ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Methods for protein complex prediction and their contributions towards understanding the organisation, function and dynamics of complexes

Sriganesh Srihari [a,*], Chern Han Yong [b], Ashwini Patil [c], Limsoon Wong [b]

[a] Institute for Molecular Bioscience, The University of Queensland, St. Lucia, Queensland 4067, Australia
[b] Department of Computer Science, National University of Singapore, Singapore 117417, Singapore
[c] Human Genome Centre, The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

ARTICLE INFO

Article history:
Received 19 March 2015
Revised 14 April 2015
Accepted 14 April 2015
Available online 23 April 2015

Edited by Wilhelm Just

Keywords:
Protein complex prediction
PPI network
Dynamic and fuzzy complexes
Complexes in diseases

ABSTRACT

Complexes of physically interacting proteins constitute fundamental functional units responsible for driving biological processes within cells. A faithful reconstruction of the entire set of complexes is therefore essential to understand the functional organisation of cells. In this review, we discuss the key contributions of computational methods developed till date (approximately between 2003 and 2015) for identifying complexes from the network of interacting proteins (PPI network). We evaluate in depth the performance of these methods on PPI datasets from yeast, and highlight their limitations and challenges, in particular at detecting sparse and small or sub-complexes and discerning overlapping complexes. We describe methods for integrating diverse information including expression profiles and 3D structures of proteins with PPI networks to understand the dynamics of complex formation, for instance, of time-based assembly of complex subunits and formation of fuzzy complexes from intrinsically disordered proteins. Finally, we discuss methods for identifying dysfunctional complexes in human diseases, an application that is proving invaluable to understand disease mechanisms and to discover novel therapeutic targets. We hope this review aptly commemorates a decade of research on computational prediction of complexes and constitutes a valuable reference for further advancements in this exciting area.

1. Introduction

Most biological processes within cells are carried out by proteins that physically interact to form stoichiometrically stable *complexes*. Even in the relatively simple model organism *Saccharomyces cerevisiae* (budding yeast), these complexes are composed of several subunits that work in a concerted manner. These complexes interact with individual proteins and other complexes to form functional modules and signalling pathways that drive the cellular machinery. Therefore, a faithful reconstruction of the entire set of complexes is essential not only to understand complex formation but also the higher level functional organisation of cells.

High-throughput experimental systems including yeast two-hybrid (Y2H), tandem affinity purification followed by mass spectrometry (TAP-MS) and protein complementation assay (PCA) have mapped a considerable fraction of interactions from model organisms including *S. cerevisiae* [1–7], *Drosophila melanogaster* [8,9] and *Caenorhabditis elegans* [10], thereby fuelling computational methods to systematically analyse these large-scale interaction data. Beginning from classical methods by Spirin and Mirny [11] and Bader and Hogue [12] that work primarily by clustering the network of protein interactions (PPI network), computational methods have come a long way, and current methods integrate diverse information with PPI networks to predict complexes. These methods have been tested extensively on data from model organisms [13,14], and are currently being extended to identify and catalogue complexes from less extensively mapped organisms including *Homo sapiens* [15].

Protein complexes represent *modular* functional units within the PPI network [11]. From a biological perspective, this modularity ensures division of labour and provides robustness against mutation and chemical attacks [16]. From a topological perspective, this modularity represents densely connected groups of proteins that function as complexes [17]. Most methods identify complexes by mining modular or dense subnetworks from PPI networks. While this general strategy looks straightforward, these methods are severely restricted by limitations in existing PPI datasets, in particular by the lack of sufficient interactions between "complexed" proteins and the presence of a large number of

* Corresponding author.
E-mail address: s.srihari@uq.edu.au (S. Srihari).

http://dx.doi.org/10.1016/j.febslet.2015.04.026
0014-5793/© 2015 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

false-positive (noisy) interactions [18,19]. Therefore, increasing the interaction coverage by integrating PPI datasets from multiple studies and reducing the noise by assessing the reliabilities of interactions (scoring of PPIs) [20–22] are crucial for accurate complex detection. To summarise, computational identification of complexes from experimental datasets involves the following steps:

(i) Integrating interactions from multiple experiments and assessing the reliabilities of these interactions;
(ii) Constructing a reliable PPI network using only the high-confidence interactions;
(iii) Identifying modular subnetworks from the PPI network to generate a candidate list of complexes; and
(iv) Evaluating these candidate complexes against *bona fide* complexes and validating and assigning roles for novel complexes.

Over the last 10 years, more than 20 different methods have been proposed in the literature for complex prediction from PPI networks. From time-to-time, surveys have evaluated these methods on datasets available at the time. For example, one of the earliest comprehensive evaluation of methods were by Brohee and van Helden [23] and Vlasblom and Wodak [24], who compared these methods on yeast Y2H datasets. Subsequently, Li et al. [13] and Srihari and Leong [14] presented a more exhaustive evaluation by including raw and scored yeast datasets from TAP-MS and PCA studies [6,7,25]. More recently, Yong and Wong [26] studied these methods specifically for the deconvolution of overlapping complexes in dense regions of PPI networks, the recovery of complexes in sparse regions of PPI networks, and the recovery of small complexes in PPI networks. With increasing coverage for human PPI datasets [27–30] these methods are now being applied to predict human complexes [15].

