RESEARCH ARTICLE

Determining the minimum number of protein-protein interactions required to support known protein complexes

Natsu Nakajima[1]*, **Morihiro Hayashida**[2], **Jesper Jansson**[3], **Osamu Maruyama**[4], **Tatsuya Akutsu**[5]*

1 Institute of Molecular and Cellular Biosciences, The University of Tokyo, 1-1-1, Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan, **2** Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College, 14-4, Nishiikumacho, Matsue, Shimane 690-8518, Japan, **3** Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, **4** Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan, **5** Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

* nakajima@kuicr.kyoto-u.ac.jp (NN); takutsu@kuicr.kyoto-u.ac.jp (TA)

OPEN ACCESS

Citation: Nakajima N, Hayashida M, Jansson J, Maruyama O, Akutsu T (2018) Determining the minimum number of protein-protein interactions required to support known protein complexes. PLoS ONE 13(4): e0195545. https://doi.org/10.1371/journal.pone.0195545

Editor: Byung-Jun Yoon, Texas A&M University College Station, UNITED STATES

Received: September 11, 2017

Accepted: March 23, 2018

Published: April 26, 2018

Data Availability Statement: All relevant data are within the paper, its Supporting Information files., and Dryad ID:doi:10.5061/dryad.8s3682g.

Funding: T. Akutsu was partially supported by JSPS, Japan, (Grants-in-Aid 26240034). M. Hayashida was partially supported by JSPS, Japan (Grants-in-Aid 16K00392). J. Jansson was partially supported by the Hakubi Project at Kyoto University, Japan. O. Maruyama was partially supported by JSPS KAKENHI Grant Number 26330330.

Abstract

The prediction of protein complexes from protein-protein interactions (PPIs) is a well-studied problem in bioinformatics. However, the currently available PPI data is not enough to describe all known protein complexes. In this paper, we express the problem of determining the minimum number of (additional) required protein-protein interactions as a graph theoretic problem under the constraint that each complex constitutes a connected component in a PPI network. For this problem, we develop two computational methods: one is based on integer linear programming (ILPMinPPI) and the other one is based on an existing greedy-type approximation algorithm (GreedyMinPPI) originally developed in the context of communication and social networks. Since the former method is only applicable to datasets of small size, we apply the latter method to a combination of the CYC2008 protein complex dataset and each of eight PPI datasets (STRING, MINT, BioGRID, IntAct, DIP, BIND, WI-PHI, iRefIndex). The results show that the minimum number of additional required PPIs ranges from 51 (STRING) to 964 (BIND), and that even the four best PPI databases, STRING (51), BioGRID (67), WI-PHI (93) and iRefIndex (85), do not include enough PPIs to form all CYC2008 protein complexes. We also demonstrate that the proposed problem framework and our solutions can enhance the prediction accuracy of existing PPI prediction methods. ILPMinPPI can be freely downloaded from http://sunflower.kuicr.kyoto-u.ac.jp/~nakajima/.

Introduction

Identification of protein complexes is important for understanding cellular mechanisms because many proteins express their functions by forming complexes. Since it is difficult to experimentally determine protein complexes, extensive studies have been done on the

Competing interests: The authors have declared that no competing interests exist.

prediction of protein complexes. Among them, many studies focused on the effective use of protein-protein interaction (PPI) data because proteins in a complex physically interact and a large amount of PPI data has been available due to developments of high-throughput experimental techniques [1, 2]. In most of these studies, protein complexes are predicted by identifying non-overlapping or overlapping clusters (i.e., certain types of connected subgraphs) in PPI networks possibly with other biological information. In order to identify clusters in PPI networks, various methods have been developed, including the Markov CLuster (MCL) method [3], the Molecular Complex Detection (MCODE) method [4], the Restricted Neighbourhood Search Clustering (RNSC) method [5], the Repeated Random Walks (RRW) method [6], the Clustering based on Maximal Clique (CMC) method [7], and the Node-Weighted Expansion (NWE) method [8]. However, it was also pointed out that known PPI data suffers from a significant amount of noise in terms of both false positives (spuriously detected interactions) and false negatives (missing interactions) [9]. Therefore, various methods have also been developed that make use of weight and/or reliability of PPIs [7, 9, 10].

When studying and analyzing these protein complex prediction methods, we encounter a fundamental question. Is the current PPI data enough to explain all known protein complexes? If not, how many additional PPIs are required? The main purpose of this paper is to tackle this fundamental question. In order to answer the question, we impose as a minimum requirement that the subgraph induced by the nodes (i.e., proteins) in each complex must be connected. The answer would enable us to deduce that many interactions to detect the associations between proteins forming all known protein complexes are missing from the current PPI data and if the number of the additional PPIs is large, more additional experiments might be necessary. Then, we define the problem of determining the minimum number of additional PPIs required to support known complexes (MinPPI) as: given a set of protein complexes (i.e., a set of sets of proteins) and a set of PPIs (i.e., a set of edges among proteins), find a minimum number of additional PPIs such that the connectivity requirement is satisfied for all given protein complexes. We also define MinPPI0 as the special case of MinPPI in which the set of given PPIs is empty.

Interestingly, the same problem has been studied in the analysis of communication and social networks under the name of the Network Construction problem and a greedy, polynomial-time approximation algorithm for it has been proposed [11]. This fact suggests that our question is a natural and general one. We modify this greedy-type algorithm so that we can start with some known PPI data, and the resulting algorithm is called GreedyMinPPI. We also develop a novel integer linear programming (ILP)-based method called ILPMinPPI that gives an exact solution.

In this paper, we compare two methods using moderate size synthetic data. Then, we apply GreedyMinPPI to three large-scale real protein complex datasets, CYC2008 [12], MIPS [13], and Aloy *et al.*'s set [1, 14] without any known PPIs, to estimate the minimum number of PPIs, and to pairs of CYC2008 and eight PPI datasets (STRING [15], MINT [16], BioGRID [17], DIP [18], BIND [19], WI-PHI [20], IntAct [21, 22], and iRefIndex [23]) to estimate the minimum number of additional PPIs.

However, as mentioned above, known PPI data suffers from a significant amount of noise. In particular, there are a large amount of missing interactions [24, 25]. Therefore, many methods have been proposed to predict PPIs from protein sequences, protein structures, and/or other biological data [26–30]. Since GreedyMinPPI outputs also unknown PPIs, it might be helpful to enhance existing PPI prediction methods by GreedyMinPPI. In order to assess the usefulness of this idea, we examine a combination of GreedyMinPPI and each of four state-of-the-art prediction methods for weighted PPIs, Struct2Net [26], ENTS [27], PIP [28], and iWRAP [29], using four PPI datasets extracted from STRING [15], MINT [16], WI-PHI [20], and IntAct [21]. Since the four databases contain interactions with confidence score based on distinct sources of evidence, we regard these PPI databases as reliable.

Problem definition and motivation

The two problems *gMinPPI* and *gMinPPI0* studied in this paper are defined formally as graph theoretic problems as follows. The input to gMinPPI is an undirected graph $G = (V, E)$, where V is a set of vertices and E is a set of edges, along with a collection $\mathcal{C}$ of subsets of V, and the output is an undirected graph $G' = (V, E \cup E')$, where E' is a set of additional edges, such that the subgraph of G' induced by each $C_i \in \mathcal{C}$ is connected and the value of $|E'|$ is minimized. gMinPPI0 is the special case of gMinPPI where $E = \emptyset$. Given any instance of gMinPPI or gMinPPI0, we use the notation $n = |V|$ and $m = |\mathcal{C}|$. When we apply gMinPPI and gMinPPI0 to the protein complex data, which are also defined as MinPPI and MinPPI0, which means that MinPPI0 is the special case of MinPPI where the set of given PPIs is empty.

In our application, the elements in V, E, and $\mathcal{C}$ represent proteins, known protein-protein interactions (PPIs), and protein complexes, respectively. The elements in E' correspond to hypothetical PPIs whose existence would guarantee that each protein complex is internally connected. See Fig 1 for two examples. Hence the value of $|E'|$ is a lower bound on the number of additional PPIs needed to support the given protein complexes. The motivation of this study is to solve MinPPI for some particular data sets from the literature and investigate their values of $|E'|$; if $|E'|$ for some data set is large, this suggests that many interactions between proteins are missing from the database and that additional experiments might be necessary to complete the picture.

