Journal of Bioinformatics and Computational Biology
Vol. 13, No. 2 (2015) 1571001 (25 pages)

DOI: 10.1142/S0219720015710018

Imperial College Press
www.icpress.co.uk

From the static interactome to dynamic protein complexes: Three challenges

Chern Han Yong*,‡ and Limsoon Wong†,§

*Graduate School for Integrative Sciences and Engineering
National University of Singapore
28 Medical Drive, Singapore 117456

†School of Computing, National University of Singapore
13 Computing Drive, Singapore 117417
‡cherny@nus.edu.sg
§wongls@comp.nus.edu.sg

Received 22 August 2014
Revised 10 November 2014
Accepted 22 December 2014
Published 4 February 2015

Protein interactions and complexes behave in a dynamic fashion, but this dynamism is not captured by interaction screening technologies, and not preserved in protein–protein interaction (PPI) networks. The analysis of static interaction data to derive dynamic protein complexes leads to several challenges, of which we identify three. First, many proteins participate in multiple complexes, leading to overlapping complexes embedded within highly-connected regions of the PPI network. This makes it difficult to accurately delimit the boundaries of such complexes. Second, many condition- and location-specific PPIs are not detected, leading to sparsely-connected complexes that cannot be picked out by clustering algorithms. Third, the majority of complexes are small complexes (made up of two or three proteins), which are extra sensitive to the effects of extraneous edges and missing co-complex edges. We show that many existing complex-discovery algorithms have trouble predicting such complexes, and show that our insight into the disparity between the static interactome and dynamic protein complexes can be used to improve the performance of complex discovery.

Keywords: Protein complex; protein interaction; dynamism.

1. Introduction

In the cell, many proteins bind physically to form stoichiometrically-stable multi-protein structures called protein complexes. Protein complexes perform a wide variety of molecular functions in many cellular processes, thus it is important to determine the set of complexes in the cell to gain an understanding of the mechanism, organization, and regulation of these processes. Since proteins in a complex interact

‡Corresponding author.

physically, many algorithms have been proposed to analyze protein–protein interaction (PPI) data to discover protein complexes.

The general strategy underlying most complex-discovery algorithms is to represent PPI data as a PPI network (PPIN), where vertices represent proteins and edges represent interactions between proteins, and then find clusters of highly interconnected proteins within the PPIN as protein complexes. Over the past decade, these algorithms have grown in sophistication and variety, and have incorporated increasing amounts of useful biological insights in their designs. However, the performance of most of these approaches, even under optimal conditions, still leaves room for improvement: For example, even in yeast with decently-comprehensive PPI data, accurate prediction of complexes at fine resolution remains difficult.

One main stumbling block is that the representations and analyses of PPIs for the purpose of complex prediction have been overwhelmingly static, even though it has been well understood that proteins and complexes exhibit a sophisticated dynamism in behavior. Proteins interact in a dynamic fashion, with a variety of interaction timings, locations, and affinities. These are mediated by a wide range of factors including cellular state, cellular processes, and the interaction environment.[1] Correspondingly, protein complexes exhibit dynamic behavior which are in fact important functional mechanisms, for example to allow complexes to be formed only at certain times, or to vary the composition of complexes to modulate or activate their functions. However, due to limitations in PPI-detection methodologies, it is difficult to interrogate the dynamics of PPIs (i.e. when, where, and how a protein interacts with others). Furthermore, this dynamism also precludes a faithful interrogation of PPIs in the cell (e.g. condition-specific PPIs may be missed, or spurious PPIs may be detected in non-physiological experimental systems). Moreover, the representation of PPIs in the PPIN does not preserve any information about the dynamics of PPIs. Thus there exists a disparity between the dynamic nature of PPIs and protein complexes on one hand, and the static representation and analysis of the PPIN on the other hand.

We identify three challenges in protein-complex discovery that arise from, or are exacerbated by, this static view of PPIs and protein complexes. First, many complexes are embedded within highly-connected regions of the PPIN, with many extraneous edges connecting a complex's member proteins to other proteins outside the complex. This arises because many proteins participate in multiple distinct complexes, resulting in complexes overlapping each other in dense regions in the PPIN. Spuriously-detected interactions further contribute to this problem. Second, many complexes exist in sparse regions of the network, so that proteins within the complexes are not densely interconnected. This arises from undetected condition-specific, location-specific, or transient PPIs. Third, many complexes are small (that is, composed of two or three proteins), making measures of important topological features, such as density, ineffectual. This is further exacerbated by extraneous or missing interactions which can embed a small complex in a larger clique, or disconnect it entirely.

In this paper, we evaluate the performance of various complex-discovery algorithms, covering different types of approaches, in the prediction of yeast and human complexes. In particular, we highlight the unsatisfactory performance in predicting complexes within highly-connected regions, complexes within sparse regions, and small complexes, and discuss how an understanding of the dynamics of protein interactions may be used to address the shortcomings of these algorithms with respect to these specific challenges.

A number of surveys on complex discovery have been published in recent years. Li *et al.*[2] in 2010 surveyed a number of complex-discovery algorithms, and categorized them according to the types of data used and the features of the algorithms. Srihari and Leong[3] in 2013 further showed that complex-discovery algorithms have evolved to incorporate increasing amounts of biological information in their designs, leading to improved performance and new biological insights. Most recently, Chen *et al.*[4] also surveyed and categorized various complex-discovery algorithms, with a distinct category for algorithms that explicitly model the dynamism of PPIs. Since descriptions and taxonomies of complex-discovery algorithms are already covered in these surveys, our paper instead emphasizes specific challenges raised by the dynamism of PPIs, and evaluates a few classic and recent algorithms with respect to these challenges.

In Sec. 2, we elaborate on protein interactions and protein complexes in the cell, with an emphasis on the dynamism of their behaviors. We give a brief background on PPI-screening technologies and their inadequacies, particularly in capturing such dynamism. In Sec. 3, we show how the three challenges in complex discovery follow from the analysis of static PPIs. In Sec. 4, we describe our experiments to evaluate five clustering algorithms in yeast and human complex discovery, with an emphasis on their shortcomings with respect to the three challenges that we have highlighted. In Sec. 5, we conclude our findings, and describe some approaches that help to address these challenges, although much room remains for improvement.

2. Background: From Interactome to Complexome

In the study of protein complexes, the interactome refers to the set of cellular physical PPIs, while the complexome describes the set of cellular complexes. Since complexes consist of physically-interacting proteins, they correspond to groups of proteins with high degrees of co-interaction in the interactome. Thus, deriving the complexome from the interactome is a fruitful strategy that has been well researched over the past decade. Many challenges have been acknowledged in this strategy, a significant portion of which we distil as the 'disparity' between the static interactome and the complexome: Due to limitations in detection technologies and methodologies (which have only recently begun to be surpassed), the views and analyses of the interactome and complexome have been overwhelmingly static, without consideration of the dynamic nature of PPIs and the corresponding dynamism of protein complexes.

2.1. *Dynamism of protein interactions*

In fact, the static interactome, understood as the set of PPIs that exist in a cell, is a mere shadow of the dynamic and complex lives of PPIs in reality, which involve a wide range of interaction timings, locations, and binding affinities.

A protein with multiple interaction partners does not necessarily interact with all of them simultaneously. When a protein interacts, and which partner it interacts with, are controlled by different cellular mechanisms. For example, co-localization of the interactors in time and space, as well as the local concentration of the interactors, are controlled by expression, mRNA degradation, protein transport, protein secretion, protein degradation; the binding affinities of different interactors are controlled through post-translational modification of the interactors, or changes to the physiochemical environment, for example by the concentration of effector molecules like ATP that may change binding affinity.[1]

Different classes of PPI binding affinities have been proposed[1,5,6]: permanent interactions, with the strongest binding affinity, are irreversible; weak transient interactions, with the weakest binding affinity, are reversible, and involve proteins that switch between both bound and unbound states *in vivo*; strong transient interactions lie between permanent interactions and weak transient interactions, and are reversible when triggered, for example by ligand binding. PPIs can also be characterized as obligate or non-obligate: proteins with obligate interactions cannot exist as stable structures on their own, and are frequently bound to their partners upon translation and folding; conversely, proteins with non-obligate interactions can exist as stable structures both in bound and unbound states.

A study of protein hubs (proteins with a large number of interaction partners) with gene-expression data has led to a proposed distinction between date hubs and party hubs[7,8]: party hubs interact with all of their partners simultaneously as a large complex, while date hubs interact with its partners in mutually-exclusive times, and are believed to link diverse biological processes together in the PPIN.

