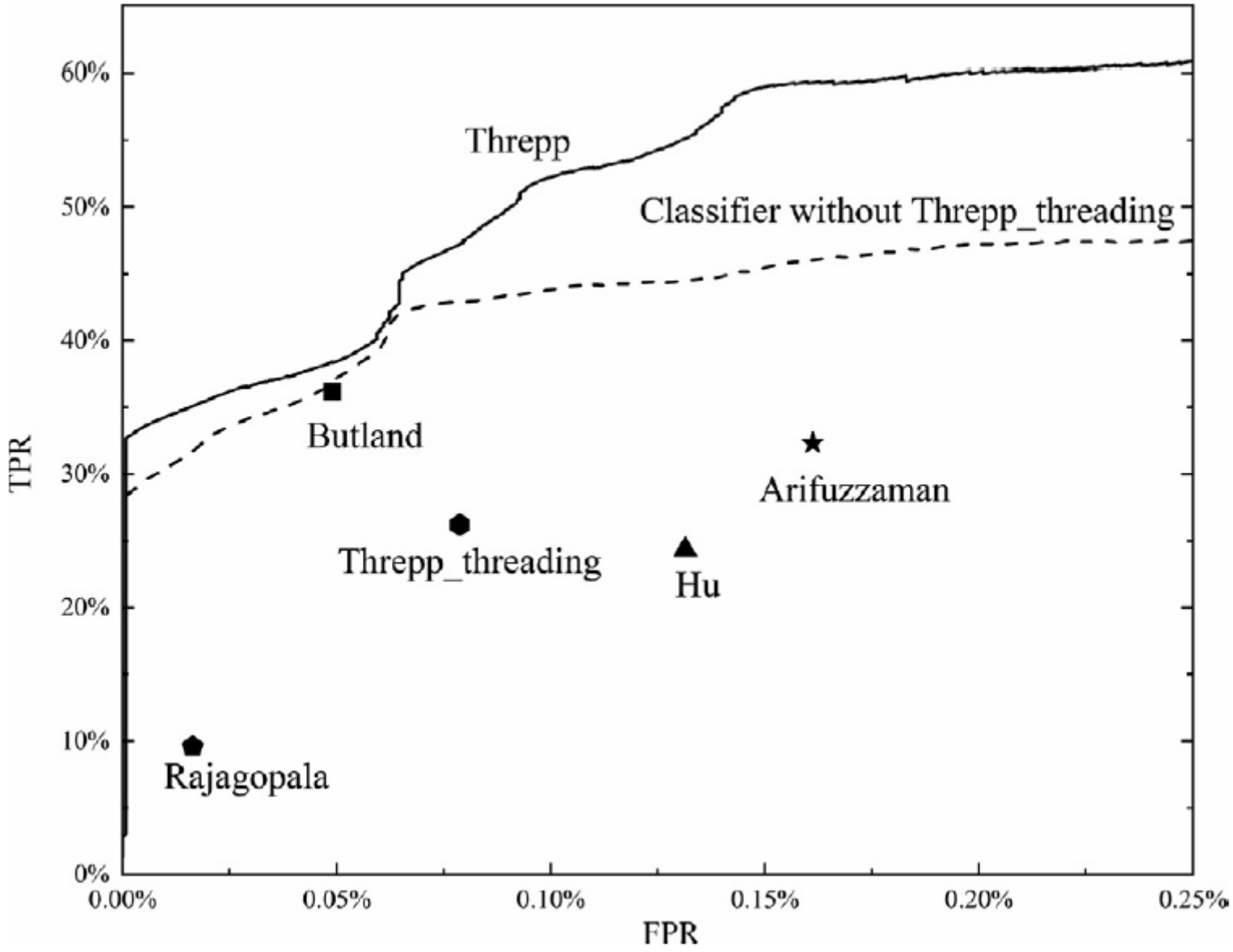
Table 1 Summary of PPI recognition by different methods.