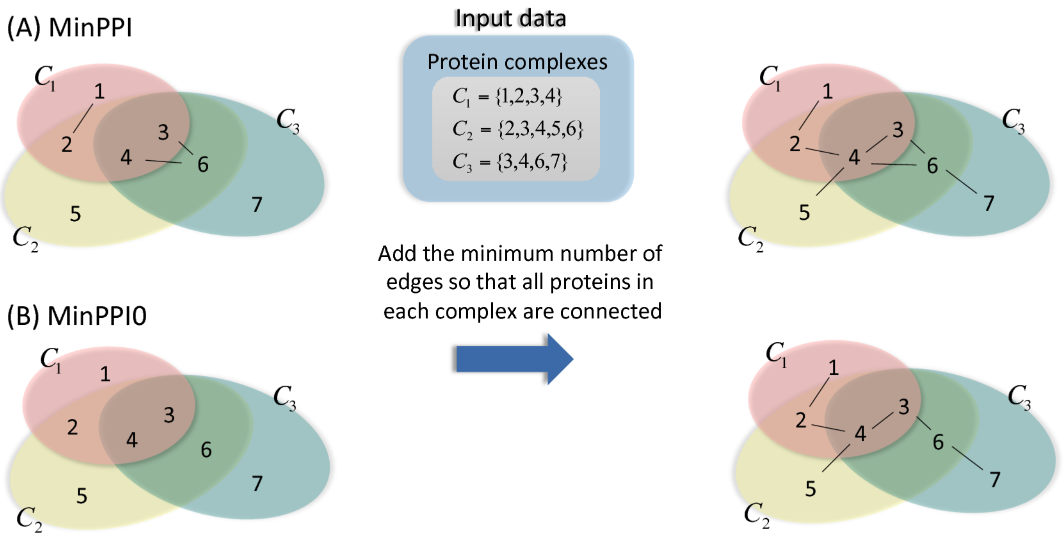


Fig 1. A description of MinPPI and MinPPI0. (A) An example of MinPPI. MinPPI corresponds to determining the minimum number of additional interactions starting with known PPI (a set of edges among proteins) data. For example, if a family of protein complexes $\mathcal{C}$ consisting of C_1, C_2 and C_3 where the total number of proteins is 7 and known PPI data are given as an input data, the objective is to find the minimum number of additional PPIs required to describe interactions in all given protein complexes such that all proteins belonging to each complex C_i ($i = 1, 2, 3$) must be connected. Since an initial PPI network has 3 edges and one protein complex overlaps with other protein complexes, the resulting graph contains 7 edges. (B) An example of MinPPI0. MinPPI0 corresponds to the determination beginning with the set of known PPIs is empty. For example, if a family of protein complexes $\mathcal{C}$ composed of C_1, C_2 and C_3 and known PPI data with no edge are given as an input data, the objective is to find the connected graph which has the minimum number of PPIs such that all proteins belonging to each complex C_i ($i = 1, 2, 3$) must be connected. Since a given PPI network has no edges and one protein complex overlaps with others, the number of additional edges is 6.

https://doi.org/10.1371/journal.pone.0195545.g001

Previous work

In the literature, gMinPPI0 is equivalent to the *Minimum Topic-Connected Overlay* problem [31] and the *Uniform Cost Network Construction* problem [11]. The more general *Network Construction* problem introduced in [11] is an extension of gMinPPI0 in which a non-negative cost of an edge c_e for each $e \in V \times V$ is allowed and aims to minimize $\sum_{e \in E'} c_e$. Restricted versions of the Network Construction problem in which the output graph G' is required to be a tree or a star were studied in [32] and [33] (note that for certain inputs, no such G' exists). Chockler *et al.* [31] proved that gMinPPI0 is NP-hard to approximate within a constant factor. They also gave a polynomial-time, greedy approximation algorithm for gMinPPI0 which starts with $E' = \emptyset$ and inserts one suitably chosen edge at a time into E' until each $C_i \in \mathcal{C}$ induces a single connected component in the resulting graph G', and showed that its approximation ratio is logarithmic in $\sum_{C_i \in \mathcal{C}} |C_i|$. Angluin *et al.* [11] recently strengthened these results by: (i) proving that the Network Construction problem and equivalently, gMinPPI0 is NP-hard to approximate within a factor of $\Omega(\log n)$; and (ii) extending the greedy approximation algorithm for gMinPPI0 and refining its mathematical analysis to obtain a polynomial-time $O(\log m)$-approximation algorithm for the Network Construction problem.

Materials and methods

Integer linear programming formulation

We propose an exact method called ILPMinPPI, for the problem of predicting PPIs beginning with the set of known PPIs is empty (corresponding MinPPI0).

MinPPI0 can be formulated using the following integer linear programming (ILP). For each complex C_p, we add the following constraints using different 0-1 variables for different C_p,

$$\begin{aligned} x_{ij}^{p,T_p} &= 1 \qquad \text{for all } i<j, \\ x_{ij}^{p,0} &\le e_{ij} \qquad \text{for all } i<j, \\ x_{ij}^{p,t+1} &\le x_{ij}^{p,t} + \sum_{k \notin \{i,j\}} x_{ij,k}^{p,t+1} \qquad \text{for all } i<j \text{ and } k \notin \{i,j\}, \\ x_{ij,k}^{p,t+1} &\le \frac{1}{2}\left(x_{ik}^{p,t} + x_{kj}^{p,t}\right) \qquad \text{for all } i<j \text{ and } k \notin \{i,j\}, \end{aligned} \tag{1}$$

where $i, j, k \in C_p$, $x_{ij}^{p,t} = x_{ji}^{p,t}$, $x_{ij,k}^{p,t} = x_{ji,k}^{p,t}$, and $T_p = \lceil \log|C_p| \rceil$. e_{ij} is a variable which indicates whether an edge exists between proteins i and j. If $e_{ij} = 1$, an interaction exists between proteins i and j. $x_{ij}^{p,t}$ reflects whether proteins i and j are connected after t ($0 \le t \le T_p$) steps and $x_{ij,k}^{p,t}$ reflects whether proteins i and j are connected through protein k after t steps. The third constraint means that if i and k are connected and k and j are connected, i and j are also connected. Hence, after enough steps T_p, proteins i and j should be connected, which means that x_{ij}^{p,T_p} takes 1. Since the connectedness of each complex is checked by using the doubling technique, it is enough to repeat this process at most $O(\log|C_p|)$ steps. Hence, these constraints state that the subgraph of $G(V, E)$ induced by nodes (proteins) in each C_p must be connected, which guarantees that each protein complex is internally connected under the condition that edges with $e_{ij} = 1$ can only be used. Then, the required ILP is formulated as

$$\text{minimize} \quad \sum_{1 \le i < j \le n} e_{ij}, \tag{2}$$

with constraints for all C_ps.

Approximation algorithm

The exact method ILPMinPPI requires exponential time and can thus only be applied to small datasets in practice. To deal with large-scale datasets, we now extend the approximation algorithm of [11] to MinPPI. The resulting method will be referred to as GreedyMinPPI. It enables us to detect the interactions among proteins with most highly overlapping and most of these interactions would be expected to be more reliable. This means that in GreedyMinPPI, the reliability or confidence score of PPIs are assigned in the order in which interactions are detected.

An $O(\log m)$-approximation algorithm for gMinPPI. Angluin *et al.* [11] proved that the Network Construction problem can be approximated within a ratio of $O(\log m)$ in polynomial time. We now show that this yields a polynomial-time, $O(\log m)$-approximation for gMinPPI.

Theorem 1. *gMinPPI can be approximated within a ratio of* $O(\log m)$ *in polynomial time.*

Proof. Let $(G; \mathcal{C})$ be any given instance of gMinPPI, where $G = (V, E)$. Create an instance $(V; \mathcal{C}; c)$ of the Network Construction problem by defining a cost function c on pairs of vertices as follows: for every $u, v \in V$ with $u \neq v$, if $\{u, v\} \notin E$ then $c(\{u, v\}) = 1$ and if $\{u, v\} \in E$ then $c(\{u, v\}) = \epsilon$, where ϵ is any constant satisfying $0 < \epsilon < 2^{-n}$ (i.e., $n = |V|$). Next, apply Angluin *et al.*'s $O(\log m)$-approximation algorithm [11] to $(V; \mathcal{C}; c)$, and denote the obtained graph by (V, E^*). Finally, output the graph $(V, E \cup E^*)$ as the approximate solution to gMinPPI.