2.2. *Dynamism of protein complexes*

Consequently, complexes display a range of dynamism in their formation, composition, and stability, which impart important functional mechanisms to the complexes' activities. For example, the highly conserved Cdc28p (a cyclin-dependent kinase or CDK) yeast protein regulates the cell-cycle by forming complexes with different cyclin proteins that phosphorylate different substrates to promote entry into different cell-cycle phases[9,10]: progressing through the cell-cycle phases, these include Cdc28p forming complexes with Cln3p to enter the cycle, with Cln1,2p in G_1 phase, with Clb5,6p to begin replication in S phase, and with Clb1,2,3,4p to enter M phase (see Fig. 7(a)). These complexes are themselves regulated through binding with cyclin-dependent kinase inhibitors (CKIs) such as Sic1p.

An integrated analysis of protein complexes with cell-cycle expression data revealed "just-in-time" assembly of most cell-cycle-related complexes in

yeast[11]: some subunits of complexes are constitutively expressed (static proteins), while other subunits are expressed only when needed (dynamic proteins), so that the entire complex can be assembled only in specific cell-cycle phases without having to transcriptionally regulate all the subunits of the complex.

The dynamism of complexes also gives them a modular architecture in function and composition, which has been described with the core-attachment model of complexes.[12] Here, the core of a complex consists of proteins that interact permanently, while attachment proteins are recruited to the core via less permanent interactions, which may modulate or activate the function of the complex.

2.3. *Interactome screening technologies*

The dynamism of PPIs, which provides such important functional mechanisms for complexes, is not captured in the static interactome. A chief reason for this is the technological limitations of past high-throughput PPI screening experiments, which has only recently begun to be surpassed.

In the past decade, the two commonly used methods for high-throughput screening of PPIs are based on the yeast two-hybrid assay (Y2H), which detects binary interactions, and the tandem affinity purification with mass spectrometry (TAP-MS) method, which detects co-complex interactions. The Y2H method uses a fragmented transcription factor to detect the interaction between a bait protein and a prey protein: when the proteins interact, they cause the transcription of a reporter gene.[13] A recent survey of advances in Y2H technology is provided by Bruckner *et al.*[14]

The Y2H assay is able to detect transient or weak interactions, but is limited to only direct physical PPIs: interactions between co-complex proteins (proteins in the same complex) that do not physically interact with each other are not detected. Y2H assays interactions at non-physiological conditions (e.g. the bait and prey proteins may be overexpressed, co-expressed, or post-translationally modified, whereas they may not be *in vivo*), so some interactions may be spuriously detected. Since interactions are interrogated in a controlled homogeneous cellular state, those that occur in other condition-specific states (such as different cell-cycle or perturbation states) may not be captured. Furthermore, some interacting proteins are unable to localize in the nucleus, or cannot interact in the nucleus' environment, so these interactions are not detected. Conversely, proteins that never co-localize *in vivo* and are thus unable to interact might be wrongly detected as interacting in the nucleus.

Aside from the above problems, Y2H also suffers from the variability inherent in interrogating biological systems, leading to poor reproducibility across multiple screens.

TAP-MS[15] allows a bait protein to complex with other proteins under physiological conditions, and washes it through affinity columns to detect its co-complex proteins (the prey proteins) via mass spectrometry. A survey of recent advances in MS-based methods is provided by Gavin *et al.*[16]

TAP-MS typically only captures strong interactions. Unlike the Y2H assay, TAP-MS retrieves proteins co-complexed with the bait protein, including those that are only indirectly-associated via bridging proteins. Furthermore, for bait proteins that form multiple distinct complexes, all the proteins that form the union of these complexes may be purified and detected. To uncover the PPIs from the purified complexes, further processing is needed,[12,17] though this may still lead to false positives (direct interactions imputed between indirectly-associated proteins) and false negatives (interactions between prey proteins not imputed).

In many TAP-MS assays, the bait protein is expressed by non-natural promoters, leading to its over-expression over physiological levels[16] (although in some studies its expression is controlled by natural promoters.[12,18]).

Under TAP-MS, protein complexes in any subcellular location can be purified. Furthermore, since a heterogeneous collection of cells are purified, complexes present in multiple cellular conditions, such as various cell-cycle and growth states, may be retrieved.[12,18] Nevertheless, complexes present only in other conditions, such as specific perturbation states, are not retrieved. Only recently have researchers begun interrogating the composition of complexes under different perturbation states, for example with affinity purification with selected reaction monitoring,[19] or affinity purification combined with sequential window acquisition of all theoretical spectra.[20] Both works represent key advances in methodologies that will allow dynamic and condition-specific views of interactomes in the near future; but for now, the range of the proteins and PPIs probed, as well as the conditions tested, remain limited.

2.4. *The static interactome*

As described above, the Y2H and TAP-MS methods do not capture timing (i.e. simultaneity) or localization information about the PPIs. For interactions whose affinities are dependent on molecular trigger events such as phosphorylation, information about such molecular triggers is lost, and moreover interactions whose triggers are not activated are not captured. Neither Y2H nor TAP-MS interrogate interactions with respect to cellular states: Under Y2H, interactions are assayed in a homogeneous cellular state which is frequently non-physiological; while under TAP-MS, interactions are frequently interrogated in heterogeneous cellular growth states, so that proteins present in complexes from various growth states are retrieved as an undifferentiated set. Moreover, complexes present only in specific perturbation conditions, which are absent from the cells, are not found. The PPIs obtained thus represent a static interactome, lacking the dynamism that imparts important functional mechanisms to the PPIs and the complexes that they comprise; at the same time, this dynamism precludes an accurate screening of the interactome.

The interactome is frequently represented as a PPIN, with vertices representing proteins and edges representing interactions. This representation itself is a simplification of the cellular organization of PPIs: Aside from missing information about interaction timing, location, affinity, and cellular state, the representation of each

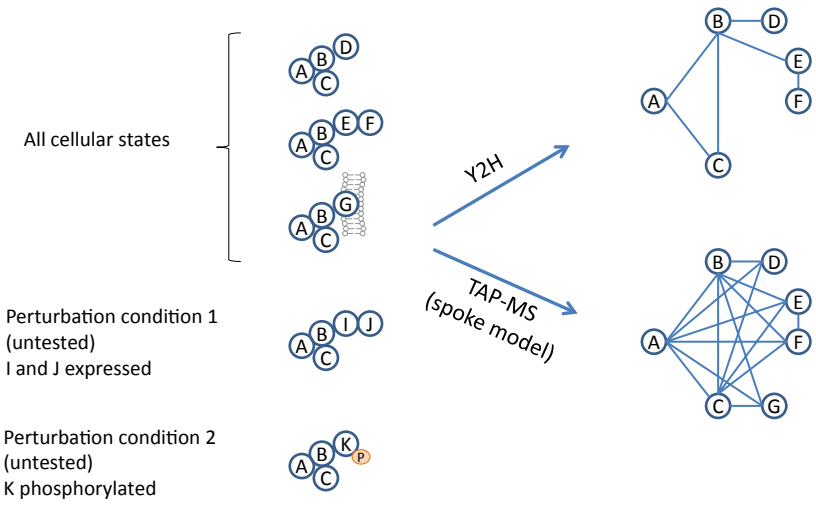


Fig. 1. Detection and representation of dynamically-behaving complexes, in an ideal scenario without spurious or missing interactions. The dynamism of protein complexes is lost after PPI screening and representation in the PPIN. Moreover, this dynamism hinders an accurate screening of PPIs.

protein as a single vertex conflates the multiple copies of each protein that exist in the cell into a single entity: In the cell, different copies of the protein may be simultaneously interacting with different partners, may exist in different cellular locations, and may be in different post-translational states, but in the PPIN all these are represented by a single vertex, and all its disparate interactions are represented as undifferentiated outgoing edges from that vertex.

Figure 1 illustrates these shortcomings of the Y2H and TAP-MS methods for detecting PPIs via a simple example; we ignore the effects of other factors such as experimental or biological variability, which in reality would lead to additional false positives (spurious edges) and false negatives (missing edges). Here, we use a simple made-up complex consisting of an A–B–C core, which forms distinct complexes with either protein D, or proteins E–F, or membrane protein G; additionally, it complexes with proteins I–J which are only expressed during perturbation condition 1, and with protein K only after phosphorylation during perturbation condition 2. We assume that all proteins are used as baits in both Y2H and TAP-MS, and in the latter we use the spoke model to obtain individual PPIs. Since the cells interrogated are never in perturbation conditions 1 or 2, proteins I, J, and K are never found to interact with A–B–C. Y2H is unable to detect the interaction with membrane protein G, while the mutually-exclusive interactions with proteins D and E–F are detected and represented as undifferentiated edges. TAP-MS likewise conflates the three distinct complexes as one large, densely-connected graph. While it appears here that the three complexes can be discerned as separate cliques in the graph, in reality the additional spurious and missing edges make this task difficult.