The resources for *bona fide* complexes against which predicted complexes are evaluated have also expanded over the years. For example, the Munich Information Centre for Protein Sequences (MIPS) (http://mips.helmholtz-muenchen.de/proj/ppi/) [31] and the Curated Yeast Complexes (CYC) 2008 (http://wodaklab.org/cyc2008/) [32] databases contain more than 400 experimentally validated complexes for yeast, whereas COmprehensive ResoUrce of Mammalian protein complexes (CORUM) (http://mips.helmholtz-muenchen.de/genre/proj/corum) [33] contains over 2000 validated 'core' mammalian complexes. Predicted complexes that have been subsequently validated have in turn contributed several novel complexes to these catalogues (*e.g.* http://human.med.utoronto.ca/) [15].

The ability to predict complexes from multiple species makes it is possible to examine the reorganisation and rewiring of complexes between these species, and thereby estimate the evolutionary conservation of complexes [9]. This could potentially have far-reaching implications, for example, in translating therapeutically relevant observations from model organisms to human [34,35]. For example, Nguyen et al. [36] note that rewiring and reorganisation of complexes from yeast to human can affect the transfer of *synthetic lethality* (SL) relationships between genes identified in yeast [37] to human; SL relationships are of therapeutic value in the context of human cancers [38].

Similarly, complexes predicted across disease conditions have revealed extensive rewiring (differential wiring) between these conditions, thereby highlighting key targetable avenues for these diseases [39]. By focusing on rewiring within complexes rather than of the entire PPI network definite dysfunctional regions could be located, thus identifying therapeutically targetable proteins.

Considering the valuable contributions of complex prediction methods, here we put together an extensive survey of methods developed to date (approximately between 2003 and 2015) and evaluate their performance on yeast PPI datasets. We build on earlier surveys [13,14,26] so as not to entirely repeat their findings, but discuss challenges faced by methods more lately, in particular detection of sub- or small and sparse complexes and discerning of overlapping complexes. We discuss these methods in the context of evolutionary conservation of complexes between species. By covering methods that integrate diverse information including gene expression and 3D structures of proteins with PPI networks, we discuss the dynamics of complex formation. Finally, we describe methods to identify dysfunctional complexes in human diseases, an application that is proving invaluable to understand disease mechanisms and to discover novel therapeutic targets.

2. Review of methods for complex prediction from PPI networks

Although in general, most methods rely on the assumption that protein complexes are embedded as densely connected proteins within the PPI network, these methods vary considerably in their algorithmic strategies and auxiliary biological information employed to identify complexes. Accordingly, these methods have been classified (Table 1) [14] as (i) those based solely on PPI network topology; and (ii) those based on PPI network topology and additional biological insights. By incorporating functional, structural, organisational or temporal information, these methods overcome some of the limitations of experimental datasets, in particular the presence of noise, thereby improving complex prediction. Several of these methods are available as easy-to-run command-line programs or *Cytoscape* [40] plug-ins (Table 1).

To begin, a PPI network is modelled as an undirected graph $G = (V,E)$ where V is the set of proteins and $E = \{(u,v)\colon u,\ v \in V\}$ is the set of interactions between these proteins. For a protein $v \in V$, the set of neighbours of v is $N(v)$ and the *degree* of v is $deg(v) = |N(v)|$. The *interaction density* of a subgraph $G'(V',E')$ of G is $\frac{2|E'|}{|V'|.(|V'|-1)}$.

2.1. Methods based solely on network clustering

Methods based solely on the topology of PPI network look for dense subnetworks or clusters in the network to identify candidate complexes. While some of these methods adopt an agglomerative approach by beginning with singleton or small sets of proteins and growing these sets based on certain cost criteria, some others adopt a partitioning approach by repeatedly breaking down larger clusters into smaller clusters.

2.1.1. Molecular COmplex Detection (MCODE)

MCODE [12] is one of the first computational methods for predicting complexes from PPI networks. MCODE adopts an agglomerative approach that works in three stages: protein (vertex) weighting, complex extraction and an optional post-processing of complexes.

In the first stage, each protein v in the network $G = (V,E)$ is weighted based on the core-clustering density of v, which is measured as the clustering coefficient of the highest k-core in the neighbourhood of v. In the second stage, the protein s with the highest clustering density is used to seed a complex. MCODE then recursively moves outward from s by including proteins into the complex whose weights are a given percentage (vertex weight parameter) away from that of s. This process stops when there are no more proteins to be added to the complex. If there are seed vertices still available, new complexes are seeded and expanded in a similar manner. The optional third stage performs a post-processing by including proteins from the neighbourhood regions

Table 1
Methods for protein complex prediction from protein interaction networks. Associated softwares are available as *Cytoscape* [40] plug-ins (Cy), command line programs (CL) or as online (OL) web servers under the mentioned links.