	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%

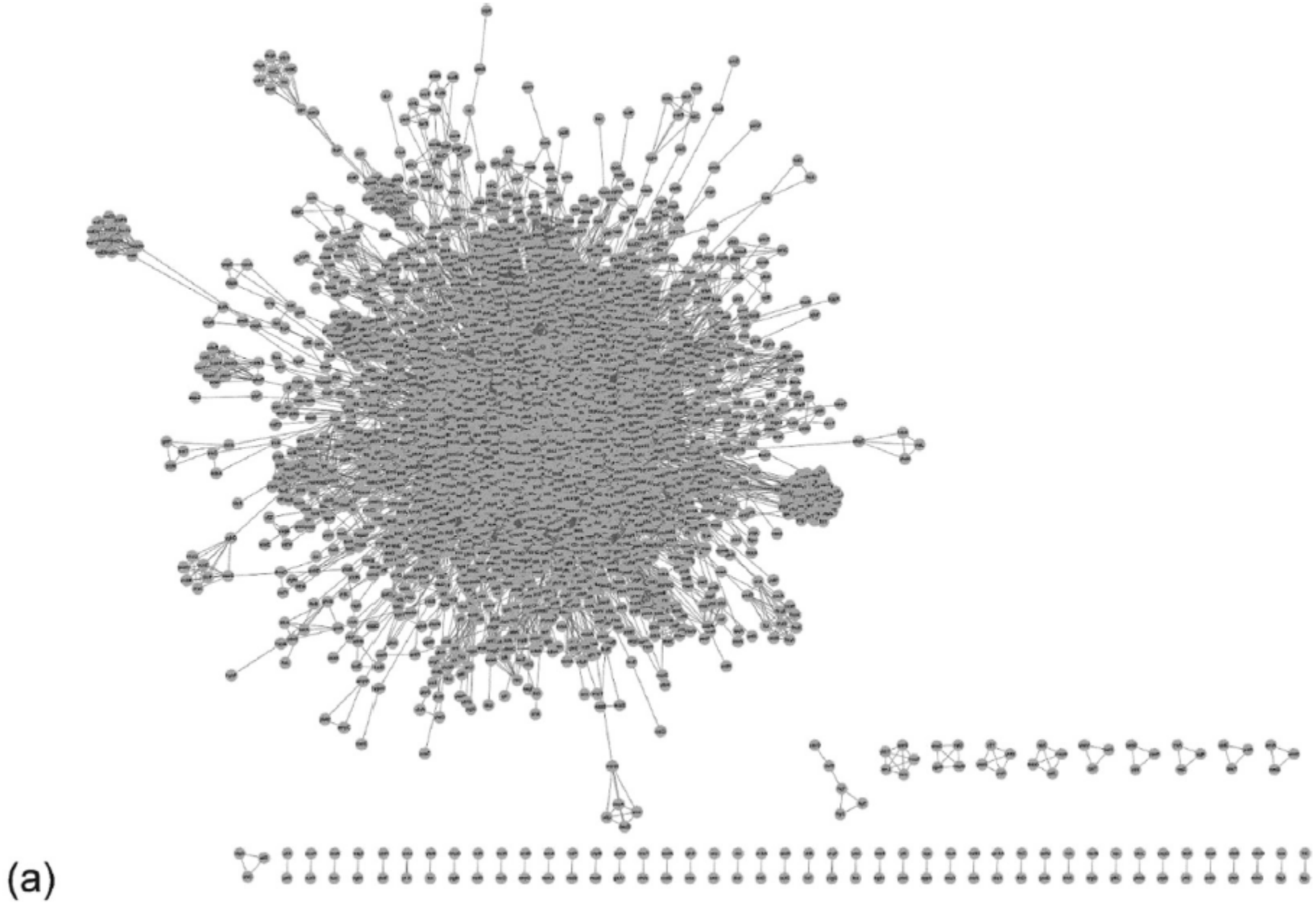
[a] With the repeated PPIs (e.g., A-B and B-A) removed from the 4 HTE datasets respectively, the numbers of PPIs become 6067, 11478, 5993 and 2191 from the original ones 6234, 11511, 5993 and 2234.

[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.

Results-Bayesian Classifier Models Increase PPI Recognition Accuracy of Individual Methods