2.5. *Augmenting the static interactome with dynamism*

Many researchers have recognized that, while the static interactome is a superficial representation of cellular protein interactions, it is still the only proteome-wide and experimentally replicated resource of PPIs that is readily available for computational analysis, and so have attempted to augment it with some degree of dynamism using other information sources.

For example, de Lichtenberg *et al.*[11] integrated yeast PPI data with gene-expression data from various cell-cycle time-points to analyze the dynamism of complex formation during the cell-cycle, and found both constitutively expressed and periodically expressed subunits of most complexes. Likewise, Srihari and Leong[21] also analyzed yeast complexes with cell-cycle expression data, and proposed that constitutively-expressed proteins are likelier to be reused across different complexes.

Other researchers have integrated PPI data with protein-domain information to identify simultaneous or mutually-exclusive interactions. Jung *et al.*[22] decomposed the PPIN into simultaneous protein interaction networks (SPINs), in which all interactions can occur simultaneously, by excluding mutually-exclusive interactions in each SPIN, and then performed complex discovery on each SPIN. Ozawa *et al.*[23] refined predicted complexes by eliminating those that included mutually-exclusive interactions.

A major shortcoming of such analyses is that they are based on the PPIN derived from high-throughput experiments such as Y2H and TAP-MS, so they cannot reveal interactions that are only active in untested conditions.[24] Nevertheless, these approaches show that incorporating this aspect of dynamism in PPIs produces complexes that match known complexes more precisely, and may even elucidate novel functional mechanisms in some complexes. However, the limitations of inferring PPI dynamism indirectly must be noted: for example, gene-expression data does not reflect post-transcriptional activities that further affect complex dynamism, such as protein degradation, transportation, or modification.

3. Three Challenges in Complex Discovery

To discover the set of protein complexes in an organism, researchers have proposed a wide variety of methods to analyze its interactome, derived from high-throughput PPI-screening technologies. A typical strategy is to impute regions of high interconnectedness in the interactome as putative complexes, since proteins within complexes interact with each other. However, since the basis of this analysis is the static interactome, which as described above lacks crucial information about the dynamism of PPIs, a comprehensive and accurate derivation of complexes becomes problematic.

First, a complex may exist within a highly-connected region of the PPIN, with many extraneous outgoing edges connecting it to other proteins outside the complex. Such a complex is challenging to find, as it is difficult to delimit its boundaries

accurately. A particular protein in the complex may have many extraneous PPI edges because it participates in other complexes as well, and the extraneous edges correspond to its interactions with the proteins in these other complexes. These distinct but overlapping (in composition) complexes may exist in different cellular locations, or may form in different cellular states which were detected by the PPI-screening technology, or may even exist in the same location and time as distinct complexes, but this information is not captured in the PPIN. These non-simultaneous interactions corresponding to distinct complexes are active in different copies of the protein, but in the PPIN these multiple copies of the protein are conflated into a single vertex, with all its non-simultaneous interactions corresponding to outgoing edges from that vertex, leading to the many extraneous edges.

The extraneous edges may also correspond to false positives due to a non-physiological environment of the assay, for example through over-expression of bait or prey proteins, or through detected interactions due to post-translational modifications that is different *in vivo*, or through Y2H-detected interactions in the nucleus where the interactors would not localize *in vivo*. Finally, the extraneous edges might simply be an artifact of experimental or other biological variability that is inherent in dealing with biological systems.

Second, a complex may be sparsely-connected in the PPIN, with few PPI edges detected between its proteins. Such a complex does not constitute a dense cluster which can be picked out by clustering algorithms. A complex may be sparse because it is condition-specific: only in certain conditions are its proteins expressed, or modified to enable binding, or co-localized, or the physiochemical environment appropriate for complex formation. If the complex only exists in a condition that was not tested during PPI screening, its proteins' co-complex interactions are not detected. PPIs could also be missing due to technological limitations. Under Y2H, proteins in the complex may not localize in the nucleus or interact in the nucleus where the interaction is assayed — in particular, PPIs in most membrane complexes are not detected. Since Y2H assays interactions in a non-physiological environment, the proteins might not have undergone post-translational modification required for binding, or the environment might be inappropriate for complex formation. Under TAP-MS, weaker interactions may not survive the double-washing step, though they may constitute important interactions within the complex. Finally, as with spurious interactions, missing interactions might also be due to variability in the experimental or biological system.

The third challenge, that of finding small complexes (defined as composed of two or three distinct proteins), is an intrinsic challenge which is exacerbated by the shortcomings of a static interactome. It has been noted that the distribution of complex sizes follows a power law distribution,[25] meaning that a large majority of complexes are small. Thus the discovery of small complexes is an important subtask within complex discovery. An inherent difficulty in this task is that the strategy of searching for dense clusters becomes problematic: fully-dense (i.e. cliques) size-2 and size-3 clusters correspond to edges and triangles, respectively, and only a few among

the abundant edges and triangles of the PPIN represent actual small complexes. Furthermore, small complexes are much more sensitive to extraneous or missing edges: For a size-2 complex, a missing co-complex interaction disconnects its two member proteins, while only two extraneous interactions are sufficient to embed it within a larger clique (a triangle). It is apparent that the challenge of small-complex discovery is exacerbated by the two problems of highly-connected regions with many extraneous edges, and sparse regions with many missing edges, in the PPIN. These problems, as described above, owe a great deal to the analysis of a static interactome to derive complexes that are dynamic in nature.

Figures 7 and 8 in the following section illustrate these challenges via two example complexes, and show how they present problems for clustering algorithms.

4. Poor Performance of Current Methods

In this section, we evaluate five clustering algorithms for the prediction of yeast and human complexes. In particular, we highlight three challenges in complex discovery: the prediction of complexes within highly-connected regions of the PPIN, the prediction of sparsely-connected complexes, and the prediction of small complexes.

4.1. *Clustering algorithms*

In this paper, we evaluate five algorithms, as representatives of different types of clustering algorithms for complex discovery: clique-based, seed-and-grow, simulation, hierarchical, and core-attachment methods. Five additional algorithms — CFinder,[26] IPCA,[27] RNSC,[28] PPSampler,[29] and MCL-CAw[30] — are also evaluated in the Supplementary Materials.

Clustering by Maximal Cliques (CMC[31]) first searches for the set of maximal cliques (cliques that are not contained within a larger clique). Then, for overlapping cliques whose overlap exceeds a threshold, CMC either merges them if they are highly interconnected, or removes the clique with the lower density.

ClusterOne[32] selects vertex seeds based on their degrees, and grows clusters greedily to maximize a cohesiveness function, defined as the ratio of the sum of edge weights within the cluster versus the sum of edge weights within the cluster as well as outgoing edges from the cluster. Furthermore, highly-overlapping clusters are merged.

Markov Clustering (MCL[33]) is based on the principle that a random walker in the PPIN will spend more time traversing a dense region before leaving it. The PPIN is represented as a transition matrix, and the probability of each node visiting every other node at each successive time step is calculated iteratively via matrix multiplication. An inflation step accentuates the differences in probabilities by raising them to a power and then re-normalizing. Regions that are densely connected, with sparse outgoing edges, are found as clusters.

Hierarchical Agglomerative Clustering with Overlap (HACO[34]) first considers all vertices as individual clusters, then iteratively merges pairs of clusters

with high connectivity between them. At each merge, the two constituting clusters are remembered; when the merged cluster A is later merged with another cluster B, it also tries to merge the remembered constituting clusters of A with the cluster B, and keeps the (possibly overlapping) resultant clusters if they are highly-connected.

Coach[35] employs a core-attachment model to detect complexes in two stages: core detection and complex formation. In the first stage, neighborhood subgraphs are induced around each vertex and its neighbors, and cores are found as vertices in each neighborhood subgraph that have higher than average local degree, and whose induced subgraph is dense. In the second stage, proteins that are connected to at least some proportion of each core's vertices are recruited as attachments to the core.

Table 1 shows the parameter settings of these five clustering algorithms used for the prediction of yeast and human complexes.