Classification	Method	Availability (URL)	Reference
Solely network clustering	MCODE (Cy)	http://apps.cytoscape.org/apps/mcode	[12]
	MCL (Cy, CL)	http://micans.org/mcl/r http://apps.cytoscape.org/apps/clustermake	[41–43]
	CMC (CL)	https://www.comp.nus.edu.sg/~wongls/projects/complexprediction/CMC-26may09/	[45]
	ClusterONE (Cy)	http://apps.cytoscape.org/apps/clusterone	[49]
	HACO (CL)	http://www.bio.ifi.lmu.de/Complexes/ProCope/	[44,50]
	PPSampler (CL)	http://imi.kyushu-u.ac.jp/~om/PPSamplerVer1.2/PPSamplerVer1_2.exe	[72]
Core-attachment structure	CORE (CL)	http://alse.cs.hku.hk/complexes/	[54]
	COACH (CL)	http://www1.i2r.a-star.edu.sg/~xlli/coach.zip	[55]
	MCL-CAw (CL)	https://sites.google.com/site/mclcaw/	[56,57]
Functional information	RNSC (CL)	http://www.cs.utoronto.ca/~juris/data/ppi04/	[59]
	PCP (CL)	https://www.comp.nus.edu.sg/~wongls/projects/complexprediction/PCP-3aug07/	[61]
Evolutionary information	NetworkBLAST (OL)	http://www.cs.tau.ac.il/~bnet/networkblast.htm	[79,149]
	NetworkBLAST-M (CL)	http://www.cs.tau.ac.il/~bnet/License-nbm.htm	[79,149]
	COCIN (CL)	https://sites.google.com/site/cocinhy/	[36]
Mutual exclusive interactions	SPIN (CL)	https://code.google.com/p/simultaneous-pin/	[68]
	DACO (CL)	http://sourceforge.net/projects/dacoalgorithm/	[70]
Sparse complexes	SWC (CL)	http://www.comp.nus.edu.sg/~wongls/projects/complexprediction/SWC-31oct14	[53]
Small complexes	SSS (CL)	http://www.comp.nus.edu.sg/~wongls/projects/complexprediction/sss-3dec2014.zip	[75]
Temporal complexes	TS-OCD (CL)	http://mail.sysu.edu.cn/home/stsddq@mail.sysu.edu.cn/dai/others/TSOCD.zip	[99]
Complexes in diseases	CONTOUR (CL)	https://sites.google.com/site/contourv1/	[39]

of complexes using a "fluff" parameter: neighbouring proteins whose clustering density is higher than this parameter are included into the complexes. The resultant complexes are then scored and ranked based on their weighted densities.

2.1.2. Markov Clustering (MCL)

MCL [41] is a fast, highly scalable graph clustering method. Applied initially to cluster protein sequences [42], MCL has proved effective for clustering large PPI networks due to its scalability [43,44].

MCL works by simulating random walks (called a *flow*) to extract dense regions from the network. To simulate the flow, MCL iteratively manipulates the adjacency matrix of the network using two operators, *expansion* and *inflation*, that control the spread and thickness of the flow, respectively. Expansion enables the flow to reach all regions of the network, whereas inflation controls the contraction of the flow, making the flow thicker in dense regions and thinner in sparse regions. In each iteration, these parameters increase the probabilities for the random walks within clusters (intra-cluster walks) and decrease the probabilities for the walks between clusters (inter-cluster walks). This process progressively separates out dense regions within the network, ultimately identifying non-overlapping clusters from the network. Since the entire process is executed as matrix operations, MCL is fast and scalable even to large networks.

2.1.3. Clustering based on merging Maximal Cliques (CMC)

CMC [45] works by repeated merging of maximal cliques extracted from the PPI network. CMC incorporates reliability scores for PPIs and therefore improves on earlier clique-merging methods, including CFinder [46] and Local Clique Merging Algorithm (LCMA) [47], that work only on unscored networks.

CMC begins by enumerating all maximal cliques in the PPI network using the fast search-space pruning-based Cliques algorithm [48]. Each clique C is assigned a score which is the weighted interaction density of C, given by $\frac{\sum_{u,v\in C} w(u,v)}{|C|.(|C|-1)}$. Cliques are ranked in non-increasing order of their weighted densities. CMC then iteratively merges highly overlapping cliques based on the extent of their inter-connectivity. The inter-connectivity I (C_1, C_2) between two cliques C_1 and C_2 is given by:

$$I(C_1, C_2) = \sqrt{\frac{\sum_{u\in(C_1-C_2)}\sum_{v\in(C_2)} w(u,v)}{|C_1 - C_2| \cdot |C_2|} \cdot \frac{\sum_{u\in(C_2-C_1)}\sum_{v\in(C_1)} w(u,v)}{|C_2 - C_1| \cdot |C_1|}}$$

If $I(C_1, C_2) \geqslant T_m$, a merge threshold, then C_2 is merged with C_1, or C_2 is simply removed if it overlaps significantly with $C_1 : |C_1 \cap C_2|/|C_2| \geq T_o$, an overlap threshold. Finally, all merged clusters are ranked by their weighted densities and output as predicted complexes. Since CMC takes into account the weights of interactions, it prioritises more reliable cliques for the merging process while eliminating the less reliable ones, thereby discounting the effects of noise in PPI datasets.

2.1.4. Clustering with Overlapping Neighbourhood Expansion (ClusterONE)

ClusterONE [49] works similar to MCODE, by seeding and greedy neighbourhood expansion. ClusterONE first identifies seed proteins and greedily expands them into groups V based on a cohesiveness measure, given by:

$$f(V) = \frac{w^{(in)}(V)}{w^{(in)}(V) + w^{(bound)}(V) + p(V)},$$

where $w^{(in)}(V)$ is the total weight of interactions within V, $w^{(bound)}(V)$ is the total weight of interactions connecting V to the rest of the network, and $p(V)$ is a penalty term to model uncertainty in the data due to missing interactions. At each step, new proteins are included into V until $f(V)$ does not increase. V is then denoted as a locally cohesive group. Highly overlapping groups are merged to produce candidate complexes. Since this step allows for overlapping complexes, ClusterONE enhances the performance of MCODE and MCL.