Clearly, the running time is polynomial. To bound the approximation ratio, let OPT be any optimal gMinPPI-solution to the given $(G; \mathcal{C})$. Denote the total number of edges in OPT by $|E| + x$. Note that $(G; \mathcal{C})$ admits a solution to gMinPPI with $|E| + x$ edges if and only if $(V; \mathcal{C}; c)$ admits a solution to the Network Construction problem whose cost (i.e., sum of costs of all edges) is in the interval $[x, x + \epsilon \cdot |E|]$. Since (V, E^*) is an $O(\log m)$-approximate solution for the latter, the output $(V, E \cup E^*)$ contains at most $|E| + (x + 1) \cdot O(\log m) = |E| + x \cdot O(\log m)$ edges and is therefore an $O(\log m)$-approximation of OPT.

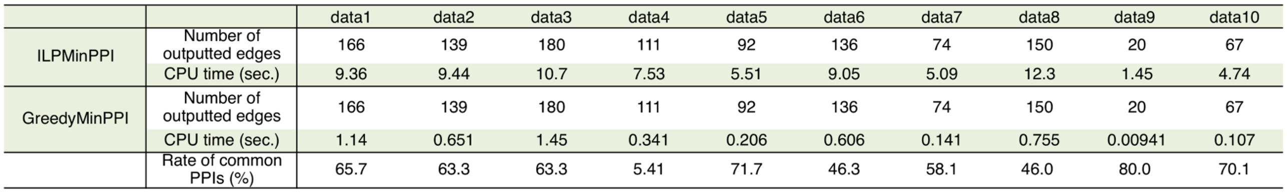
Results

Computer environment

We evaluated the performance of both ILPMinPPI and GreedyMinPPI using both synthetic data and real protein-protein interaction data. An integer programming solver, CPLEX Interactive Optimizer 12.4.0.0 (http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/) was used to compute an exact optimal solution and 'Epi' library (version 1.1.67) in R (version 3.2.1) was used to plot the ROC curve. The implementation of ILPMinPPI and GreedyMinPPI was done by C/C++ code. The C/C++ programs are used to generate linear programs which are then fed to CPLEX. All experiments were performed on a PC with Intel Core i7-2600 CPU (3.40 GHz) with 7.7 GB RAM running under the Fedora 21 with Linux kernel 3.14.2 and with Xeon E5-2667 CPU (3.30GHz×8) with 62.9 GiB memory running under the Mint 17.1 Cinnamon with Linux kernel 3.13.0-37-generic. We used the same CPU to compare the CPU time in each experiment. ILPMinPPI can be freely downloaded from http://sunflower.kuicr.kyoto-u.ac.jp/~nakajima/.

Results on ILPMinPPI using synthetic data and real data

Comparison of ILPMinPPI and GreedyMinPPI using synthetic data. At the beginning, we compared GreedyMinPPI with ILPMinPPI using two types of synthetic datasets. We randomly built two datasets (syndata 1, syndata 2), each of which is composed of 10 artificial protein complex datasets (data1–data10), where the maximum number of total proteins, complexes and proteins within a complex are 10, 20 and 5 for syndata 1, and both 100 and 4 for syndata 2, respectively. Furthermore, in real world datasets, since the proteins apparently



		data1	data2	data3	data4	data5	data6	data7	data8	data9	data10
ILPMinPPI	Number of outputted edges	166	139	180	111	92	136	74	150	20	67
	CPU time (sec.)	9.36	9.44	10.7	7.53	5.51	9.05	5.09	12.3	1.45	4.74
GreedyMinPPI	Number of outputted edges	166	139	180	111	92	136	74	150	20	67
	CPU time (sec.)	1.14	0.651	1.45	0.341	0.206	0.606	0.141	0.755	0.00941	0.107
	Rate of common PPIs (%)	65.7	63.3	63.3	5.41	71.7	46.3	58.1	46.0	80.0	70.1

Fig 2. Performance evaluation of ILPMinPPI and GreedyMinPPI using synthetic data (syndata 2).

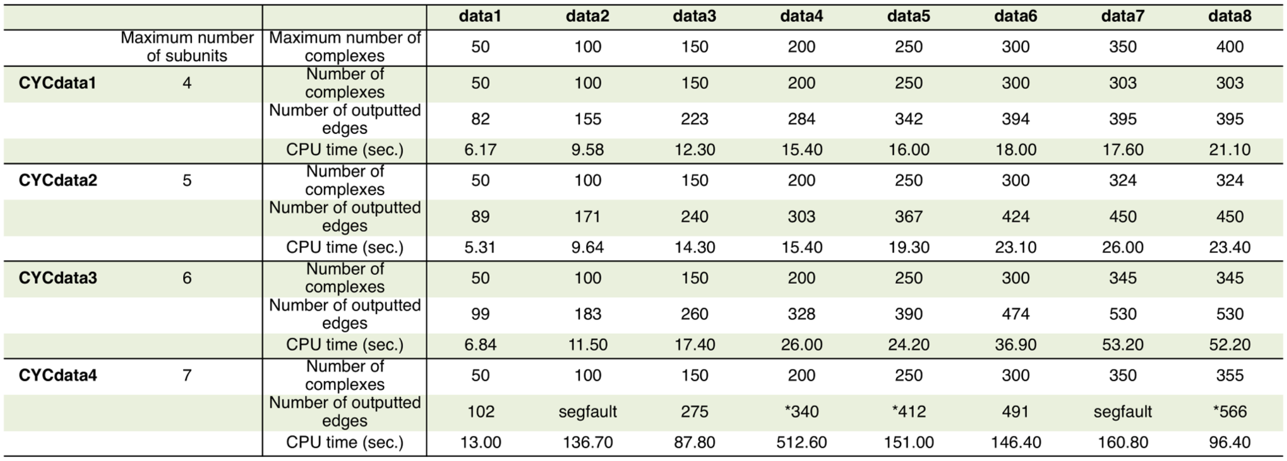
https://doi.org/10.1371/journal.pone.0195545.g002

outnumbered the protein complexes, we also performed the simulation experiments in three types of practical setting (see details in Table A1(a) in S1 File). For each protein complex dataset, the corresponding ILP formulation is written in the LP-format required by CPLEX. Then, we evaluated and compared the performance of ILPMinPPI with that of GreedyMinPPI by measuring the total number of interactions (edges) and the CPU time (real time) as shown in Fig 2 and Table A1 in S1 File.

For example, the results indicate that ILPMinPPI outputted 166 edges requiring 9.36 sec using data1 of syndata 2 shown in Fig 2. For syndata 1, since the objective values provided by the two methods are not all the same, we count the number of common PPIs. Table A1(b) in S1 File shows that two methods do not always output the same objective values but the optimal values are relatively close. On the other hand, both ILPMinPPI and GreedyMinPPI provide the same objective values in all cases of syndata 2, 3 and 4, however, the rate of the common PPIs between the two methods is not so high.

Although ILPMinPPI outputs the same objective value for the same complex datasets, this property is not guaranteed for GreedyMinPPI because the objective value may depend on the ordering of the input data. In order to validate whether or not GreedyMinPPI outputs the same prediction results among runs with the same input data but different orderings, we also examined GreedyMinPPI on MinPPI0 using randomly generated datasets, where the maximum numbers of proteins and complexes were 1600 and 400, respectively, and the maximum number of proteins within one complex was 5. The configuration of each dataset was changed by shuffling the complexes and the subunits each 10 times. Therefore, 100 configurations were examined for each dataset. The results are summarized in Table A2(a) in S1 File. We found that all approximate solutions were exactly the same including the addition order of edges. It is reasonable because GreedyMinPPI repeatedly detects the protein pair with the highest overlap by using the confidence score and thus it is not plausible that multiple pairs have the same confidence score. However, there existed one exceptional case when GreedyMinPPI was applied to real protein complex datasets (see Table A2(b) in S1 File). For the STRING dataset, 51 protein pairs were identified in 91 trials whereas 139 protein pairs were identified in 9 trials. However, the number of bad cases (i.e., 139 protein pairs) is small. Therefore, it is expected that even in non-preferred cases, we can obtain a reasonably good solution by examining multiple configurations and taking the best solution.