4.2. *Data sources*

4.2.1. *PPI data*

A number of repositories for PPI data are available, covering a range of organisms, interactions types (genetic interactions or physical PPIs), interactions sources (such as curated PPIs, experimental PPIs, or predicted PPIs), and experimental detection methods. In our work, we obtain our yeast and human PPIs by taking the union of physical PPIs from three repositories: BioGRID,[36] IntAct,[37] and MINT.[38] In yeast we also incorporate the Consolidated PPI dataset.[17]

We unite these datasets, and score and filter the PPIs, using a simple reliability metric based on the Noisy-Or model to combine experimental evidences (also used by Chua *et al.*[39]). For each experimental detection method e, we estimate its reliability as the fraction of interactions detected where both interacting proteins share at least one high-level cellular-component Gene Ontology term. Then the score of an interaction (a, b) is estimated as:

$$\text{score}(a, b) = 1 - \prod_{i \in E_{a,b}} (1 - \text{rel}_i)^{n_{i,a,b}},$$

Table 1. Five clustering algorithms tested, and their parameters used for discovery of yeast and human complexes.

	Category	Parameters
CMC	Clique-based	Yeast: ov = 0.5, mg = 0.5
		Human: ov = 0.5, mg = 0.75
ClusterOne	Seed-and-grow	Yeast and human: default
MCL	Optimization	Yeast: –I 2.5
		Human: –I 4
HACO	Hierarchical	Yeast: –c c 0.75 –g 0.1
		Human: –c c 0.75 –g 0.5
Coach	Core-attachment	Yeast and human: default

where rel_i is the estimated reliability of experimental method i, $E_{a,b}$ is the set of experimental methods that detected interaction (a, b), and $n_{i,a,b}$ is the number of times that experimental method i detected interaction (a, b). We avoid duplicate counting of evidences across the datasets by using their publication IDs.

Most clustering algorithms perform better when a smaller subset of high-quality PPIs are used. We determined that a score cutoff of 0.8 and 0.7 in yeast and human large complexes respectively, and 0.99 in small complexes, gave decent performance in most clustering algorithms.

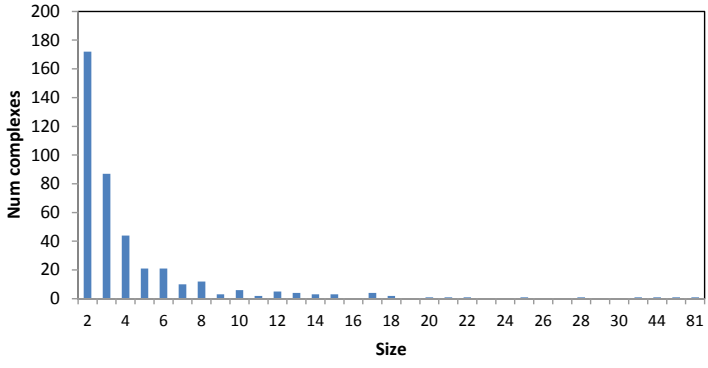
4.2.2. *Reference complexes for yeast and human*

To evaluate the performance of complex discovery algorithms, we use reference complexes that have been manually validated via literature curation. For yeast, we use the CYC2008[40] set, which consists of 408 yeast complexes. For human, we use the CORUM[41] set, which consists of 1829 human complexes.

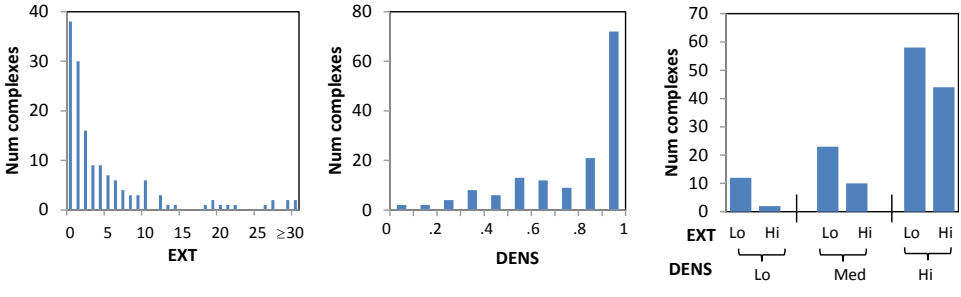
To investigate the performance of the clustering algorithms with respect to the three highlighted challenges, we stratify the reference complexes in terms of their sizes, extraneous edges, and densities. First, to quantify whether a complex is embedded within a highly-connected region of the PPIN, we derive EXT, the number of external proteins that are highly-connected to it, defined as being connected to at least half of the proteins in the complex. Second, to quantify how sparse a complex is, we derive DENS, the density of each complex, defined as the number of PPI edges in the complex divided by the total number of possible edges in the complex. In our analysis, we stratify the complexes into large and small complexes, and further stratify the large complexes into low, medium, and high DENS (corresponding to DENS of $[0, 0.35]$, $(0.35, 0.7]$, and $(0.7, 1]$, respectively), and low and high EXT (corresponding to EXT ≤ 3 and > 3, respectively), to give seven total strata (one for small complexes, and six for large complexes).

Figure 2 illustrates the size distribution of the yeast complexes, and the distributions of EXT, DENS, and our six analysis strata (stratified by EXT and DENS), among the large yeast complexes. Figure 3 shows the corresponding distributions for human complexes. In both yeast and human, the sizes of complexes follow the power-law distribution,[25] which highlights the important subtask of predicting small complexes (of size two and three): among both yeast and human complexes, about 60% are small complexes (259 out of 408 in yeast, 1029 out of 1829 in human).

Among large complexes in both yeast and human, about 40% of complexes have high EXT. We expect the prediction of these complexes to be extremely challenging, as it would be difficult to accurately delimit their borders from their highly-connected surroundings (the highly-connected external proteins are likely to be recruited into the predicted complexes). Only 10% of large complexes in yeast have low density. On the other hand, in human about 35% of large complexes are sparsely-connected with low DENS. We expect these sparsely-connected complexes to also be difficult to



(a) Size distribution



(b) Large complexes

Fig. 2. Statistics of the yeast reference complexes, from the CYC2008 database. (a) The size distribution of the complexes. (b) EXT (number of highly-connected external proteins) and DENS (density) distributions of large complexes.

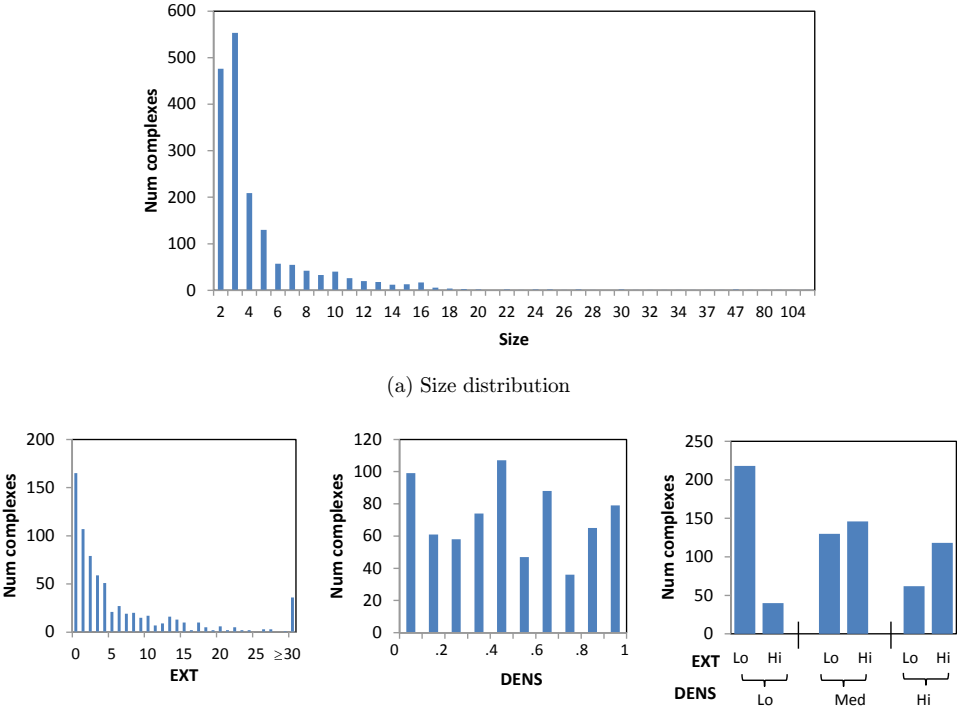
predict, as they do not form dense clusters that are picked out by most clustering algorithms.