2.1.5. Hierarchical Agglomerative Clustering with Overlaps (HACO)

HACO [50] modifies the classical hierarchical agglomerative clustering (HAC) [51] to identify overlapping complexes. The standard HAC algorithm with average linkage [52] maintains a pool of candidate sets to be merged. The distance between two non-overlapping sets S_1 and S_2 is given by:

$$d(S_1, S_2) = \frac{1}{|S_1||S_2|} \sum_{p\in S_1, q\in S_2} d(p, q).$$

where $d(p,q)$ is the negative of the affinity between proteins p and q. In each step of HAC, two non-overlapping sets S_1 and S_2 with the closest distance are iteratively merged to generate a new set S_{12}, while S_1 and S_2 are removed. The algorithm terminates when there are no remaining sets to merge.

In HACO, the sets S_1 and S_2 are retained for later use as required, the intuition being that if there is another set S_3 whose distance to S_1 is only slightly greater than that of S_2 then the decision to merge S_1 and S_2 could be arbitrary and unstable. In this case, HACO produces two merged sets S_{12} and S_{13} by retaining S_1 based on a divergence decision: if S_1 is considerably different from S_{12} then S_1 is retained (in order to generate S_{13}), otherwise S_1 is removed while keeping S_{12}. This procedure results therefore in overlapping complexes.

2.1.6. Ensemble clustering

Yong et al. [53] developed that an ensemble clustering approach to aggregate clusters generated from multiple clustering algorithms (including MCL, CMC, ClusterONE and HACO) using a majority voting-based scoring. The intuition behind aggregating clusters from different methods is to improve the coverage of complexes while maintaining the quality of the resultant clusters by scoring higher those predicted by multiple methods.

2.2. Methods based on network clustering combined with biological insights

Incorporating auxiliary information with the analysis of PPI networks overcomes some of the inherent limitations of PPI datasets, in particular noise, thus enhancing the performance of complex prediction methods.

2.2.1. Methods incorporating core-attachment structure

CORE [54], COACH [55], MCL-CAw [56,57] and CACHET [58] look for clusters that adhere to the *core-attachment* organisation, noted originally in yeast complexes by Gavin et al. [6]. Large-scale pull-down of yeast complexes using TAP-MS in [6] revealed that proteins within complexes are organised as two distinct sets: *cores* that constitute central functional units of complexes, and *attachments* that aid core proteins in their functions. Consequently, by specifically looking for clusters that adhere to this organisation, complexes could be identified with better accuracies.

In CORE [54], the probability for two proteins u and v with degrees d_u and d_v, respectively, to belong to the same core is determined by the number of common neighbours $|N(d_u) \cap N(d_v)|$ between u and v. The probability that u and v have at least m common neighbours participating in i interactions is computed under the null hypothesis that d_u interactions connecting u and d_v interactions connecting v are assigned to random neighbours in the PPI network. This probability is used to arrive at a P-value for u and v to belong to the same core, given by:

$$\begin{aligned} P\_\mathrm{value}(u,v) &= \Pr(\geq i \text{ interactions and} \geq\ m \text{ neighbours}) \\ &= \sum_{i \leq j \leq |E|, m \leq k \leq \min\{d_1,d_2\}-j} P_{\mathrm{interact}}(j||V|,d_1,d_2) \cdot P_{\mathrm{common}}(k||V|,d_1,d_2,j), \end{aligned}$$

where P_{interact} and P_{common} are computed under the null hypothesis. The P-value for (u,v) is then compared to P-values from all pairs involving u and v, and if (u,v) is ranked the highest among all these pairs (*i.e.*, (u,v) has the lowest P-value), then (u,v) is considered to belong to a two-core $\{u,v\}$.

CORE then repeatedly merges cores of sizes two, three and so on until further increase in core size is not possible, to produce the final set of cores. Subsequently, a protein p is added as an attachment to a core if p interacts with at least half the members of the core, to produce a complex.

COACH [55] works by identifying small dense neighbourhoods around proteins with high degrees in the PPI network. These dense subnetworks are then merged to generate cores. Attachments are added to these cores in a similar way as CORE to produce complexes.

MCL-CAw [56,57], on the other hand, refines clusters produced from MCL [41] by identifying core and attachment sets of proteins within each cluster to build complexes. A set of densely connected proteins within each MCL cluster is designated as a core, and attachment proteins are then included based on their connectivity to this core to produce a complex. MCL-CAw ensures that these attachment proteins can originate from outside the cluster and can be assigned to multiple cores, thus allowing for overlapping complexes.

CACHET [58] is different from the above methods in that it is specialised for reliability-weighted bipartite graphs of bait-prey interactions produced from TAP experiments. TAP uses immobilised baits proteins to capture prey proteins that interact, thus preserving co-complex relationships among these proteins; such relationships are typically lost when the TAP data are converted to pairwise interactions in PPI networks. CACHET first extracts maximal non-overlapping bicliques from the input bipartite graph as cores, and then assembles, in a similar way as CORE, the attachment proteins of these cores.