Additionally, the CPU time of GreedyMinPPI is much less than that of ILPMinPPI only if the dataset is small. It must be noted that ILPMinPPI and GreedyMinPPI could possibly work when one protein complex consists of relatively few proteins, because the CPU time depends on the maximum number of proteins as described in Table A3 in S1 File. Actually, ILPMinPPI could not work on syndata 5. GreedyMinPPI provides optimal or near-optimal solutions while reducing the CPU time and it might therefore be effective for large datasets. This result also suggests that ILPMinPPI still has room for improvement by extending the ILP formulation to avoid combinatorial explosion.



			data1	data2	data3	data4	data5	data6	data7	data8
	Maximum number of subunits	Maximum number of complexes	50	100	150	200	250	300	350	400
CYCdata1	4	Number of complexes	50	100	150	200	250	300	303	303
		Number of outputted edges	82	155	223	284	342	394	395	395
		CPU time (sec.)	6.17	9.58	12.30	15.40	16.00	18.00	17.60	21.10
CYCdata2	5	Number of complexes	50	100	150	200	250	300	324	324
		Number of outputted edges	89	171	240	303	367	424	450	450
		CPU time (sec.)	5.31	9.64	14.30	15.40	19.30	23.10	26.00	23.40
CYCdata3	6	Number of complexes	50	100	150	200	250	300	345	345
		Number of outputted edges	99	183	260	328	390	474	530	530
		CPU time (sec.)	6.84	11.50	17.40	26.00	24.20	36.90	53.20	52.20
CYCdata4	7	Number of complexes	50	100	150	200	250	300	350	355
		Number of outputted edges	102	segfault	275	*340	*412	491	segfault	*566
		CPU time (sec.)	13.00	136.70	87.80	512.60	151.00	146.40	160.80	96.40

Fig 3. Performance comparison of ILPMinPPI with four protein complex datasets. Summary of four real protein complex datasets from CYC2008 and performance comparison of ILPMinPPI with these datasets. For example, 'data7' of CYCdata1 is composed of 303 complexes (see 'Number of complexes'), where the number of subunits is at most 4 (see 'Maximum number of subunits'). When using this data, ILPMinPPI outputs 395 edges and requires 17.6 seconds. A 'segfault' refers to the segmentation fault which occurs when a program accesses an invalid memory address and the number of outputted edges with asterisk(*) means that CPLEX outputs not an optimal solution but a feasible solution because of exceeding the memory limit.

https://doi.org/10.1371/journal.pone.0195545.g003

Results using protein complex dataset. Although ILPMinPPI is not efficient for large data in general, it would be insightful to examine whether it can provide optimal solutions for real datasets. We thus applied ILPMinPPI to MinPPI0 using CYC2008 protein complexes as a benchmark set reported by Pu *et al.* [12]. CYC2008 consists of 408 protein complexes, some of which are composed of dozens of distinct subunits shared between the different complexes.

To examine how the CPU time is affected by the size of a dataset and the number of subunits, we prepare four datasets derived from the benchmark complexes according to the number of proteins within a complex. Unfortunately, the results on synthetic data indicate that ILPMinPPI can be applicable for the limited datasets including complexes composed of only a few subunits. Therefore, we randomly constructed the datasets that are limited to the maximum number of proteins within a complex as ranging from 4 to 10 and then examine how the performance of ILPMinPPI evolves as the maximum number of complexes increases ranging from 50 to 400 with 50 intervals in order from the beginning of CYC2008. For example, as shown in Fig 3, CYCdata1 consists of 8 components (data1–data8), each of which is composed of a large set of protein complexes formed by at most 4 proteins.

Fig 3 indicates that there may be some correlation between the CPU time and the maximum number of subunits. If a complex is formed by at most 6 subunits, ILPMinPPI requires much less or slowly increasing CPU time, regardless of the size of the complex. Compared with the performance on synthetic data, almost the same CPU time is required for this method when using the syndata 2 shown in Table A1(c) in S1 File. In contrast, in case that the number of subunits is more than 7, it cannot provide an optimal solution by solving the MIP problem on CPLEX due to the high memory consumption. Indeed, MinPPI0 on these datasets almost always lead to exponential increase in CPU time or segmentation faults, some of which are not listed in Fig 3 if the complex is composed of 8–10 subunits. Therefore, these results reveal that the computational complexity of ILPMinPPI increases exponentially because of depending not on the total number of complexes but on the maximum number of subunits.

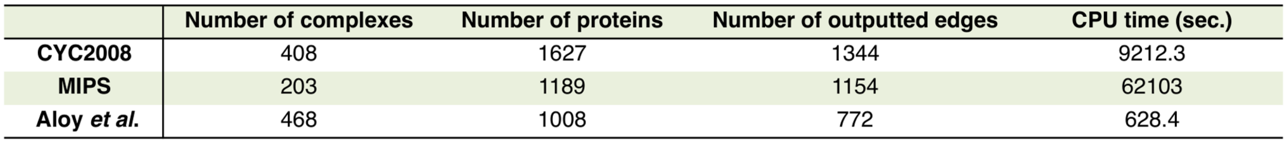
In this experiment, an optimal solution can be provided within a reasonable CPU time only if the relatively small number of subunits are included in the complex, note however that, in reality, there also even exist several complexes formed by more than 10 subunits in CYC2008 and more than 30 subunits in MIPS [13] as summarized in Table A3(b) in S1 File. The current ILPMinPPI has room for improvement to reduce the computational costs. One possibility is that there needs to be a reduction of the size of the MIP problem by omitting the unnecessary constraints or to increase the maximum memory usage.

Results on GreedyMinPPI using real data

The number of PPI for three protein complex datasets. We only tested GreedyMinPPI to solve MinPPI0 using real large-scale protein complex datasets because the current ILPMinPPI is not suitable for application to large-scale datasets (it is only applicable to datasets of small size consisting of 100-150 nodes). We used a dataset derived from CYC2008 [12] which consists of 408 protein complexes involving 1627 proteins in *S. cerevisiae*, and also used two protein complex datasets obtained from MIPS [13] which was publicly available from the study of [34] and Aloy *et al.* [1, 14].

We applied GreedyMinPPI to three protein complex datasets and evaluated the performance in terms of the total number of outputted PPIs and the execution time (CPU time). To test the correlation between the CPU time and the protein complex formation, we counted the number of complexes consisting of p_n proteins. The results are presented in Fig 4 and Table A3 in S1 File. It is observed that the number of added edges is less than the number of proteins in all cases. It is reasonable because three protein complex datasets may consist of the disjoint sets of protein complexes which are a set of unions of proteins (overlapping proteins). For example, MIPS dataset is composed of a small number of disjoint sets of complexes in which a lot of proteins overlap and this may be related to the increased CPU usage as described below. In contrast, Aloy *et al.*'s dataset consists of a large number of disjoint sets of complexes that are a small number of overlapping complexes, despite containing lots of complexes. Similarly, as you can see in Fig 4 and Table A3(b) in S1 File, the increase on required CPU time may also be caused by the formation of protein complexes, which means that it may be caused by the existence of some protein complexes that are composed of a lot of proteins and overlap with other complexes. For example, the MIPS dataset consists of many complexes consisting of more than 10 proteins (p_n>10), on the other hand, Aloy *et al.*'s dataset contains 98% complexes consisting of a lower number of proteins (p_n≤10). Certainly, it is observed that the computation with MIPS dataset requires a lot of CPU time despite including a small number of complexes among three datasets, but it requires much less CPU time for the prediction with Aloy *et al.*'s dataset. Therefore, these results suggest that the CPU utilization depends not only on the number of complexes or proteins but also on the total number of proteins involved in one protein complex.

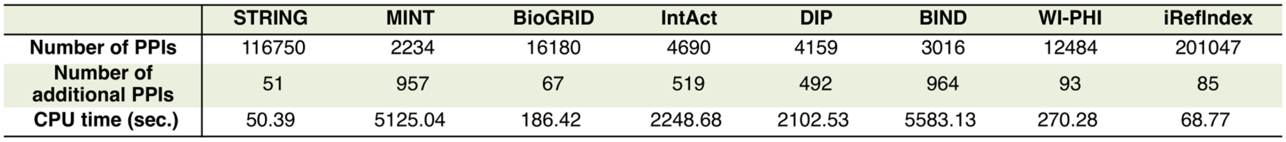
Results using known PPI datasets. We examined GreedyMinPPI for MinPPI from known PPIs obtained from eight protein interaction databases such as STRING [15], MINT [16], BioGRID [17, 34], DIP [18], BIND [19, 35], WI-PHI [20], IntAct [21] and iRefIndex [23, 36], and using protein complexes in CYC2008 dataset. Although the same PPIs should be contained in different databases based on the different experimental techniques, known PPIs are limited to the extracted PPIs composed of 1627 proteins formed of CYC2008 protein



	Number of complexes	Number of proteins	Number of outputted edges	CPU time (sec.)
CYC2008	408	1627	1344	9212.3
MIPS	203	1189	1154	62103
Aloy *et al.*	468	1008	772	628.4

Fig 4. Results with three protein complex datasets.

https://doi.org/10.1371/journal.pone.0195545.g004



	STRING	MINT	BioGRID	IntAct	DIP	BIND	WI-PHI	iRefIndex
Number of PPIs	116750	2234	16180	4690	4159	3016	12484	201047
Number of additional PPIs	51	957	67	519	492	964	93	85
CPU time (sec.)	50.39	5125.04	186.42	2248.68	2102.53	5583.13	270.28	68.77

Fig 5. Summary of eight databases and results on MinPPI by GreedyMinPPI.

https://doi.org/10.1371/journal.pone.0195545.g005

complexes stored in these databases. The performance of GreedyMinPPI was evaluated by measuring the total number of additional interactions and the CPU time used to solve MinPPI, as shown in Fig 5.