4.3. *Evaluation methods*

We say that a cluster (i.e. a predicted complex) P matches a known complex C at a given match threshold match_thresh if Jaccard $(P, C) \geq$ match_thresh, where Jaccard (P, C) is the Jaccard similarity between the proteins contained in P and C:

$$\text{Jaccard}(P, C) = \frac{|V_P \cap V_C|}{|V_P \cup V_C|},$$

where V_X is the set of proteins contained in X. For large complexes, we use a stringent matching criteria of match_thresh $= 0.75$ in matching yeast complexes, and a rougher matching criteria of match_thresh $= 0.5$ in matching human complexes, as the latter task is much more difficult. For small complexes, we use the most stringent criteria of match_thresh $= 1$, as it is easier for a small cluster to match a small complex by chance. Given a set of clusters $\mathbf{P} = \{P_1, P_2, \ldots\}$, and a set of reference



(a) Size distribution

(b) Large complexes

Fig. 3. Statistics of the human reference complexes, from the CORUM database. (a) The size distribution of the complexes. (b) EXT (number of highly-connected external proteins) and DENS (density) distributions of large complexes.

complexes $\mathbf{C} = \{C_1, C_2, \ldots\}$, the precision and recall are calculated as:

$$\text{Precision} = \frac{|\{P_i \in \mathbf{P} | \exists\, C_j \in \mathbf{C}, P_i \text{ matches } C_j\}|}{|\mathbf{P}|},$$

$$\text{Recall} = \frac{|\{C_i \in \mathbf{C} | \exists\, P_j \in \mathbf{P}, P_j \text{ matches } C_i\}|}{|\mathbf{C}|}.$$

The precision–recall graph is another useful measure of performance, as it indicates the quality of predictions at different recall levels. It is obtained by applying varying thresholds on the predicted complexes' weighted densities, and plotting the precision and recall at each threshold. For brevity, we use the area under the precision–recall (AUPR) graph as a summarizing statistic (the actual precision-recall graphs are shown in the Supplementary Materials).

4.4. *Results*

Figures 4(a) and 4(b) show the performance of the clustering algorithms on prediction of large yeast and human complexes, at a finer matching level of

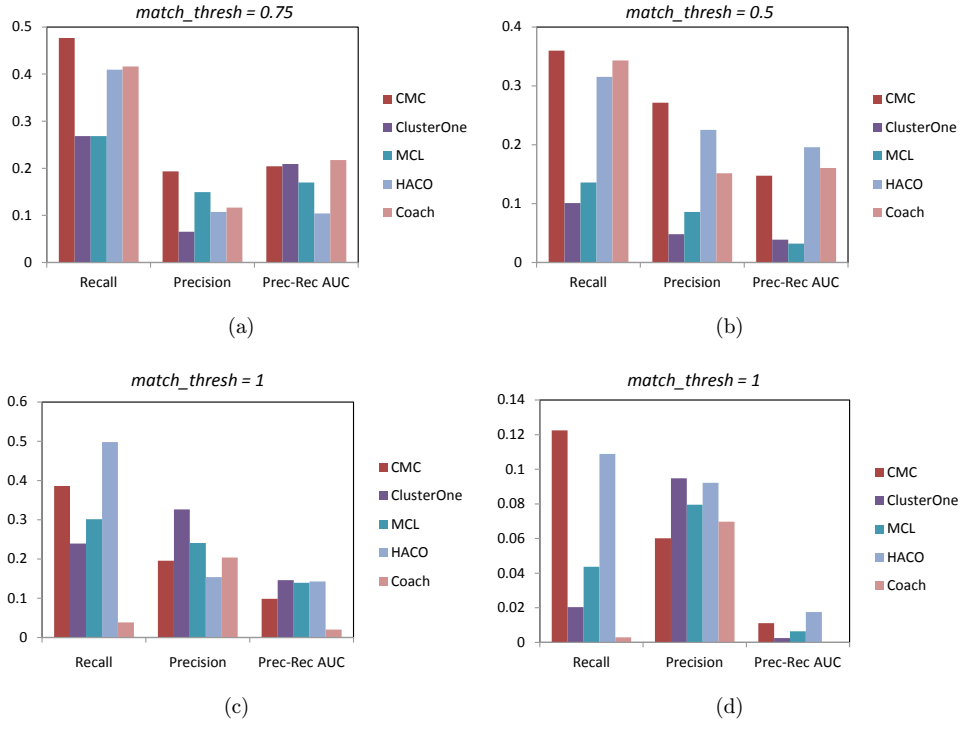


Fig. 4. Performance of the clustering algorithms on prediction of (a) large yeast complexes, (b) large human complexes, (c) small yeast complexes and (d) small human complexes.

match_thresh = 0.75 for yeast, and a rougher match_thresh = 0.5 for human (as prediction of human complexes is a more difficult task: at match_thresh = 0.75, the highest recall achieved is only about 10%). CMC, HACO, and Coach stand out with the highest recall levels for both yeast and human. In yeast, they achieve recalls over 40%, but CMC attains only 20% in precision, while HACO and Coach do worse with 10% precision. In human, even with a rougher matching criteria, CMC, HACO, and Coach achieve only over 30% recall, with 15% to 25% precision. MCL attains low recall in yeast and human, as it does not generate overlapping clusters. ClusterOne attains the lowest recall and precision, in both yeast and human.

Figures 4(c) and 4(d) show the performance of the clustering algorithms on the prediction of small yeast and human complexes, at a perfect matching requirement of match_thresh = 1.0. Coach predicts fewer than 5% of small yeast complexes, and almost no small human complexes at all, as it is problematic to define tightly connected cores with less-connected attachments when only two or three vertices are available. CMC and HACO again achieve high recall, but at the expense of generating many false positives, as they attain low precision. Again, MCL has low recall as it does not generate overlapping clusters. ClusterOne achieves the highest precision levels for small complexes, but at the expense of the lowest recalls. Note that the

performance for small human complexes is dismal: the best clustering algorithm only attains 12% recall.

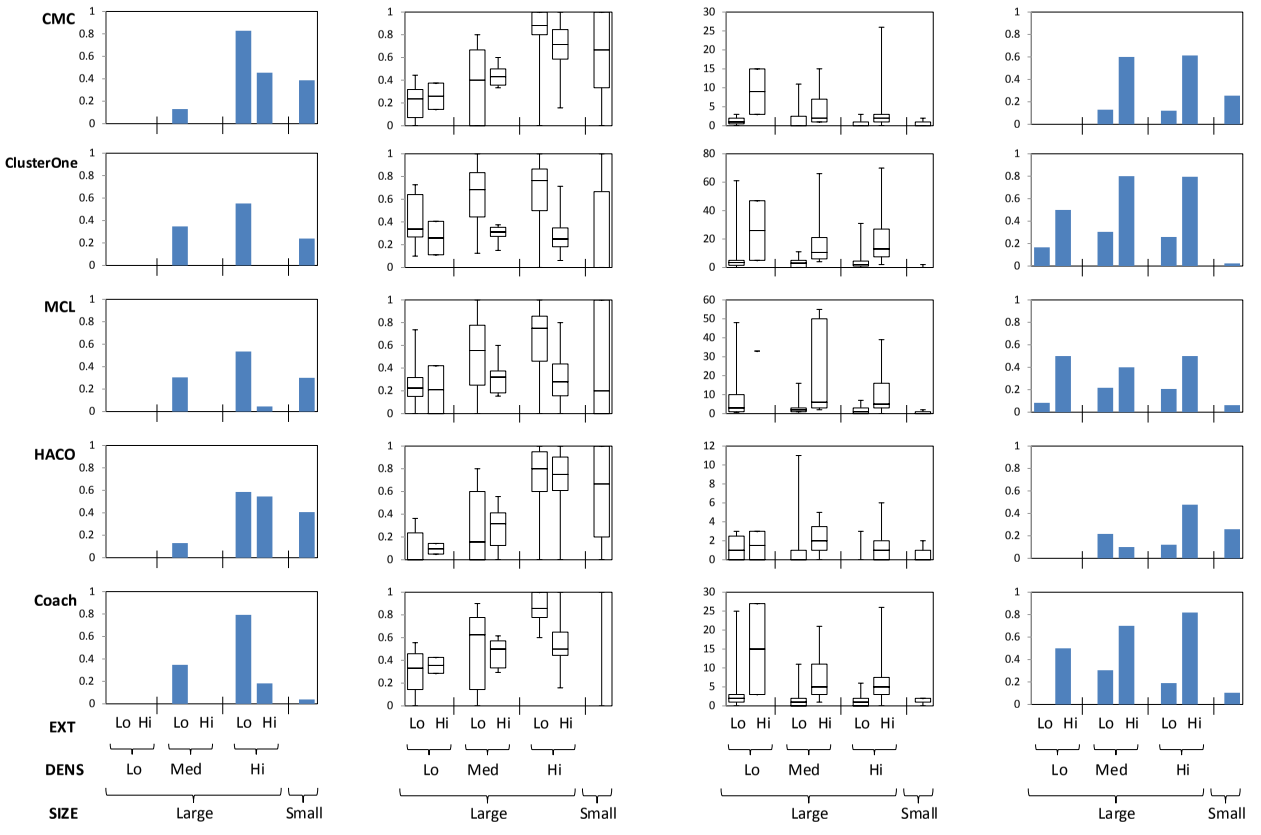
To investigate which complexes are problematic to predict, we study the performance of the complex discovery algorithms on the complexes stratified in terms of their sizes, extraneous edges, and densities. As described above, the complexes are stratified into small and large complexes, and large complexes are further stratified by density (DENS) and number of highly-connected external proteins (EXT), to give seven groups of complexes (see Figs. 2 and 3 for the distribution of size, DENS, and EXT of yeast and human complexes).