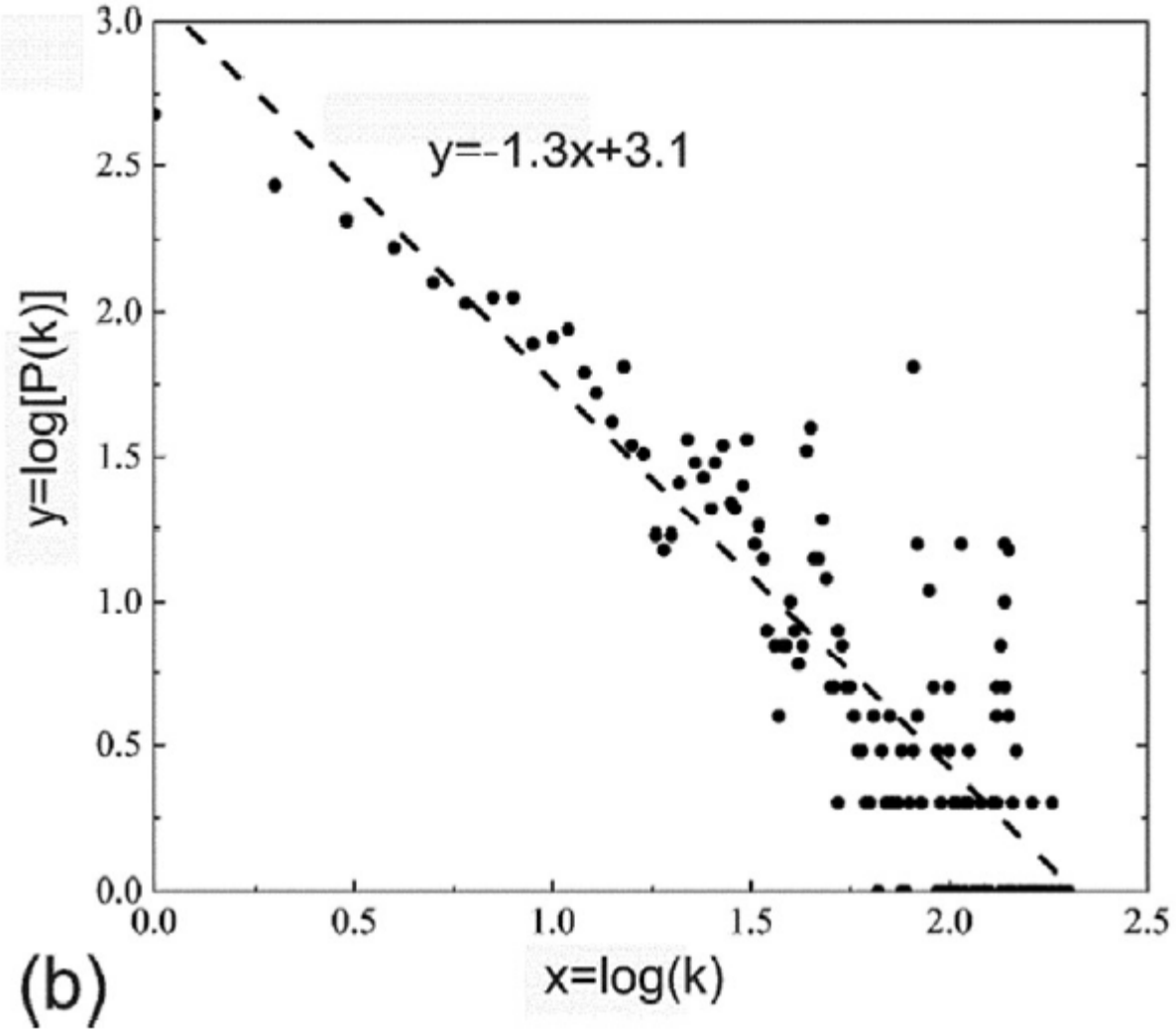


Results-PPI Networks Reveal Dominant Roles of Essential Proteins in *E. coli*

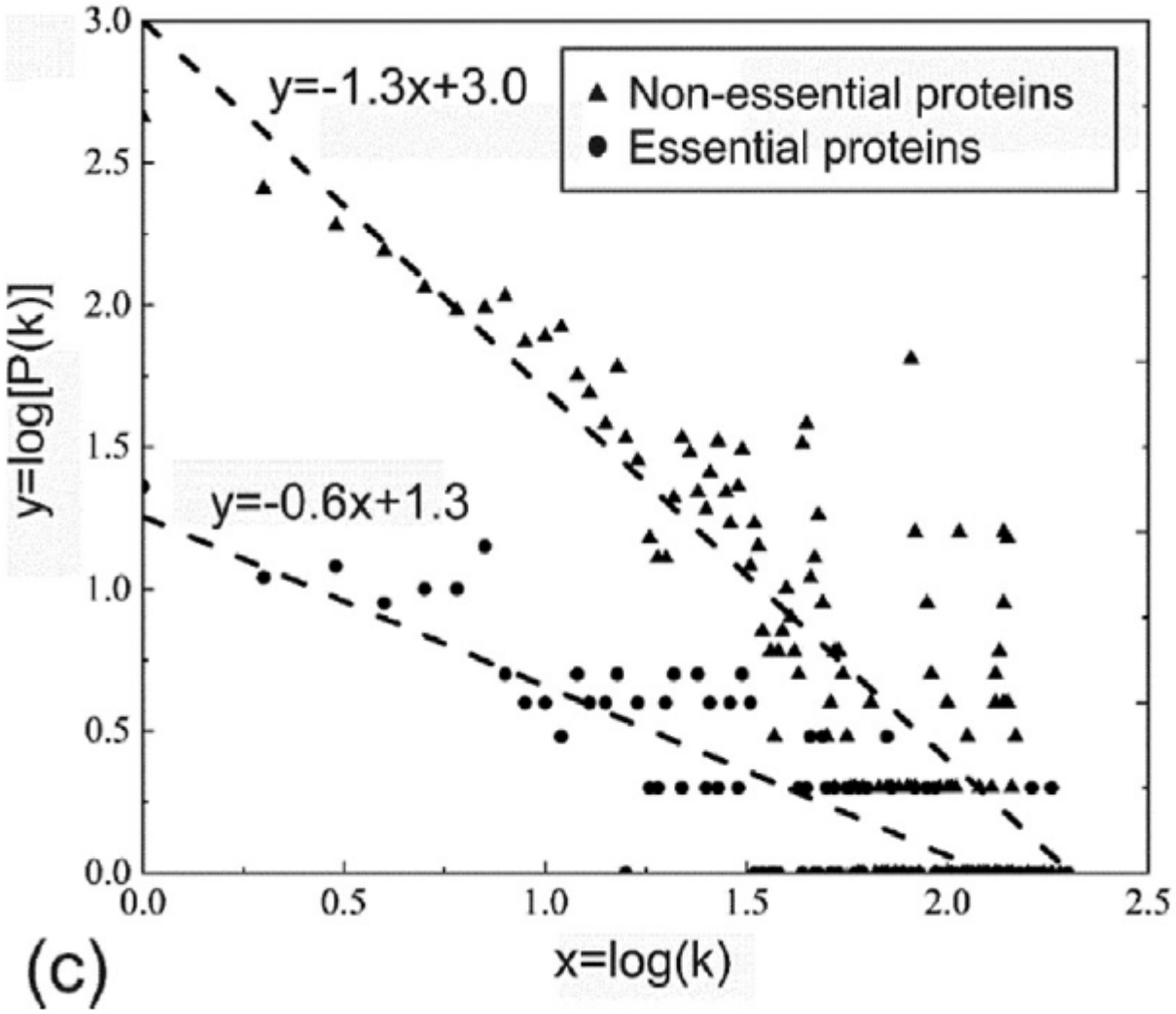


(a)