It is shown that the minimum number of additional PPIs ranges from at least 51 (STRING) to 964 (BIND). In particular, in the case of BioGRID database, although it contains 16180 interactions, it is observed that 67 interactions are missing to support all interactions in the CYC2008. The results also mean that even the four largest PPI databases, STRING (51) in Table A4(a) in S1 File, BioGRID (67) in Table A7(a) in S1 File, WI-PHI (93) in Table A10(a) in S1 File and iRefIndex (85) in Table A13(a) in S1 File, do not have enough PPIs to form all CYC2008 protein complexes. For example, it is seen from Table A4 in S1 File that Dcs1p/Dcs2 heterodimer and Rad17p/Ddc1p/Mec3p complexes in CYC2008 are not fully covered by PPIs in STRING (Table A4(b) in S1 File) and GreedyMinPPI identifies 51 additional PPI pairs for supporting all CYC2008 complexes (Table A4(a) in S1 File). Table A5 in S1 File also presents the additional protein pairs and the complexes including each pair.

Furthermore, we analyzed the biological significance of additional protein pairs identified for four databases, STRING, BioGRID, WI-PHI and iRefIndex, using the PANTHER system (http://www.pantherdb.org/), which provides the classification of proteins and their genes according to family, subfamily, biological process, cellular component and molecular function. To evaluate the frequency distribution of the assigned categories, we counted the frequency in each category (see Tables A6, A9, A12 and A15 in S1 File). As for the molecular function categories, the top five across four databases are protein binding, oxidoreductase activity, RNA binding, pyrophosphatase activity and structural constituent of ribosome. However, from the PANTHER analysis, we could not observe a clear trend on the characteristic biological function for the additional proteins, which suggests that various types of complexes are not fully covered by the PPIs currently in the four databases.

In order to examine the plausibility of the predicted interactions, we performed in silico experiments. Among various excellent tools developed for prediction of protein interactions [37, 38], we employed PSOPIA [39] because it needs only sequence information, is easy to use, and is reported to have good prediction performance [39]. PSOPIA predicts the interaction between two protein sequences based on information from known homologous PPIs using Averaged One-Dependence Estimators. The interaction of two protein pairs is estimated by calculating the three features, sequence similarities (F_{Seq}), statistical propensities (F_{Dom}) and a sum of edge weights between homologous proteins (F_{Net}). We performed the prediction for 67 protein pairs on BioGRID. Table A16 in S1 File summarizes the estimated scores for the additional protein pairs. Although the confidence is not high because of the high false positive rates in prediction of PPIs in PSOPIA and many other tools, it is found that many identified additional protein pairs possibly interact with each other ($0.1 \leq S_{all} < 0.5$) and 8 among them were considered as highly probable ($S_{all} \geq 0.5$). Furthermore, even if there is some data configuration change with real complex and PPI datasets, almost all approximate solutions are the same and the additional proteins are detected in the same order as mentioned in the prediction with synthetic data (see Table A2(b) in S1 File).

In addition, GreedyMinPPI only requires a small amount of CPU time regardless of the number of additional interactions and it might be effective for an application with large-scale datasets. To summarize, the current PPI data identified by eight databases is incomplete and does not adequately describe all PPIs involved in CYC2008; on the positive side, GreedyMinPPI enables us to detect the minimum number of additional interactions required to support all regulatory interactions among proteins which constitute the corresponding protein complex.

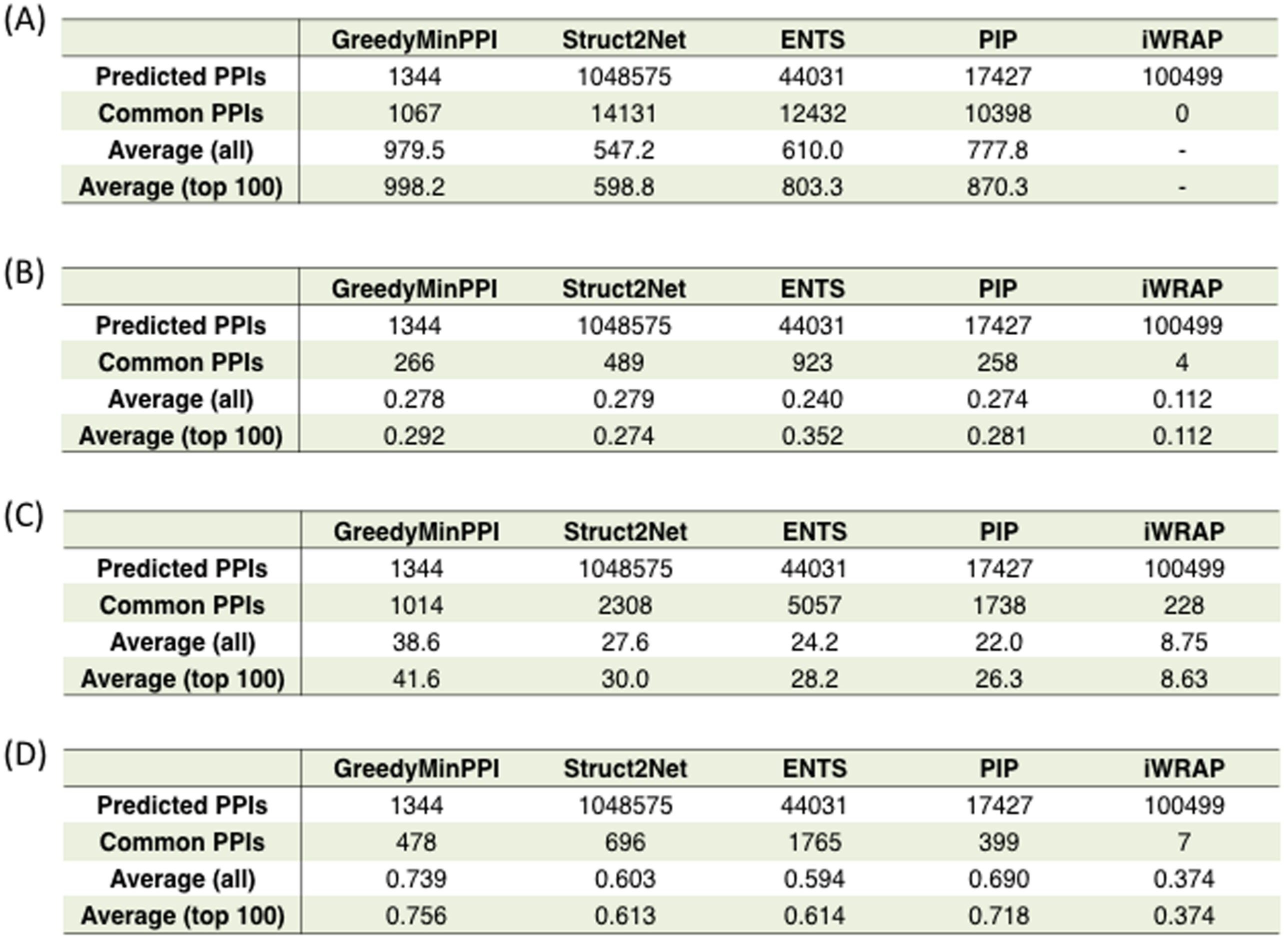
Comparison of prediction performance

Performance comparison with weighted PPI datasets. To investigate the accuracy of GreedyMinPPI, we compared our results with those from existing PPI prediction methods using four different weighted PPI datasets predicted from Struct2Net [26], ENTS [27], PIP [28] and iWRAP [29]. Struct2Net is a web server for predicting PPIs based on the structural features using protein sequence data as input data. ENTS is a random forest based PPI prediction method only from the primary sequence data. PIP is developed based on a naïve Bayes classifier to stochastically predict whether each protein pair is present in the same complex regardless of their direct interaction. iWRAP is a threading-based prediction method for detecting de novo cancer related interactions. Since all interactions predicted from Struct2Net, ENTS, PIP and iWRAP were assigned the confidence score, we assessed by measuring the ROC curve and AUC (Area Under the Curve) score. In order to plot the ROC curve, we regarded the PPIs obtained from STRING, MINT, WI-PHI and IntAct databases as the gold standard (refer to S1 File) [40]. It must be noted that, in comparative experiments with those databases, we limited PPIs to those of extracted PPIs composed of 1627 proteins stored in those databases.