Figure 5(a) shows that yeast complexes with lower density are much harder to predict than those with higher density: no complex with low DENS are predicted at all by any clustering algorithm, while complexes with high DENS are predicted much more frequently. Furthermore, complexes with higher EXT are harder to predict than those with lower EXT: In each density strata, complexes with high EXT have lower recall than those with low EXT. Small complexes are also challenging to predict: Most clustering algorithms do not predict more than 40% of small complexes. As expected, the easiest complexes to predict are the large complexes with high DENS and low EXT.

Figure 5(b) shows that complexes with higher density can be predicted with better-matching clusters: Within each EXT strata, the match score increases with density. Furthermore, complexes with lower EXT are predicted with better-matching clusters: among complexes with medium or high DENS, match score is higher among those with low EXT than high EXT.

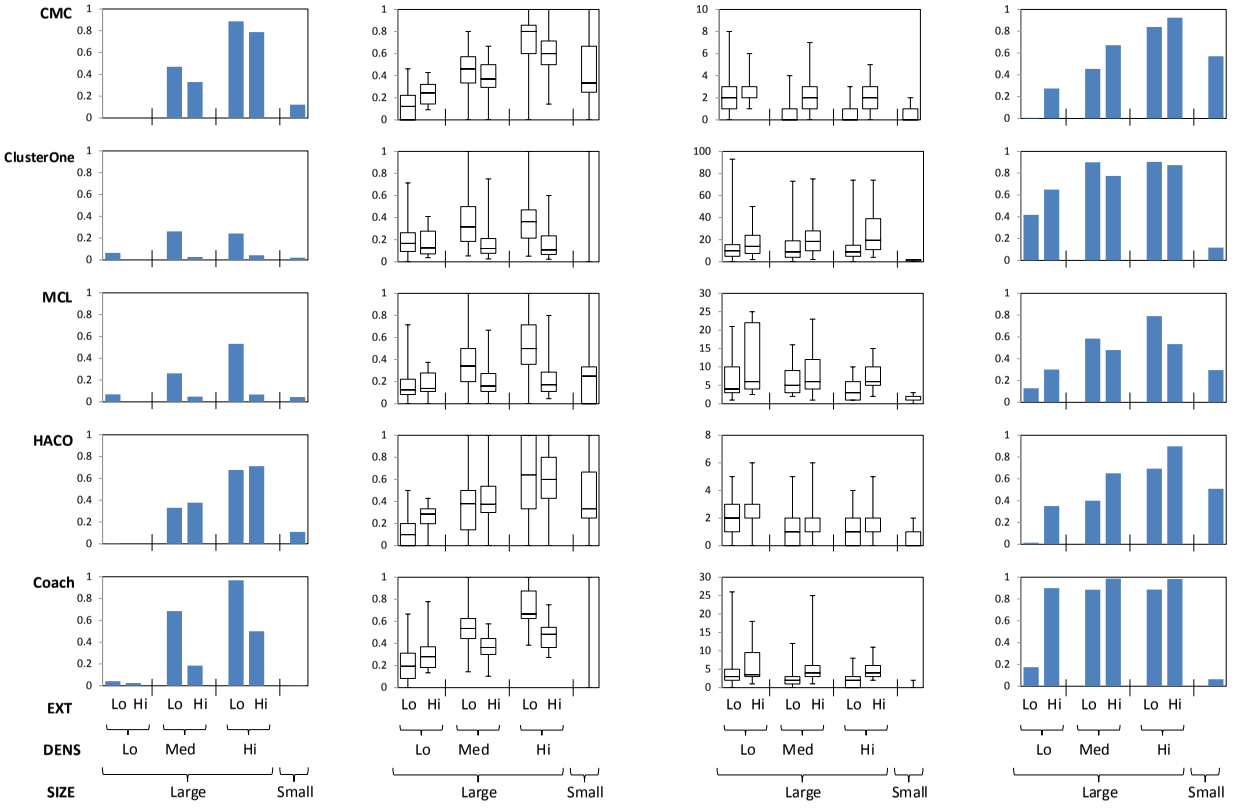
Figures 5(c) and 5(d) reveal why complexes with higher EXT are difficult to predict. Figure 5(c) shows that clustering algorithms tend to include many extraneous proteins when predicting complexes with higher EXT: Across all DENS strata, complexes with higher EXT have greater number of extra proteins in their best-matched clusters (intuitively, the extraneous proteins are likely to be those highly-connected external proteins). Figure 5(d) shows that clustering algorithms tend to merge together complexes with higher EXT: Across all DENS stratas, complexes with higher EXT tend to be found in clusters merged with other complexes.

Figure 6 shows the corresponding performance of the clustering algorithms on the stratified human complexes. Similar conclusions can be drawn here as from yeast complexes. Small complexes are challenging to predict, with most clustering algorithms predicting less than 10% of them. Complexes with lower density are harder to predict than those with higher density, and are predicted with clusters that match them less well; likewise, complexes with higher EXT are also harder to predict than those with lower EXT, and are also predicted with clusters that match them less well (Figs. 6(a) and 6(b)). However, Fig. 6(b) shows that, within the low-DENS stratum, complexes with high EXT attain slightly higher match scores than those with low EXT, because these low-density complexes with high EXT are likely to slightly overlap with clusters consisting of complex proteins with the external proteins that



(a) Recall (b) Match score of best cluster (c) Extra proteins in best cluster (d) Merged complexes

Fig. 5. Performance of complex discovery algorithms on yeast complexes, stratified by size, DENS, and EXT. The x-axis of each chart corresponds to the different stratified groups of complexes, given at the bottom of the figure.



(a) Recall (b) Match score of best cluster (c) Extra proteins in best cluster (d) Merged complexes

Fig. 6. Performance of complex discovery algorithms on human complexes, stratified by size, DENS, and EXT. The x-axis of each chart corresponds to the different stratified groups of complexes, given at the bottom of the figure.

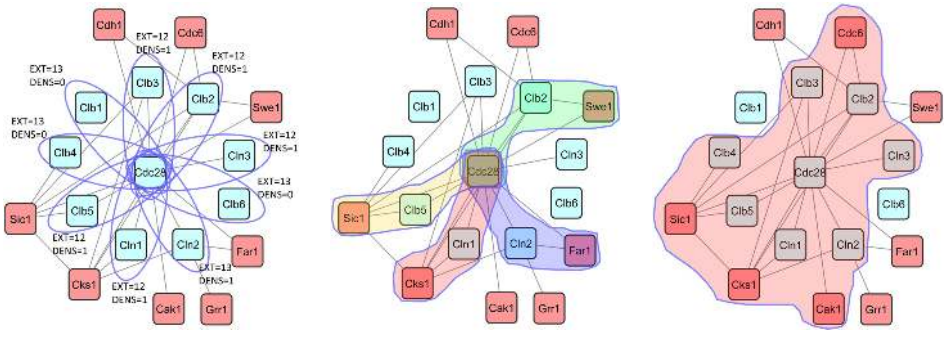
they are highly-connected to; indeed, in these cases the match scores are mostly under 0.5.

Figure 6(c) shows that, as in yeast, human complexes with high EXT are predicted with clusters that include many more extraneous proteins. Figure 6(d) shows that complexes with higher EXT tend to be merged together in clusters.

4.5. *Example complexes*

Here we highlight some example complexes that are known to behave dynamically, and show how their static interactomes exhibit characteristics (such as high EXT and low DENS) which result from their static representation, and which make them difficult to predict.