Results-Node Degree Distribution is Scale Free



Results-Essential Proteins Interact with More Partners than Non-essential ones

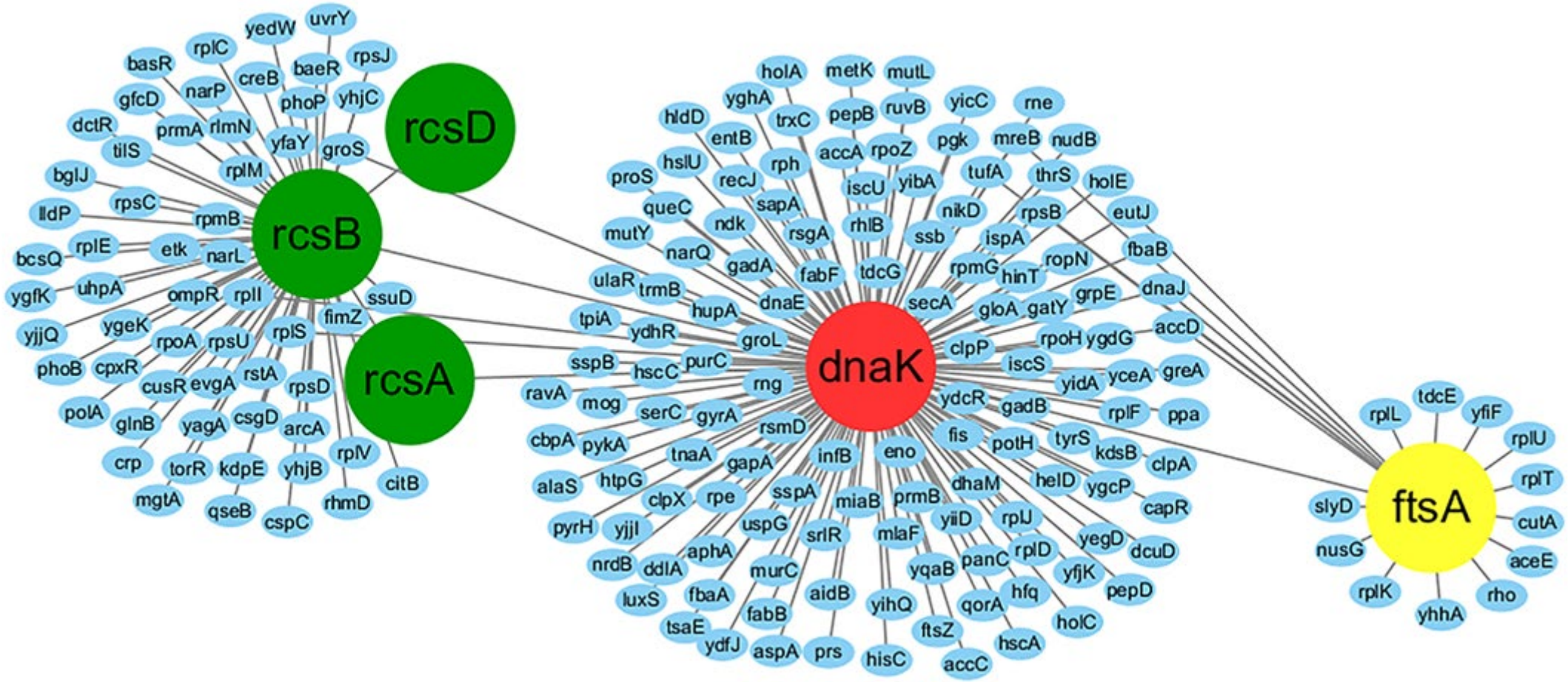


Results-Betweenness Centrality

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

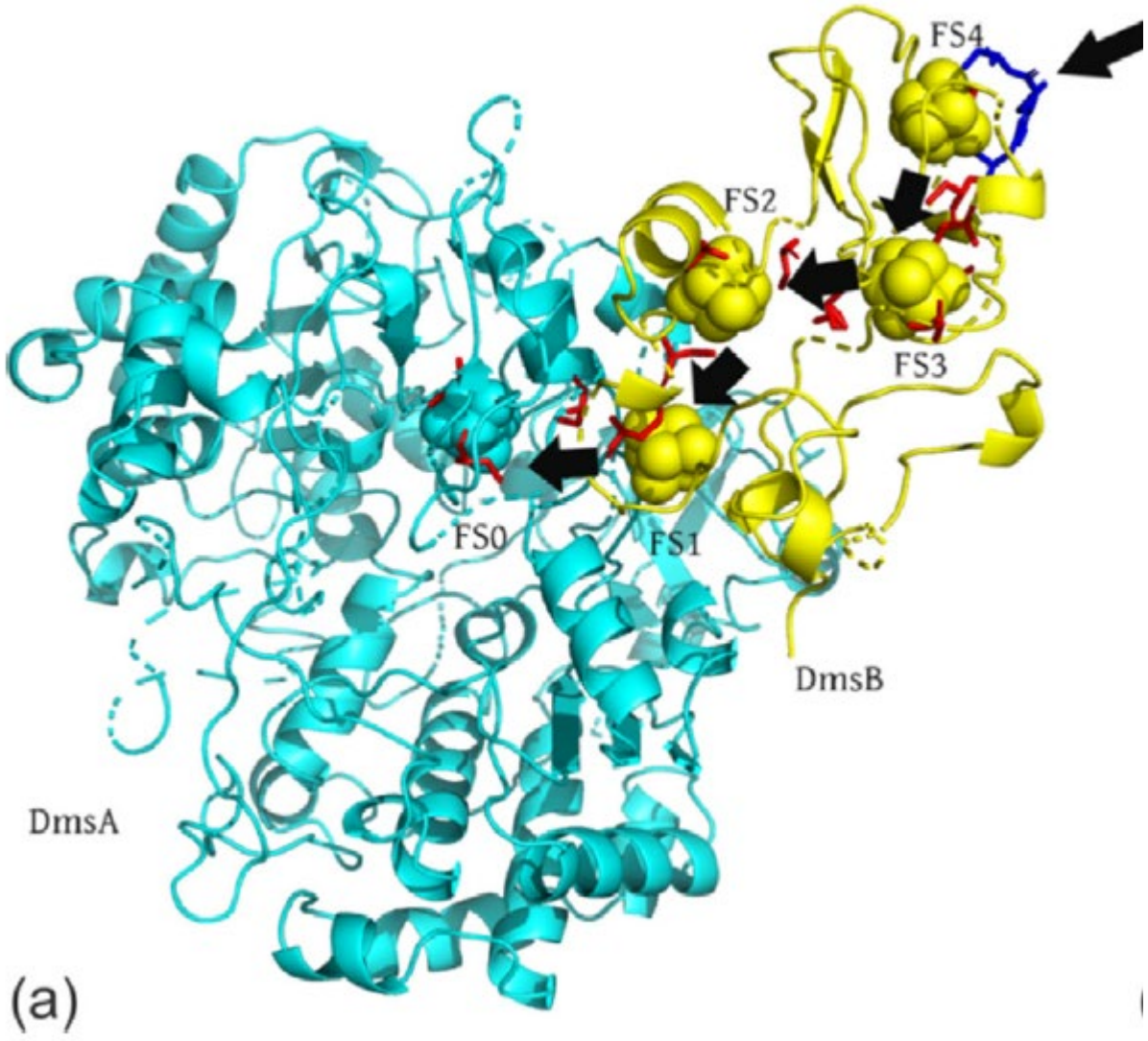
Results-Betweenness Centrality



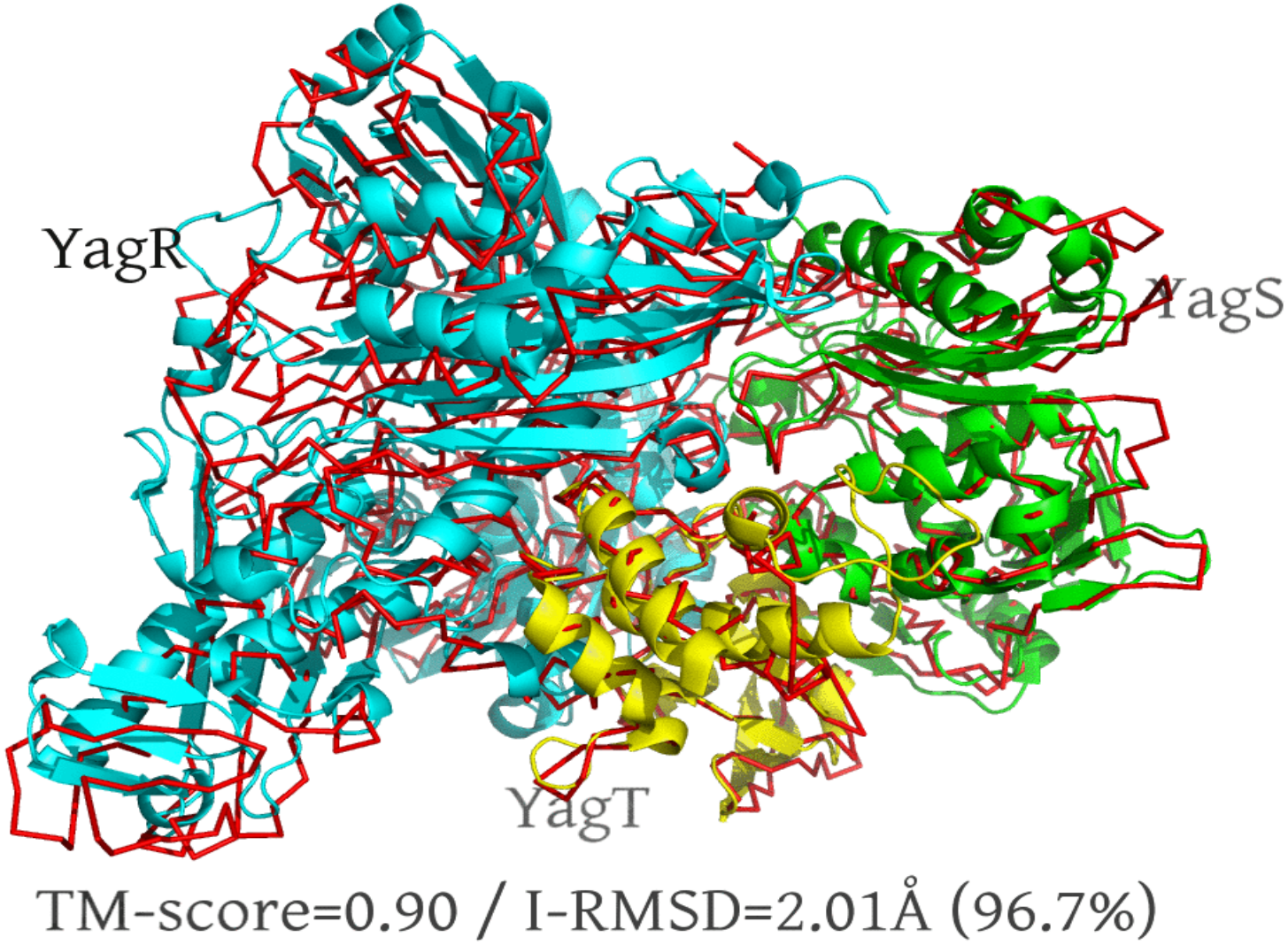
Ref: Zhu X., *et al*., ***Science*** 1996, 272:1606-1614.

Carballès, Fabrice., *et al*., ***Molecular microbiology*** 1999, 34:442-450.

Results-Structural Modeling of Protein Interactome in *E. coli* -DmsAB



Results-Structural Modeling of Protein Interactome in *E. coli* -YagRST



Results-Structural Modeling of Protein Interactome in *E. coli*-

Comparison of Threpp Models on 39 Solved PPI Complexes

- There are in total 39 out of the 35,125 protein-protein complexes whose structures have been experimentally solved in PDB since 2016

- The average TM-score of the Threpp models is 0.73 for these experimental structures