As for the confidence score assignment, for each interaction by GreedyMinPPI, the scores between [1-1344] were calculated to reflect the reliability where the firstly added interaction has the highest confidence which equals to 1344. Four existing methods also provided individual confidence scores, Struct2Net score ranging [0.25-0.98], ENTS score ranging [0.50-0.97], PIP score ranging [300.07-146673.28] and iWRAP score ranging [0.90-1.00].

Fig 6 shows the number of predicted and common PPIs and the averages for all confidence scores and in the 100 highest scoring group for each database. The results show that the number of common PPIs that are shared between the PPIs predicted from GreedyMinPPI and those derived from STRING, WI-PHI are 1067, 1014, respectively, and the averages for all and the top 100 are the highest scoring among other methods. It should be noted that in case of IntAct, although the percentage of the number of common interactions is less than 50, the average for the top 100 by GreedyMinPPI is higher than those by other methods. It is because GreedyMinPPI has an advantage of being able to detect the minimum number of PPIs or additional PPIs by firstly adding edges among proteins that are most highly overlapping. In contrast, the averages of confidence scores of Struct2Net and ENTS are higher than those of GreedyMinPPI using MINT. Note also that since the iWRAP PPIs were predicted using the dataset of yeast cancer related genes and iWRAP detected an interaction between XPA (RAD14) and SMARCA5, whose overexpression leads to cell proliferation [29, 41], it provides only limited prediction and does not have many common interactions with four databases. In this way, although GreedyMinPPI does not guarantee the optimality of its solution, it enables us to provide high confidence protein interactions.

Comparison of distribution of PPI confidence score. Since the four databases provide confidence scores, each of which was computed by evidences from their own sources and all five scored PPIs were sorted in descending order in advance [42–44], we examined the



(A)

	GreedyMinPPI	Struct2Net	ENTS	PIP	iWRAP
Predicted PPIs	1344	1048575	44031	17427	100499
Common PPIs	1067	14131	12432	10398	0
Average (all)	979.5	547.2	610.0	777.8	-
Average (top 100)	998.2	598.8	803.3	870.3	-

(B)

	GreedyMinPPI	Struct2Net	ENTS	PIP	iWRAP
Predicted PPIs	1344	1048575	44031	17427	100499
Common PPIs	266	489	923	258	4
Average (all)	0.278	0.279	0.240	0.274	0.112
Average (top 100)	0.292	0.274	0.352	0.281	0.112

(C)

	GreedyMinPPI	Struct2Net	ENTS	PIP	iWRAP
Predicted PPIs	1344	1048575	44031	17427	100499
Common PPIs	1014	2308	5057	1738	228
Average (all)	38.6	27.6	24.2	22.0	8.75
Average (top 100)	41.6	30.0	28.2	26.3	8.63

(D)

	GreedyMinPPI	Struct2Net	ENTS	PIP	iWRAP
Predicted PPIs	1344	1048575	44031	17427	100499
Common PPIs	478	696	1765	399	7
Average (all)	0.739	0.603	0.594	0.690	0.374
Average (top 100)	0.756	0.613	0.614	0.718	0.374

Fig 6. Comparison of the PPIs predicted by the five methods with STRING (A), MINT (B), WI-PHI (C) and IntAct (D) databases.

https://doi.org/10.1371/journal.pone.0195545.g006

distributions of database confidence score of the top 100 interactions that are calculated by five existing methods with STRING score ranging [150-999], MINT score ranging [0.091-0.984], WI-PHI score ranging [6.624-146.551] and IntAct score ranging [0.216-0.963] as shown in Fig 7 and A1–A4 Figs in S1 File.

In particular, the distribution with STRING exhibits that GreedyMinPPI detected the PPIs with uniformly high confidence score which suggests that the prediction is performed with a high reliability as shown by the high average score (Fig 7). Since iWRAP and STRING have no common edges, the distribution of iWRAP cannot be plotted. In case of WI-PHI, GreedyMinPPI can detect the interactions with relatively high confidence score than other methods, on the other hand, ENTS could predict interactions with higher scores than GreedyMinPPI using MINT. In case of IntAct, the distribution of the confidence score on GreedyMinPPI looks similar to the distribution by PIP, and certainly there is only a slight difference in the range of averages of top 100 interactions by GreedyMinPPI and PIP. Similarly, the distribution by Struct2Net and ENTS can appear to be almost the same (A1–A4 Figs in S1 File).

Therefore, the results suggest that the top 100 interactions predicted by GreedyMinPPI have higher confidence scores than those by existing methods but not providing the uniformly high scores except in the case of STRING. However, since it enables us to detect the interactions

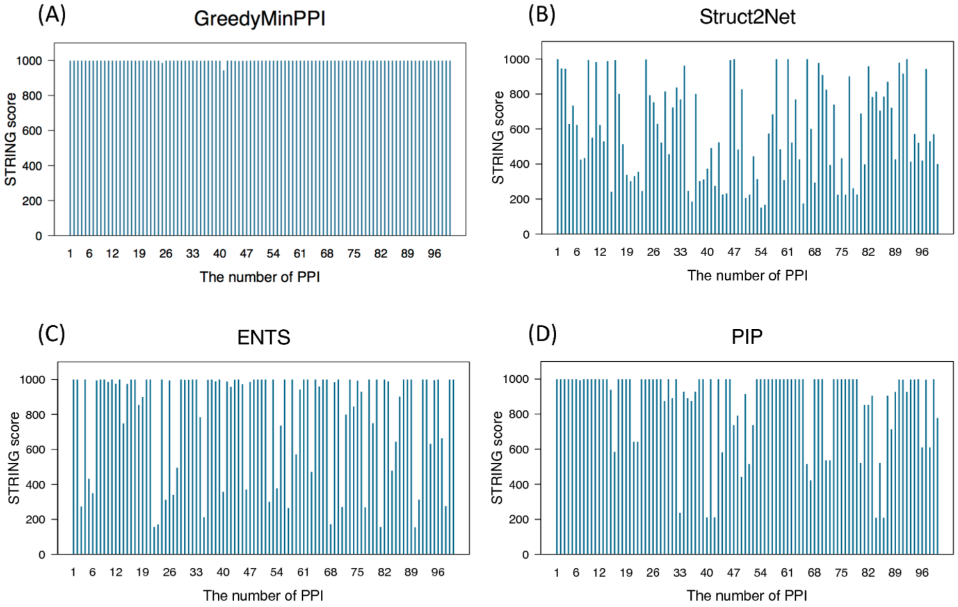


Fig 7. Distribution of PPI confidence score using STRING. The distribution of the confidence scores of PPIs predicted by GreedyMinPPI (A), Struct2Net (B), ENTS (C) and PIP (D).

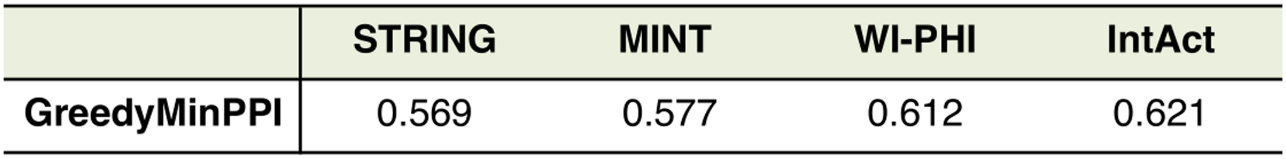
https://doi.org/10.1371/journal.pone.0195545.g007

among proteins with most highly overlapping, most of these interactions would be expected to be more reliable.