The Cdc28p yeast protein, as described above, complexes with various cyclin proteins (Cln1p to Cln3p, Clb1p to Clb6p) to regulate the cell-cycle. The proper complexes are formed at each point of the cell-cycle via gene-expression and post-translational controls.[9,10] Figure 7(a) shows the interactome around these proteins and their neighbors, with the nine different complexes formed by Cdc28p circled. Although these interactions occur at different times during the cell-cycle, they are collapsed into the same static interactome, resulting in a highly-connected region around Cdc28p and its cyclin partners: note that the EXT for each of the complexes range from 12 to 13. Furthermore, PPIs are missing between Cdc28p and some of its cyclin partners, giving a density of 0 to these complexes. In fact, these PPIs exist in our source datasets, but with slightly fewer experimental evidences to back them up compared to other Cdc28p PPIs; thus they scored slightly lower in reliability and they were filtered from our PPIN. While it is possible to lower our reliability score cutoff to include these PPIs, this would also include many spurious PPIs and make the discovery of other complexes even more difficult.

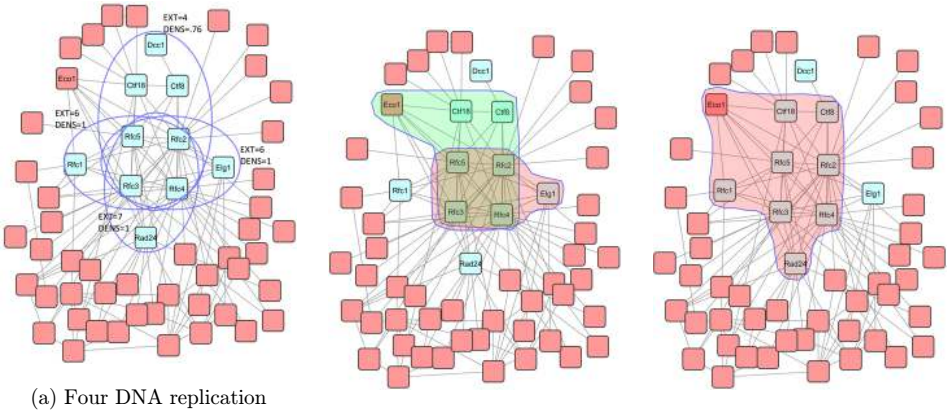


(a) Nine complexes formed by Cdc28p

(b) Clusters from CMC

(c) Cluster from MCL

Fig. 7. (a) Cdc28p is involved in nine distinct complexes, which overlap and have many highly-connected external proteins (EXT). Three of the complexes are disconnected (DENS = 0). (b) CMC includes extraneous proteins in its clusters. (c) MCL merges the complexes.



(a) Four DNA replication factor complexes

(b) Clusters from CMC

(c) Cluster from MCL

Fig. 8. (a) A common core is shared among four DNA replication factor complexes, which contributes to a high number of external proteins (EXT) in each complex. (b) CMC finds only one of the four complexes. (c) MCL merges three of the four complexes.

Figures 7(b) and 7(c) show the clusters predicted by CMC and MCL, respectively. CMC found four clusters that overlap with four Cdc28p complexes, but with one extraneous protein in each case, while MCL found one large cluster that covered Cdc28p, seven of the nine cyclin proteins, and four extraneous proteins.

The four Replication Factor C (RFC) complexes in yeast are structurally similar complexes involved in DNA metabolism. Each of these complexes consist of a core of four subunits (Rfc2p to Rfc5p) and distinct attachment proteins, and perform different biological functions related to DNA metabolism.[42] The interactome of the RFC complexes and their neighbors are shown in Fig. 8(a), with the four complexes circled. Here again, conflating the four distinct complexes in the static interactome results in many extraneous edges and high connectivity to proteins outside each complex: the EXT for the four complexes range from 4 to 7.

Figures 8(b) and 8(c) show the clusters predicted by CMC and MCL, respectively. CMC predicted one of the RFC complexes perfectly, while predicting a second cluster that matched another complex less well; MCL predicted a large cluster that overlapped with three of the RFC complexes.

Note that MCL does not allow overlaps in its predicted clusters, so in the above examples it predicts clusters that merge the overlapping and highly-connected complexes together. While CMC allows overlapping clusters, the many extraneous edges and high connectivity to external proteins make it difficult to delimit the overlapping complexes precisely.

5. Discussion and a Call to Arms

Protein interactions behave in a dynamic fashion, with a variety of interaction timings, locations, and affinities. The cellular control of this dynamism gives

important functional mechanisms to protein complexes, allowing complexes to assemble at specific times, or to vary in composition to activate or modulate their functions. Interaction detection technologies are limited in their ability to capture such dynamics; furthermore, this dynamism also impedes accurate and comprehensive screening of interactions. Moreover, the representation of interactions in a PPIN does not preserve any information about interaction dynamism, allowing only a static analysis of a dynamic reality.

In Sec. 3, we identified three challenges in complex prediction that result from, and are exacerbated by, the analysis of the static interactome to derive complexes that behave dynamically in nature. First, many proteins participate in multiple complexes, leading to overlapping complexes embedded within highly-connected regions of the PPIN with many extraneous edges connecting them to external proteins. This makes it difficult to accurately delimit the boundaries of such complexes. Second, many condition- and location-specific PPIs are not detected, leading to sparsely-connected complexes that cannot be picked out by clustering algorithms. Third, the majority of complexes are small complexes (made up of two or three proteins), which are extra sensitive to the effects of extraneous edges and missing co-complex edges.

In Sec. 4, we presented results of five clustering algorithms for prediction of large and small complexes in yeast and human, and showed that only complexes with high density and few highly-connected external proteins can be consistently predicted: more than 80% of such large complexes can be predicted in yeast and human (with match_thresh = 0.75 and 0.5, respectively), and more than 60% of such small complexes can be predicted in yeast and human (with match_thresh = 1). Complexes with low density frequently could not be predicted at all, while those with many highly-connected external proteins tended to be predicted in clusters with many extraneous proteins or merged complexes. Furthermore, small complexes with such characteristics are especially challenging to predict, particularly in human for which recall rates are extremely low.

Drawing on our insight into the causes of these challenges, we suggest a few approaches for addressing them.

First, to address the problem of complexes in highly-connected regions with many extraneous edges, Liu *et al.*[43] proposed a technique to decompose the PPIN into spatially- and temporally-coherent subnetworks. First, hub proteins with large numbers of interaction partners are removed before complex discovery, as they tend to correspond to date hubs with non-simultaneous interactions. Next, cellular-location Gene Ontology terms[44] are used to decompose the PPIN into spatially-coherent subnetworks. By splitting dense regions of the PPIN into less-dense but coherent subnetworks, complex-discovery performance is improved, with the biggest improvements among complexes in highly-connected regions. Another reasonable idea to tackle this problem is to make use of information such as protein domains to identify non-simultaneous interactions, which can be used to improve complex discovery. For example, Jung *et al.*[22] decomposed the PPIN into subnetworks of

simultaneous interactions, from which temporally-coherent complexes can be extracted; while Ozawa *et al.*[23] refined predicted complexes by eliminating those with non-simultaneous interactions.

Second, to address the problem of sparse complexes, an approach like Supervised Weighting of Composite Networks (SWC[45]) is promising. It integrates PPI data with additional data sources — in particular, functional associations and co-occurrence in literature — using a supervised approach to weight edges with their posterior probability of belonging to a complex. By integrating diverse data sources that may support co-complex relationships between proteins, SWC fills in the missing edges in many sparse complexes, while reducing the amount of spurious non-co-complex edges. Using this approach, improvements are obtained in both precision and recall for yeast and human complex discovery, especially among the sparse complexes.

Third, to address the problem of predicting small complexes, an approach like Size-Specific Supervised weighting (SSS[46]) is promising. It integrates PPI data with two additional data sources, functional associations and co-occurrence in literature, along with their topological features, using a supervised approach to weight edges with their posterior probabilities of belonging to small complexes versus large complexes. SSS then extracts small complexes from the weighted network, and scores them using the probabilistic weights of edges within, as well as surrounding, the complexes. This approach achieves significant improvements in precision and recall in discovering small complexes.

While these approaches perform better than the clustering algorithms surveyed here, there remains much room for improvement, especially for the prediction of human complexes where high levels of accuracy and resolution are still unattainable. We hope that this paper is able to call the bioinformatics community into action to address these highlighted problems, as well as the more general challenge of predicting dynamic protein complexes.