- Threpp achieves average TM-scores of 0.71 and 0.81 for homo- and hetero-dimers

Conclusion

- The Threpp recognizes and models structure of protein-protein interactions in organisms

- Threpp 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, which was found critical to the robustness of genome evolution and the centrality of functionally important proteins that are essential to *E. coli* survival

Conclusion

- Complex structure models were constructed for all predicted *E. coli* PPIs based on Threpp, where 6,771 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

- Two examples from DmsAB and 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

Server-https://zhanglab.ccmb.med.umich.edu/Threpp/-Homepage

THREPP ON-LINE

Multimeric Threading Based Protein-protein Interaction Predictor

Threpp (Multimeric Threading based Protein-protein Interaction Predictor) is a computational algorithm for protein-protein interaction (PPI) prediction. Starting from a pair of query sequences, Threpp first threads them against a non-redundant complex structure library to examine the probabily for them to interact through a naive Bayes classifier model which combines the Threpp threading score and available high-throughput experimental (HTE) data. The quaternary stuctural models of the PPIs are then constructed by reassembling the monomeric threading templates with the identified PPI frameworks. Large-scale benchmark tests showed that Threpp can significantly improve the precision and recall of both HTE and multimeric threading, and therefore reduce the false positive rate for the current PPI modeling approaches. The performance of the current Threpp server is optimal for predicting PPIs in *E. coli*, for which the integrated HTE datasets are constructed. We are still working on extending Threpp to other species by including HTE datasets from non-*E. coli* species.