Performance comparison with unweighted PPI datasets. In this subsection, we validated the predictive performance of GreedyMinPPI with three unweighted PPI prediction methods, PIPE [30], SPPS [45] and InteroPORC [46]. PIPE predicts PPIs only using information derived from the primary sequence of *S. cerevisiae* which consists of 6304 protein sequences. SPPS server is developed by combining Support Vector Machines (SVM) with the posterior probability which is derived from a sigmoid function, to search the possible interacting partners with the query protein sequence. InteroPORC provides an automated prediction based on the interolog concept, which reconstructs the orthology relationships across multiple species.

The performances of the three unweighted datasets were assessed by measuring six criteria, recall, precision, specificity, accuracy (ACC), *F*-measure and matthew's correlation coefficient (MCC) (see S1 File). Recall is the proportion of actual positives which are correctly predicted; precision measures the proportion of positive predictions which are actual positives and there is a trade-off between recall and precision; specificity measures the proportion of actual negatives that are correctly predicted; accuracy is the proportion of the total number of correct predictions; *F*-measure is the harmonic mean of recall and precision; matthew's correlation coefficient is the correlation coefficient between the actual and predicted binary classes, that takes from −1 to 1.

Needless to say, since the objective of MinPPI0 or MinPPI is to detect the minimum number of interactions or additional interactions so that all proteins belonging to each complex are connected, the number of predicted PPIs from GreedyMinPPI should mostly be lower than those from existing methods. This leads to the lower AUC performance of GreedyMinPPI compared to those of existing methods, as shown in Fig 8 and refer to the "original" in Tables A18–A21 in S1 File. Therefore, we tested whether the performance of existing methods can be improved by combining the interactions predicted from GreedyMinPPI. Firstly, we examined



	STRING	MINT	WI-PHI	IntAct
GreedyMinPPI	0.569	0.577	0.612	0.621

Fig 8. AUC scores on GreedyMinPPI with the four databases.

https://doi.org/10.1371/journal.pone.0195545.g008

the six criteria of GreedyMinPPI and three methods using the four databases, STRING, MINT, WI-PHI and IntAct as the gold standard. Secondly, we also examined them to compare the performance of six criteria on their own datasets from PIPE, SPPS and InteroPORC with those on the combined datasets, each of which was constructed by adding all interactions from GreedyMinPPI. All measurement results are summarized in Table A17 in S1 File. Certainly, the four criteria except for specificity and ACC of GreedyMinPPI are relatively low for each database. However, almost all criteria of combined datasets for three existing methods are slightly improved in case of using STRING, MINT and WI-PHI databases, although some of them on the combined datasets perform a little bit worse than those on their own dataset using IntAct. From this experiment, it is found that since GreedyMinPPI predicts additional interactions including unknown PPIs, it might be helpful for enhancement of existing PPI prediction methods.

AUC performance comparison with weighted PPI datasets. To validate the performance enhancement of GreedyMinPPI with the weighted PPIs predicted from Struct2Net, ENTS, PIP and iWRAP in terms of AUC, we examined MinPPI using the dataset, combining all interactions and the corresponding confidence scores from GreedyMinPPI with those from the four existing methods. The combined scores were computed by,

$$S_c = SPE \times S_g + S_e \tag{3}$$

where *SPE* reflects the specific weight of scores predicted from GreedyMinPPI when combining the scores. S_c is the combined score and S_g and S_e are the individual scores that are calculated from GreedyMinPPI and existing methods, respectively.

Comparison results with the four databases are summarized in Tables A18–A21 in S1 File. Since iWRAP does not have any common edges with STRING, the AUC score of iWRAP on the original dataset could not be computed. In case of any database, regardless of the value of *SPE*, the AUC scores on the combined datasets increased than those on their own datasets from GreedyMinPPI. Because although GreedyMinPPI is based only on the connected components from graph theory, MinPPI tends to output pairs of proteins that are contained in many protein complexes. In other words, in this paper, we assume that the proteins consisting of highly overlapping probably interact with each other and these interactions might be missing interactions. In particular, the AUC on iWRAP was greatly improved when combining all interactions because fewer shared interactions exist in case of each database. On the other hand, the AUC performance was not improved using the combined datasets when combining top 200 and 500 interactions from GreedyMinPPI. These results suggest that GreedyMinPPI is also suitable for enhancement of the predictive performance of existing methods when all outputted interactions are utilized.

Discussion and conclusion

In this paper, we have introduced MinPPI, which is a problem of determining the minimum number of PPIs required to support known protein complexes. For solving this problem, we have developed a novel integer linear programming-based method (ILPMinPPI) and a greedy-type method (GreedyMinPPI) based on an existing greedy-type approximation algorithm. The comparison of these two methods using moderate size synthetic data suggests that GreedyMinPPI outputs optimal or near-optimal solutions for practical instances. Since ILPMinPPI cannot be applied to large-scale data, we have applied GreedyMinPPI to pairs of a protein complex dataset and eight PPI datasets. Our findings show that the minimum number of additional required PPIs ranges from 51 (STRING) to 964 (BIND). Significantly, this suggests that even the four best PPI databases, STRING (51), BioGRID (67), WI-PHI (93) and iRefIndex (85) do not have enough PPIs to form all CYC2008 protein complexes. We have also applied GreedyMinPPI to enhance the existing PPI prediction methods. The results suggest that it is also useful for that purpose.

Although ILPMinPPI and GreedyMinPPI output PPIs with scores, they output PPIs only based on the minimum requirement that each complex must constitute a connected subgraph in a PPI network. Therefore, these methods are not optimized for prediction of PPIs and thus should be used only as auxiliary methods to enhance existing PPI prediction methods. However it might be possible to modify the formalization of MinPPI as a kind of machine-learning problem to infer PPIs from protein complexes. Although such a variant would still not be enough to be used as a stand-alone prediction method, it would be useful to further enhance the prediction accuracy of existing PPI prediction methods. Another important future work is to improve ILPMinPPI so that it can be applied to real protein complex datasets, because it is unclear whether GreedyMinPPI outputs near-optimal solutions for large-scale protein complex datasets and the theoretically guaranteed $O(\log m)$ approximation ratio is not enough to estimate the minimum number.

Supporting information

S1 File. Fig A1:Distribution of PPI confidence score using STRING. Fig A2:Distribution of PPI confidence score using MINT. Fig A3:Distribution of PPI confidence score using WI-PHI. Fig A4:Distribution of PPI confidence score using IntAct. Table A1:Performance evaluation of ILPMinPPI and GreedyMinPPI using synthetic data. Table A2:Prediction results on GreedyMinPPI with different data configuration. Table A3:Results with three protein complex datasets. Table A4:The additional protein pairs for supporting CYC2008 protein complex with STRING. Table A5:The additional protein pairs and the included complexes on STRING. Table A6, A9, A12, A15:Frequency distribution of the assigned categories of additional proteins on STRING from PANTHER analysis. Table A7:The additional protein pairs for supporting CYC2008 protein complex with BioGRID. Table A8:The additional protein pairs and the included complexes on BioGRID. Table A10:The additional protein pairs for supporting CYC2008 protein complex with WI-PHI. Table A11:The additional protein pairs and the included complexes on WI-PHI. Table A13:The additional protein pairs for supporting CYC2008 protein complex with iRefIndex. Table A14:The additional protein pairs and the included complexes on iRefIndex. Table A16:67 additional protein pairs on BioGRID and S_{all} which is the estimated score by PSOPIA. Table A17:Comparison of prediction performance with the four databases using unweighted datasets. Table A18:Comparison of prediction performance using AUC with STRING. Table A19:Comparison of prediction performance using AUC with MINT. Table A20:Comparison of prediction performance using AUC with WI-PHI. Table A21:Comparison of prediction performance using AUC with IntAct.
(PDF)

Acknowledgments

T. Akutsu was partially supported by JSPS, Japan, (Grants-in-Aid 26240034). M. Hayashida was partially supported by JSPS, Japan, (Grants-in-Aid 16K00392). J. Jansson was partially supported by the Hakubi Project at Kyoto University, Japan. O. Maruyama was supported by JSPS KAKENHI Grant Number 26330330.

Author Contributions

Conceptualization: Osamu Maruyama, Tatsuya Akutsu.

Data curation: Natsu Nakajima, Morihiro Hayashida.

Formal analysis: Natsu Nakajima, Jesper Jansson, Osamu Maruyama, Tatsuya Akutsu.

Methodology: Natsu Nakajima, Jesper Jansson, Osamu Maruyama, Tatsuya Akutsu.

Software: Natsu Nakajima.

Writing – original draft: Natsu Nakajima, Jesper Jansson, Tatsuya Akutsu.