2.2.2. Methods incorporating functional information

Proteins within a complex are generally enriched for the same or similar functions. Therefore, combining functional annotations for proteins where available with the topology of PPI networks could improve complex identification. Following on this idea, the Restricted Neighbourhood Search Clustering (RNSC) [59], Dense neighbourhood Extraction using Connectivity and conFidence Features (DECAFF) [60] and Protein Complex Prediction (PCP) [61] make use of functional annotations from Gene Ontology [62] to predict complexes.

RNSC [59] employs a cost minimisation strategy to partition the PPI network by iteratively moving proteins between clusters until an integer-valued cost function is optimised. To prevent settling into poor local minima, RNSC periodically shuffles the clustering by dispersing the contents of a cluster at random. Finally, RNSC assigns a P-value to each of the clusters based on the functional coherence of the constituent proteins, and outputs only the clusters with $P < 0.001$ as the list of complexes. DECAFF [60] follows a clique-identification and merging procedure to identify clusters from the PPI network, and then filters these clusters using functional coherence of the proteins. On the other hand, PCP [61] uses the functional annotations to assign weights to interactions in the network, and uses these weighted interactions to cluster the network based on clique merging to generate complexes.

2.3. Comparative assessment of complex detection methods

Here we compare some of the complex prediction methods described above for predicting complexes from the yeast interactome. We obtain PPI data by combining physical interactions from the BioGRID [27], IntAct [63,64] and MINT [65] repositories. These repositories catalogue interactions detected from a multitude of studies, *e.g.* [1,2] (which employ Y2H), [5] (PCA) and [6,7] (TAP-MS). To assess the reliabilities of these interactions detected using different experimental techniques, we compute the reliability for each pair against a common independent criteria; here using similarities between Gene Ontology [62] annotations for these proteins. Specifically, each interaction (a,b) is weighted using a metric based on the number and the type of experiments that detected the interaction, given by:

$$\text{Reliability weight }(a,b) = 1 - \prod_{i \in E_{a,b}} (1 - rel_i)^{n_{i,a,b}},$$

where $E_{a,b}$ is the set of experimental technique that detected interaction (a,b); rel_i is the estimated reliability of experimental technique i calculated as the fraction of interactions detected by i such that both partners share at least one high-level Cellular Component term from Gene Ontology [62]; and $n_{i,a,b}$ is the number of times that experimental technique i detected interaction (a,b). A weighted PPI network was constructed using the top 20000 interactions, covering 3680 proteins (average node degree 10.87).

A predicted complex (or a cluster) P matches a known complex C if the Jaccard similarity between P and C, Jaccard $(P,C) \geqslant 0.5$, where:

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

Given the set of reference complexes $\mathbf{C} = \{C_1, C_2, \ldots C_n\}$, the precision, recall, and F-score of a set of predicted clusters $\mathbf{P} = \{P_1, P_2, \ldots P_m\}$ are given by:

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

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

$$F = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Predicted complexes are scored by their weighted densities and ranked. We calculate the area under the curve (AUC) of the precision–recall curve.

We employ the CYC2008 catalogue [32] (accessed 2012) as our set of reference yeast complexes, consisting of 408 complexes. We evaluate the methods only for prediction of large complexes (consisting of at least four proteins), of which there are 149 in CYC2008. This is because practically all methods find it difficult to detect small complexes (consisting of fewer than four proteins) and hence explicitly exclude these complexes from their predictions (*e.g.* see [55]). Besides, the possibility of a predicted complex matching a reference complex that is small purely by chance is relatively high [61], and therefore evaluating the methods becomes challenging (further discussed in Section 2.4.3).

Fig. 1 shows the performance of nine methods using precision, recall, F-measure and AUC. We see that methods incorporating biological information achieve higher recall and also generate ranked predictions with higher AUC compared to those based solely on network clustering. Methods that leverage reliability weights (MCL, CMC, ClusterONE, HACO and MCL-CAw) achieve higher recall than those that ignore these weights (MCODE). These results agree with evaluations from earlier studies [13,14]. Finally, ensemble clustering attains the highest recall while maintaining high AUC. The best-performing methods on an average predict about 75% of the complexes.

Fig. 2 shows the neighbourhood subnetworks around two example complexes predicted by some of the complex discovery methods. Fig. 2a shows the CBF3 complex, which consists of four proteins; these proteins are connected to a number of external proteins outside the complex making it difficult for some methods to recover this complex with high accuracy. CMC and COACH both recover the complex accurately, whereas RNSC recovers only three proteins and ClusterONE includes one extra (noisy) protein into the prediction. Fig. 2b shows the mRNA cleavage factor complex consisting of five proteins. Again, these proteins are connected to many external proteins; furthermore, one of the complex proteins, Hrp1p, is not directly connected to the rest of the complex. As a result, none of the methods predict the entire complex accurately: CMC and COACH both predict four of the five proteins, RNSC predicts three and MCL predicts two along with an external protein.

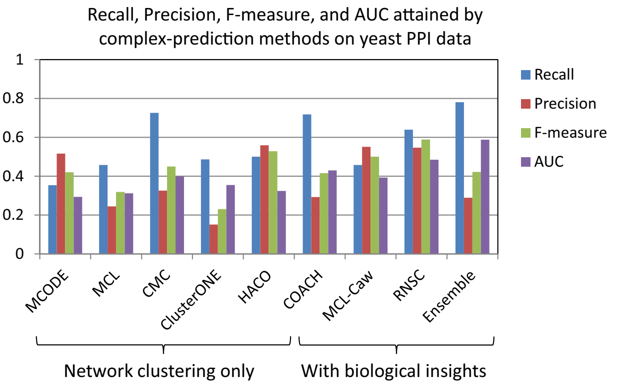

Fig. 1. Comparative assessment of complex-prediction methods on yeast PPI datasets. Ensemble refers to ensemble clustering by combining MCL, CMC ClusterONE, IPCA, COACH and RNSC [53].