Threpp On-Line Server (An example of the Threpp output):

- Input your first sequence in FASTA format here: Example input

  Or upload the sequence from your local computer:

  选择文件 未选择任何文件

- Input your second sequence in FASTA format here: Example input

  Or upload the sequence from your local computer:

  选择文件 未选择任何文件

Server-https://zhanglab.ccmb.med.umich.edu/Threpp/-Homepage

- Email: (mandatory, where results will be sent to)
- ID: (optional, your given name of the protein)

Run Threpp | Clear form

Threpp Download

- Click package.zip to download the standalone package of Threpp progam and the template library.
- Click Ecoli3D.zip to download structural models of all PPIs predicted by Threpp in *E. coli* genome, where Ecoli3D.txt contains a summary table of the structural modeling results.
- Click solved_structures.zip to dwonload the model and native structures of 39 protein complexes whose complex structures are experimentally determined after the Threpp structural modeling.
- Click HTE.zip to download the high throughput experiment (HTE) datasets used by Threpp and the script to search the query protein pairs through the HTE dataset.

Reference:

- Weikang Gong, Aysam Guerler, Chengxin Zhang, Elisa Warner, Chunhua Li, Yang Zhang. *Integrating Multimeric Threading With High-throughput Experiments for Structural Interactome of Escherichia coli* . Journal of Molecular Biology, 433: 166944 (2021). [PDF] [Supporting Information]

Server-https://zhanglab.ccmb.med.umich.edu/Threpp/-Example

Threpp results for TPP9

[Click result.zip to download all results on this page]

Input Sequence in FASTA format

>chain A (99 residues) [Download]
MALTKAEMSEYLFDKLGLSKRDAKELVELFFEEIRRALENGEQVKLSGFGNFDLRDKNQR
PGRNPKTGEDIPITARRVVTFRPGQKLKSRVENASPKDE
>chain B (90 residues) [Download]
MNKSQLIDKIAAGADISKAAAGRALDAIIASVTESLKEGDDVALVGFGTFAVKERAARTG
RNPQTGKEITIAAAKVPSFRAGKALKDAVN

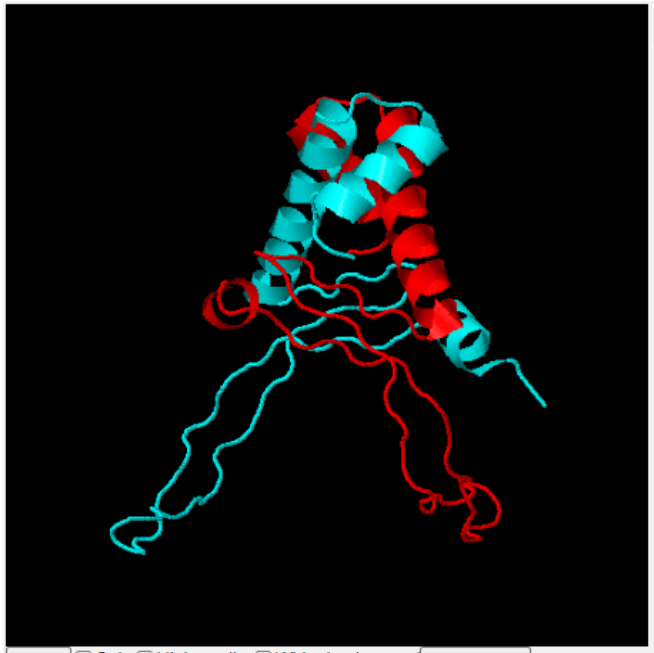
Top 20 dimer threading templates

Rank	PDB hit	BioUnit Num	Chain A	Chain B	Threpp score	Iden	Cov	Norm. Prob.	Download alignment
1	4qjn	1	0	1	16.721	0.446	0.889	100.0	Template1
2	4qjn	1	1	0	16.684	0.439	0.915	100.0	Template2
3	1p51	1	2	3	16.184	0.385	0.963	100.0	Template3
4	2np2	1	2	3	16.155	0.317	0.952	100.0	Template4
5	1p51	1	3	2	16.119	0.385	0.963	100.0	Template5
6	2np2	1	3	2	16.099	0.317	0.952	100.0	Template6
7	1mul	2	0	1	15.655	0.480	0.804	99.9	Template7
8	1mul	2	1	0	15.655	0.480	0.804	99.9	Template8
9	4p3v	1	0	1	15.619	0.632	0.762	99.9	Template9
10	4p3v	1	1	0	15.619	0.632	0.762	99.9	Template10
11	1huu	1	0	1	15.012	0.475	0.847	99.9	Template11
12	1huu	1	1	0	15.012	0.475	0.847	99.9	Template12
13	2o97	1	1	0	14.877	0.447	0.746	99.9	Template13
14	2o97	1	0	1	14.838	0.660	0.746	99.9	Template14
15	4pt4	1	0	1	14.486	0.366	0.968	99.9	Template15
16	4pt4	1	1	0	14.479	0.366	0.968	99.9	Template16
17	1ihf	1	4	3	14.296	0.308	0.963	99.9	Template17
18	1ihf	1	3	4	14.089	0.667	0.952	100.0	Template18
19	3rhi	1	1	0	13.798	0.394	0.873	99.9	Template19
20	3rhi	1	0	1	13.773	0.406	0.899	99.9	Template20