Writing – review & editing: Natsu Nakajima, Morihiro Hayashida, Jesper Jansson, Osamu Maruyama, Tatsuya Akutsu.

References

1. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006; 440(7084):631–636. https://doi.org/10.1038/nature04532 PMID: 16429126
2. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006; 440(7084):637–643. https://doi.org/10.1038/nature04670 PMID: 16554755
3. Enright AJ, Dongen SV, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research. 2002; 30(7):1575–1584. https://doi.org/10.1093/nar/30.7.1575 PMID: 11917018
4. Bader BD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003; 4:2. https://doi.org/10.1186/1471-2105-4-2 PMID: 12525261
5. King AD, Przulj N, Jurisica I. Protein complex prediction via cost-based clustering. Bioinformatics. 2004; 20(17):3013–3020. https://doi.org/10.1093/bioinformatics/bth351 PMID: 15180928
6. Macropol K, Can T, Singh AK. Repeated random walks on genome-scale protein networks for local cluster discovery. BMC Bioinformatics. 2009; 10:283. https://doi.org/10.1186/1471-2105-10-283 PMID: 19740439
7. Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics. 2009; 25(15):1891–1897. https://doi.org/10.1093/bioinformatics/btp311 PMID: 19435747
8. Maruyama O, Chihara A. NWE: Node-weighted expansion for protein complex prediction using random walk distances. Proteome Science. 2011; 9 Suppl 1:S14. https://doi.org/10.1186/1477-5956-9-S1-S14 PMID: 22165822
9. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, et al. High quality binary protein interaction map of the yeast interactome network. Science. 2008; 322(5898):104–110. https://doi.org/10.1126/science.1158684 PMID: 18719252
10. Ruan P, Hayashida M, Maruyama O, Akutsu T. Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels. BMC Bioinformatics, 2014; 15 Suppl 2:S6. https://doi.org/10.1186/1471-2105-15-S2-S6 PMID: 24564744
11. Angluin D, Aspnes J, Reyzin L. Network construction with subgraph connectivity constraints. Journal of Combinatorial Optimization. 2015; 29(2):418–432. https://doi.org/10.1007/s10878-013-9603-2
12. Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research. 2009; 37(3):825–831. https://doi.org/10.1093/nar/gkn1005 PMID: 19095691
13. Mewes HW, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Research. 2002; 30(1):31–34. https://doi.org/10.1093/nar/30.1.31 PMID: 11752246

14. Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, et al. Structure-based assembly of protein complexes in yeast. Science. 2004; 303(5666):2026–2029. https://doi.org/10.1126/science.1092645 PMID: 15044803

15. Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Research. 2003; 31(1):D258–D261. https://doi.org/10.1093/nar/gkg034

16. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the molecular interaction database. Nucleic Acids Research. 2007; 35:D572–D574. https://doi.org/10.1093/nar/gkl950 PMID: 17135203

17. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006; 34:D535–D539. https://doi.org/10.1093/nar/gkj109 PMID: 16381927

18. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP: the database of interacting proteins. Nucleic Acids Research. 2000; 28(1):289–291. https://doi.org/10.1093/nar/28.1.289 PMID: 10592249

19. Bader GD, Betel D, Hogue CW. BIND: the biomolecular interaction network database. Nucleic Acids Research. 2003; 31(1):248–250. https://doi.org/10.1093/nar/gkg056 PMID: 12519993

20. Kiemer L, Costa S, Ueffing M, Cesareni G. WI-PHI: a weighted yeast interactome enriched for direct physical interactions. Proteomics. 2007; 7(6):932–943. https://doi.org/10.1002/pmic.200600448 PMID: 17285561

21. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, et al. IntAct: an open source molecular interaction database. Nucleic Acids Research. 2004; 32:D452–D455. https://doi.org/10.1093/nar/gkh052 PMID: 14681455

22. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res. 2010; 38:D525–D531. https://doi.org/10.1093/nar/gkp878 PMID: 19850723

23. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics. 2008; 9:405. https://doi.org/10.1186/1471-2105-9-405 PMID: 18823568

24. Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, Lappe M, et al. Estimating the size of human interactome. Proc Natl Acad Sci U S A. 2008; 105(19):6959–6964. https://doi.org/10.1073/pnas.0708078105 PMID: 18474861

25. Sambourg L, Thierry-Mieg N. New insights into protein-protein interaction data lead to increased estimates of the S. cerevisiae interactome size. BMC Bioinformatics. 2010; 11:605. https://doi.org/10.1186/1471-2105-11-605 PMID: 21176124

26. Singh R, Park D, Xu J, Hosur R, Berger B. Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Research. 2010; 38:W508–W515. https://doi.org/10.1093/nar/gkq481 PMID: 20513650

27. Rodgers-Melnick E, Culp M, DiFazio SP. Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. BMC Genomics. 2013; 14:608. https://doi.org/10.1186/1471-2164-14-608 PMID: 24015873

28. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003; 302(5644):449–453. https://doi.org/10.1126/science.1087361 PMID: 14564010

29. Hosur R, Xu J, Bienkowska J, Berger B. iWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. Journal of Molecular Biology. 2011; 405 (5):1295–1310. https://doi.org/10.1016/j.jmb.2010.11.025 PMID: 21130772

30. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC Bioinformatics. 2006; 7:365. https://doi.org/10.1186/1471-2105-7-365 PMID: 16872538

31. Chockler G, Melamed R, Tock Y, Vitenberg R. Constructing scalable overlays for pub-sub with many topics. Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing, Portland, OR, USA. ACM New York, 2007; p.109–118.

32. Korach E, Stern M. The clustering matroid and the optimal clustering tree. Mathematical Programming, Series B. 2003; 98(1);385–414. https://doi.org/10.1007/s10107-003-0410-x

33. Korach E, Stern M. The complete optimal stars-clustering-tree problem. Discrete Applied Mathematics. 2008; 156(4);444–450. https://doi.org/10.1016/j.dam.2006.12.004

34. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods. 2012; 9(5):471–472. https://doi.org/10.1038/nmeth.1938 PMID: 22426491

35. Chen XW, Jeong JC, Dermyer P. KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. Nucleic Acids Research. 2011; 39:D750–D754. https://doi.org/10.1093/nar/gkq943 PMID: 20952400

36. Schmitt T, Ogris C, Sonnhammer EL. FunCoup 3.0: database of genome-wide functional coupling networks. Nucleic Acids Research. 2014; 42:D380–D388. https://doi.org/10.1093/nar/gkt984 PMID: 24185702

37. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife. 2014; 3:e02030. https://doi.org/10.7554/eLife.02030 PMID: 24842992

38. Padhorny D, Kazennov A, Zerbe BS, Porter KA, Xia B, Mottarella SE, et al. Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds. Proc Natl Acad Sci U S A. 2016; 113 (30):E4286–E4293. https://doi.org/10.1073/pnas.1603929113 PMID: 27412858

39. Murakami Y, Mizuguchi K. Homology-based prediction of interactions between proteins using Averaged One-Dependence Estimators. BMC Bioinformatics. 2014; 15(213). https://doi.org/10.1186/1471-2105-15-213 PMID: 24953126

40. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Computational Biology 2007; 3(3):e42. https://doi.org/10.1371/journal.pcbi.0030042 PMID: 17397251

41. Jin Q, Mao X, Li B, Guan S, Yao F, Jin F. Overexpression of SMARCA5 correlates with cell proliferation and migration in breast cancer. Tumour Biology. 2015; 36(3):1895–1902. https://doi.org/10.1007/s13277-014-2791-2 PMID: 25377162

42. Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature. 1989; 340 (6230):245–246. https://doi.org/10.1038/340245a0 PMID: 2547163

43. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B. A genetic protein purification method for protein complex characterization and proteome exploration. Nature biotechnology. 1999; 17 (10):1030–1032. https://doi.org/10.1038/13732 PMID: 10504710

44. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Bioinformatics. 2007; 6(3):439–450.

45. Liu X, Liu B, Huang Z, Shi T, Chen Y, Zhang J. SPPS: A sequence-based method for predicting probability of protein-protein interaction partners. PLoS One. 2012; 7(1):e30938. https://doi.org/10.1371/journal.pone.0030938 PMID: 22292078

46. Michaut M, Kerrien S, Montecchi-Palazzi L, Chauvat F, Cassier-Chauvat C, Aude JC, et al. InteroPORC: automated inference of highly conserved protein interaction networks. Bioinformatics. 2008; 24 (14):1625–1631. https://doi.org/10.1093/bioinformatics/btn249 PMID: 18508856