2.4. Open challenges in complex detection

The above examples highlight some of the major challenges in complex discovery: many complexes either do not form dense subnetworks or are too small to be recovered accurately.

2.4.1. Detection of sparse complexes

Existing methods rely on the assumption that complexes are embedded as dense subnetworks within the PPI network and hence adopt density-based clustering for identifying complexes. In an analysis of complexes identifiable from a yeast PPI network, it was noted that only about 65% of complexes with at least four proteins in the network could be identified with Jaccard similarity $\geqslant 0.50$ [66]. The remaining 35% missed complexes did not meet the denseness criteria due to lack of sufficient interactions between member proteins. Even in the well-studied organism yeast, about 30% of the interactome still remains to be mapped of an estimated 25000–35000 interactions [67], thus posing a severe challenge to methods that are based on dense subnetworks to identify complexes. To overcome this limitation, [66] proposed to add functional interactions including association between proteins based on functional similarity to enhance the density of complexed regions within PPI networks, and thereby aid existing methods in identifying sparse complexes. Doing so enhanced the performance of MCL, MCL-CAw, CMC and HACO by up to 47% on average across a number of yeast PPI networks.

Supervised Weighting for Composite Networks (SWC) [53] integrates even more data sources including functional association data derived from multiple evidence such as co-occurrence in the literature, to build a composite protein network which fills in the missing interactions within sparse complexes. To reduce the noise introduced into the network, SWC weights the edges using a supervised-learning approach. This improved the performance of most clustering algorithms in yeast and human complex prediction, with sparse complexes benefitting the most.

2.4.2. Discerning overlapping complexes

Many proteins participate in multiple distinct complexes, resulting in complexes that overlap in the PPI network. These overlapping complexes are frequently highly inter-connected to each other, making it difficult for clustering algorithms to correctly decipher their boundaries [26]; approximately 40% of yeast complexes overlap with at least one other complex.

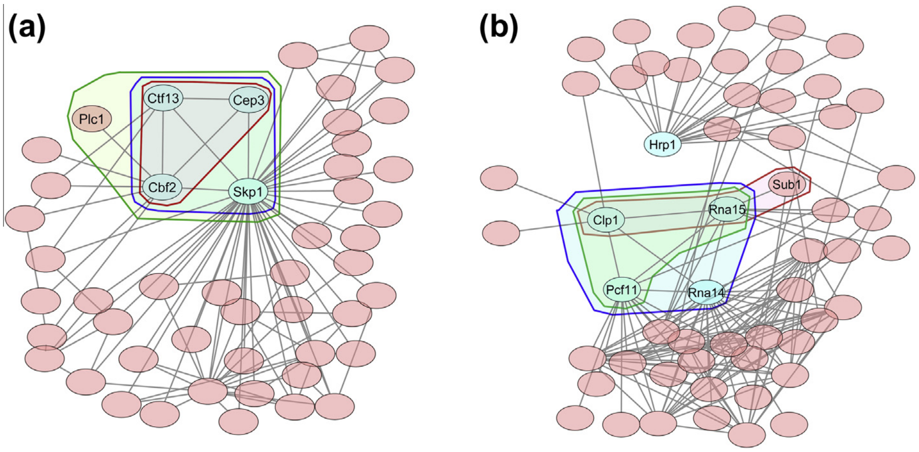

Fig. 2. Examples of complexes predicted by different methods. Blue nodes are proteins within the complex and red nodes are proteins not in the complex. (a) CBF3 complex, consisting of four proteins, compared with that predicted by CMC and COACH (blue), RNSC (red), ClusterONE (green). (b) mRNA cleavage factor complex, consisting of five proteins, compared with that predicted by CMC and COACH (blue), RNSC (green), MCL (red).

Some proteins use the same binding surface to interact with multiple partners so that these interactions do not occur simultaneously. Such mutually exclusive interactions can be used to discount simultaneously occurring interactions, which can help to deconvolute overlapping complexes and produce finer clusters in general. For example, Jung et al. [68] used structural data of protein binding interfaces to construct a simultaneous PPI network (SPIN) containing only cooperative interactions and exclude mutually exclusive interactions. MCODE and LCMA displayed considerable improvement on SPIN relative to the original PPI network. Ozawa et al. [69] used domain–domain interactions (DDIs) to identify conflicting pairs of protein interactions and used these to refine the clusters from MCODE and MCL. The accuracies of these methods improved by at least twofold. Similarly, Will and Helms [70] integrated PPI networks and DDIs, taking into account the connectivity constraints due to sharing of domains, to identify transcription-factor (TF) complexes in yeast.

Liu et al. [71] reasoned that proteins with many neighbours in the PPI network are unlikely to interact with all of them simultaneously. Such proteins, or hubs, were thus removed before clustering, and added back to the generated clusters to which these were highly connected. Furthermore, since a set of interactions can occur simultaneously only if all interacting partners are in the same cellular compartment, the PPI network was decomposed into spatially coherent subnetworks before clustering. This technique improved the performance of MCL, RNSC, IPCA, and CMC, in part because overlapping complexes could be more easily separated and extracted.