Acknowledgments

This work is supported in part by a Singapore Ministry of Education grant MOE2012-T2-1-061 and a National University of Singapore NGS scholarship.

References

1. Nooren IMA, Thornton JM, Diversity of protein-protein interactions, *EMBO J* **22** (14):3486–3492, 2003.
2. Li X, Wu M, Kwoh CK, Ng SK, Computational approaches for detecting protein complexes from protein interaction networks: A survey, *BMC Genomics* **11**(Suppl 1):S3, 2010.
3. Srihari S, Leong HW, A survey of computational methods for protein complex prediction from protein interaction networks, *J Bioinform Comput Biol* **11**(2):1230002, 2013.

4. Chen B, Fan W, Liu J, Wu FX, Identifying protein complexes and functional modules — From static PPI networks to dynamic PPI networks, *Brief Bioinform* **15**(2):177–194, 2014.
5. Jones S, Thornton JM, Principles of protein-protein interactions, *Proc Natl Acad Sci USA* **93**(1):13–20, 1996.
6. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C, Transient protein-protein interactions: Structural, functional, and network properties, *Structure* **18**(10):1233–1243, 2010.
7. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M, Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature* **430**:88–93, 2004.
8. Batada NN, Hurst LD, Tyers M, Evolutionary and physiological importance of hub proteins, *PLoS Comput Biol* **2**(7):e88, 2006.
9. Mendenhall A, Hodge A, Regulation of cdc28 cyclin-dependent protein kinase activity during the cell-cycle of the yeast *Saccharomyces cerevisiae, Microbiol Mol Biol R* **62** (4):1191–1243, 1998.
10. Enserink JM, Kolodner RD, An overview of Cdk1-controlled targets and processes, *Cell Div* **5**(11), 2010.
11. de Lichtenberg U, Jensen LJ, Brunak S, Bork P, Dynamic complex formation during the yeast cell-cycle, *Science* **307**(5710):724–747, 2005.
12. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B *et al.*, Proteome survey reveals modularity of the yeast cell machinery, *Nature* **440**:631–636, 2006.
13. Fields S, Song O, A novel genetic system to detect protein-protein interactions, *Nature* **340**(6230):245–246, 1989.
14. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U, Yeast two-hybrid, a powerful tool for systems biology, *Int J Mol Sci* **10**(6):2763–2788, 2009.
15. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B, A generic protein purification method for protein complex characterization and proteome exploration, *Nat Biotechnol* **17**(10):1030–1032, 1999.
16. Gavin AC, Maeda K, Kühner S, Recent advances in charting protein-protein interaction: Mass spectrometry-based approaches, *Curr Opin Biotech* **22**:42–49, 2011.
17. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS, Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae, Mol Cell Proteomics* **6**(3):439–450, 2007.
18. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP *et al.*, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature* **440**:637–643, 2006.
19. Bisson N, James DA, Ivosev G, Tate SA, Bonner R, Taylor L, Pawson T, Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor, *Nat Biotechnol* **29**(7):653–658, 2011.
20. Collins BC, Gillet LC, Rosenberger G, Röst HL, Vichalkovski A, Gstaiger M, Aebersold R, Quantifying protein interaction dynamics by SWATH mass spectrometry: Application to the 14-3-3 system, *Nat Methods* **10**(12):1246–1253, 2013.
21. Srihari S, Leong HW, Temporal dynamics of protein complexes in PPI networks: A case study using yeast cell-cycle dynamics, *BMC Bioinformatics* **13**(Suppl 17):S16, 2005.
22. Jung SH, Hyun B, Jang WH, Hur HY, Han DS, Protein complex prediction based on simultaneous protein interaction network, *Bioinformatics* **26**(3):385–391, 2010.

23. Ozawa Y, Saito R, Fujimori S, Kashima H, Ishizaka M, Yanagawa H, Miyamoto-Sato E, Tomita M, Protein complex prediction via verifying and reconstructing the topology of domain-domain interactions, *BMC Bioinformatics* **11**:350, 2010.
24. Ideker T, Krogan NJ, Differential network biology, *Mol Syst Biol* **8**:565, 2012.
25. Tatsuke D, Maruyama O, Sampling strategy for protein complex prediction using cluster size frequency, *Gene* **518**(1):152–158, 2012.
26. Adamcsek B, Palla G, Farkas I, Derenyi I, Vicsek T, CFinder: Locating cliques and overlapping modules in biological networks *Bioinformatics* **22**(8):1021–1023, 2006.
27. Li M, Chen J, Wang J, Hu B, Chen G, Modifying the DPClus algorithm for identifying protein complexes based on new topological structures, *BMC Bioinformatics* **9**:398, 2008.
28. Przulj N, Wigle DA, Functional topology in a network of protein interactions, *Bioinformatics* **20**(3):340–348, 2003.
29. Widita CK, Maruyama O, PPSampler2: Predicting protein complexes more accurately and efficiently by sampling, *BMC Syst Biol* **7**(Suppl 6):14, 2013.
30. Srihari S, Ning K, Leong HW, MCL-CAw: A refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure, *BMC Bioinformatics* **11**:504, 2010.
31. Liu G, Wong L, Chua HN, Complex discovery from weighted PPI networks, *Bioinformatics* **25**(15):1891–1897, 2009.
32. Nepusz T, Yu H, Paccanaro A, Detecting overlapping protein complexes in protein-protein interaction networks, *Nat Methods* **9**:471–472, 2012.
33. van Dongen S, Graph clustering by flow simulation, PhD Thesis, University of Utrecht, 2000.
34. Wang H, Kakaradov B, Collins SR, Karotki L, Fiedler D, Shales M, Shokat KM, Walther TC, Krogan NJ, Koller D, A complex-based reconstruction of the Saccharomyces cerevisiae interactome, *Mol Cell Proteomics* **8**(6):1361–1381, 2009.
35. Wu M, Li X, Kwoh CK, Ng SK, A core-attachment based method to detect protein complexes in PPI networks, *BMC Bioinformatics* **10**:169, 2009.
36. Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, ODonnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M, The BioGRID interaction database: 2013 update, *Nucleic Acids Res* **41**(Database Issue):D816–D823, 2013.
37. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, del Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H, The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases, *Nucleic Acids Res* **42**(Database Issue):D358–D363, 2014.
38. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, Castagnoli L, Cesareni G, MINT, the molecular interaction database: 2012 update, *Nucleic Acids Res* **40**(Database Issue):D857–D861, 2012.
39. Chua HN, Sung WK, Wong L, An efficient strategy for extensive integration of diverse biological data for protein function prediction, *Bioinformatics* **23**(24):3364–3373, 2007.
40. Pu S, Wong J, Turner B, Cho E, Wodak SJ, Up-to-date catalogues of yeast protein complexes, *Nucleic Acids Res* **37**(3):825–831, 2009.
41. Ruepp A, Waegele B, Lechner M, Brauner B, Dk I, Fobo G, Frishman G, Montrone C, Mewes H, CORUM: The comprehensive resource of mammalian protein complexes–2009, *Nucleic Acids Res* **38**:D497–D501, 2010.

42. Bylund GO, Majka J, Burgers PMJ, Overproduction and purification of RFC-related clamp loaders and PCNA-related clamps from *Saccharomyces cerevisiae, Metholds Enzymol* **409**:1–11, 2006.
43. Liu G, Yong CH, Chua HN, Wong L, Decomposing PPI networks for complex discovery, *Proteome Sci* **9**(Suppl 1):S15, 2011.
44. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*, Gene ontology: Tool for the unification of biology, *Nat Genet* **25**:25–29, 2000.
45. Yong CH, Liu G, Chua HN, Wong L, Supervised maximum-likelihood weighting of composite protein networks for complex prediction, *BMC Syst Biol* **6**(Suppl 2):S13, 2012.
46. Yong CH, Maruyama O, Wong L, Discovery of small protein complexes from PPI networks with size-specific scoring, *BMC Syst Biol* **8**(Suppl 5):S3, 2014.

Chern Han Yong received his BS and MS degrees in Computer Science from the University of Texas at Austin. He is currently working towards a Ph.D. at the National University of Singapore. His main research interests include computational biology, artificial intelligence, and machine learning.

Limsoon Wong is a Professor in the School of Computing at the National University of Singapore. He currently works mostly on knowledge discovery technologies and their application to biomedicine. He is a Fellow of the ACM, named in 2013 for his contributions to database theory and computational biology. He serves on the editorial boards of several journals, including Journal of Bioinformatics and Computational Biology.