(a) Templates are ranked in descending order of ThreppScore of dimer threading.
(b) *BioUnit Num* is the biological assembly (i.e. biounit) number.
(c) *Chain A* and *Chain B* are the PDB chains in biological assembly files that aligns the first and second query sequence, respectively.
(d) *ThreppScore*, also known as SPRING-score, is a combination of monomeric threading Z-score, interface contact statistical potential, and TM-align match between monomer-to-dimer templates.
(e) *Iden* is the sequence identity of the templates in the threading aligned region with the query sequence.
(f) *Cov* is the coverage of threading alignment. It is equal to the number of aligned residues divided by the length of two query proteins.
(g) *Norm Prob* is the percentage probability of correct dimer template identified.
(h) *Download alignment* provides the 3D coordinates of template aligned regions.
(f) Template residues identical to query sequence are highlighted in color. Upper and lower case letters denote residues from first and second chain, respectively.
(g) The full template table is available here.

Server-https://zhanglab.ccmb.med.umich.edu/Threpp/-Example

Top 20 structure models



Reset ☐ Spin ☐ High quality ☐ White background Save image

Click to view	Rank[a]	Download[b]
◉	1	model1.pdb.gz
○	2	model2.pdb.gz
○	3	model3.pdb.gz
○	4	model4.pdb.gz
○	5	model5.pdb.gz
○	6	model6.pdb.gz
○	7	model7.pdb.gz
○	8	model8.pdb.gz
○	9	model9.pdb.gz
○	10	model10.pdb.gz
○	11	model11.pdb.gz
○	12	model12.pdb.gz
○	13	model13.pdb.gz
○	14	model14.pdb.gz
○	15	model15.pdb.gz
○	16	model16.pdb.gz
○	17	model17.pdb.gz
○	18	model18.pdb.gz
○	19	model19.pdb.gz
○	20	model20.pdb.gz

(a) Threpp models are ranked in the same order as the respective dimer templates, as listed here.
(b) Download structure models.

Combination with high-throughput experiment (HTE)

#	chain A (seqID) [a]	chain A (seqID) [b]	PPI [c]	Dataset source
0	2o97A/1_0_0 (0.253)	2o97A/1_1_7 (0.789)	Y	Dimer theading
1	b1712 (100.0%)	b0440 (100.0%)	Y	Hu et al
2	b1712 (100.0%)	b0440 (100.0%)	N	Rajagopala et al
3	b1712 (100.0%)	b0440 (100.0%)	N	Arifuzzaman et al
4	b1712 (100.0%)	b0440 (100.0%)	Y	Butland et al

Likelihood ratio of interaction = 8.16 ≥ 1.87, the protein pair is predicted to interact.

(a) b-number of chain A in the HTE dataset, and the percentage sequence identity between query chain and the sequence used in HTE.
(b) b-number of chain B in the HTE dataset, and the percentage sequence identity between query chain and the sequence used in HTE.
(c) Whether the protein pair is found to interact in the HTE: Y for interacting pair; N for non-interacting pair; NA for a protein pair not included in the HTE.

[Click result.zip to download all results on this page]

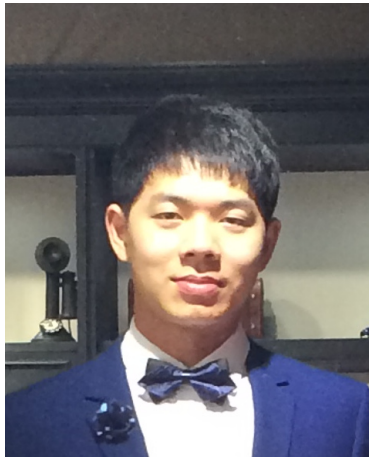
References:
1. Weikang Gong, Aysam Guerler, Chengxin Zhang, Elisa Warner, Chunhua Li, Yang Zhang. *Integrating multimeric threading and high-throughput experiments for structural interactome of Escherichia coli*. submitted (2020).

Thanks to

Prof. Yang Zhang

Prof. Chunhua Li

Dr. Aysam Guerler



Dr. Chengxin Zhang

Thanks for your attention !