Tatsuke and Maruyama [72] observed that the sizes of protein complexes tend to follow a characteristic power-law distribution wherein the majority of complexes are small whereas the larger complexes occur less frequently. This insight was used to randomly partition the PPI network into complexes (clusters) of different sizes using Markov chain Monte-Carlo sampling [73]. Interestingly, this sampling-based approach (PPSampler) could recover several known complexes from CYC2008. The most recent version Repeated Simulated Annealing of Partitions of Proteins (ReSAPP) [74] uses simulated annealing method to optimise the sampling by returning the partition with the highest probability. ReSAPP combines clusters from multiple sampling runs and thereby can also identify overlapping complexes.

2.4.3. Detection of small complexes

Small complexes (consisting of fewer than four proteins) comprise the majority of complexes in yeast and human, but their prediction is especially susceptible to inaccuracies in the PPI network: missing interactions could easily disconnect a small complex whereas spurious interactions could embed the complex within a larger subnetwork. Topological measures such as interaction density applicable to large complexes are less effective for detecting small complexes – *e.g.* from a network with n proteins there are $O(n^3)$ triplets (with density 1) that could be predicted as three-protein complexes. Furthermore, evaluation measures such as Jaccard match become less effective for evaluating small complexes – *e.g.* a mismatch of only one protein in a three-protein complex renders the prediction inaccurate or less useful despite achieving a Jaccard of 0.50. As a result, most methods fare poorly in detecting small complexes (evaluated in [26,75]) or explicitly exclude small complexes from their predictions (*e.g.* see [55]). Detection of small complexes therefore requires specialised methods.

Yong et al. [75] propose one such specialised method called size-specific supervised weighting (SSS). SSS integrates functional associations and literature co-occurrences with PPI data, along with various topological characteristics, using a supervised approach to weight each interaction with its probability of belonging to a small complex. Small complexes are extracted and scored with their cohesiveness-weighted density, which incorporates interactions both within and surrounding each complex. SSS attains better performance in small-complex prediction compared to traditional clustering approaches, deriving about 50% more small complexes at equivalent precision levels.

Ruan et al. proposed two methods for predicting size-two and size-three complexes separately [76,77]. Both methods use weights of the interactions around putative small complexes as well as the number of domains in the constituent proteins to derive features for a kernel-based supervised approach. These methods outperform traditional clustering approaches in predicting heterodimeric and heterotrimeric complexes.

Protein sub-complexes can be considered as an interesting special case of overlapping and/or small complexes in which a subset of proteins from a larger complex forms a smaller but by itself a distinct complex. This can be related to cores in which the set of core proteins interact with different sets of attachments to form distinct complexes [6]. Since these sub-complexes overlap with multiple complexes, most general clustering methods either merge all complexes to produce less discernable large clusters. TAP data (*e.g.* [6,7]) is valuable here because bait-prey pairs from sub-complexes tend to appear multiple times as part of (larger) complexes. Zaki and Nora [78] found that CACHET [58], which is specialised to TAP data, was highly effective in identifying these sub-complexes.

Based on this idea, these authors developed TRIBAL (TRIad-Based Algorithm) which preserves the multi-edge nature of these bait-prey interactions in TAP data to identify sub-complexes.

2.5. Detecting evolutionarily conserved complexes

With rapid increase in the number of resources for human PPIs over the last several years [27–30], applying complex prediction methods to identify human complexes has become feasible, and recently a number of studies have attempted to reconstruct complexes from different human tissues and across diseases states (Section 4). Among the interesting observations is that many human complexes are ancient and slowly evolving, with roughly a quarter of the human complexes overlapping with those from lower-order organisms [15]. This has inspired several studies to look at the evolutionary conservation of complexes between human and lower-order organisms. While some of these studies have mainly looked into the evolutionary convergence or divergence of complexes, others have employed these insights to further enhance complex prediction.

Among the seminal works in this direction were by Kelley et al. [79] and Sharan et al. [80] who constructed *orthology networks* using conserved interactions between species (initially between *S. cerevisiae* and the bacteria *Helicobacter pylori* and later extended to human) based on protein-sequence homology, and clustered these networks to identify conserved complexes between these species. The complexes so-identified were involved in protein translation, DNA-damage response (DDR) and nuclear transport, suggesting that complexes from core cellular processes tend to be evolutionarily conserved.

van Dam and Snel [81] studied rewiring of protein complexes between yeast and human by mapping PPI networks onto *bona fide* complexes, and concluded that a majority of co-complexed protein pairs retained their interactions from yeast to human, thereby indicating that evolutionary changes in complexes were not due to extensive rewiring of complexed PPIs but instead due to gain or loss of protein subunits from yeast to human. Hirsh and Sharan [82] devised a probabilistic model of protein evolution and employed it to identify conserved complexes between species. Similar to observations by [79,80], these authors found that complexes involved in core cellular processes including pre-mRNA processing, replication, cytoskeleton maintenance and proteasome were highly conserved.

In an interesting work integrating 3D-protein structural information with PPI networks, Marsh et al. [83] characterised the evolutionary conservation of 'pathways of assembly' for complexes. The authors observed that evolutionary events optimised complex assembly by simplifying the topologies of complexes, and thereby demonstrated an evolutionary conservation of the assembly order. In particular, gene fusion events reduced the number of assembly steps by at least one, thereby generating fewer intermolecular interfaces in the resultant complex. These events also optimised network topologies by reducing the number of discrete protein interactions, leading to conservation of complexed regions within networks [84].

Nguyen et al. [36] integrated protein domain information with PPI networks to construct *domain-interolog networks* and studied conservation of complexes between yeast and human. These authors noted that although several proteins are conserved by sequence similarity between yeast and human (*e.g.* RAD9 and hRAD9), there are many others that did not show any sequence conservation (*e.g.* BRCA1 in human) and yet performed core functions (*e.g.* cell cycle and DDR) that were conserved. These proteins in fact retained conserved functional domains – for example, the BRCT domain present in yeast RAD9 and human hRAD9 is also present in the non-conserved human BRCA1 and 53BP1; all these proteins play vital roles in DDR [85]. Therefore, considering *functional conservation* by integrating domain similarity rather than mere sequence similarity is important to understand conservation patterns of complexes. Based on domain conservation, the authors found that several human complexes had in fact reorganised *via* creation of "mosaic" proteins that accumulated conserved domains from multiple yeast proteins.

Methods that detect coevolution of interacting proteins could also be used to detect complexes – *e.g.* using insights from studies such as [86] on the coevolution of entire protein sequences and specific interaction sites in the context of protein interactions (also see reviews [87,88]). Sets of interacting proteins that coevolve either tend to conserve their interacting domains or adapt to compensatory changes in binding surfaces of partners, thus suggesting evolutionary pressure possibly to conserve specific functions. Therefore, some of these groups of coevolving proteins could potentially constitute conserved complexes.

3. Integrating contextual information with PPI networks for predicting dynamic protein complexes

Many, if not all, protein complexes are *dynamic* entities, which assemble at a specific sub-cellular space and time to perform a specific function and disassemble after that. For example, cyclin-CDK complexes involving cyclin-dependent kinases (CDKs) are activated based on the concentration levels of cyclins in a cell-cycle dependent manner [89]. However, due to the lack of specific contextual (temporal and spatial) information in currently available PPI datasets, it is challenging to decipher the dynamics of complexes solely from PPI networks [71]. This limitation severely impacts the performance of computational methods and more critically our understanding of complex organisation and function [90].

3.1. Identifying temporal complexes

Several methods have looked into novel ways of integrating contextual information with PPI networks to understand the dynamics of complexes. One of the earliest attempts was by Han et al. [91] who integrated expression levels of genes with yeast PPI network to study *hub* proteins. Han et al. noted two distinct kinds of hubs that are transiently expressed and interact with other proteins to form dynamic modules – *date hubs*, which interact only with singleton or a small set of proteins at any given time, and *party hubs* which simultaneously interact with several proteins. Although initially contested [92] this finding is now widely accepted [93,94], with Komurov and White [95] further extending the concept to include *family hubs* that constitutively express and interact with other (constitutively expressed) proteins to form static modules.

By integrating PPI networks with the expression levels of cell-cycle proteins, de Lichtenberg et al. [96] studied the dynamics of complex assembly and disassembly during the yeast cell cycle. Eukaryotic complexes are composed of both constitutively expressed as well as dynamically expressed proteins, which enable them to assemble "just-in-time" to perform functions. Most subunits of complexes are pre-synthesized and pre-assembled whereas the remaining subunits are synthesized only when required, thereby tightly regulating the final complex assembly: by holding off on the last components, cells prevent accidental 'switching on' of complexes at wrong times.

Similarly, by integrating protein expression levels from the yeast cell cycle with cores and attachments within complexes, [97] found that attachments are enriched significantly higher for dynamically expressed proteins compared to cores, whereas the cores are enriched for constitutively expressed proteins. This pattern reflects the "reusability" of cores during complex formation: cores being reused across multiple complexes are maintained constitutively throughout the cell cycle, whereas attachments being required just-in-time are expressed dynamically when required during complex assembly.

Li et al. [98] identified temporal complexes by clustering PPI networks constructed using gene expression data from different experimental time points. Using yeast datasets, Li et al. found that about 60% of complexes existed only at one time point (*i.e.* more dynamic) whereas about 24% of complexes existed in more than three time points (*i.e.* more constitutive). By segregating the PPI network based on time-based profiles, dynamic sub-complexes could be separated from larger static clusters, thereby improving overall complex prediction. Similarly, [99] proposed a method Time Smooth Overlapping Complex Detection (TS-OCD) for joint analysis of PPI networks and time-series gene expression profiles to detect dynamic complexes at each time point. Analysis using yeast datasets showed that significantly many complexes could be detected compared to static methods, and in particular their method could identify complexes that share proteins dynamically to perform time-dependent functions.

Goh et al. [100] found that miRNAs with widely different expression profiles (*i.e.*, anti-coexpressed) strongly target hub-spokes in PPI networks but yet avoid targeting the same set of hub-spokes. This suggests that anti-coexpressed miRNAs play an important role in controlling the formation of protein complexes that are mutually exclusive. It is tantalizing to speculate on the possibility of inferring mutual exclusivity proteins which are targets of anti-coexpressed miRNAs, and exploiting this information *via* a SPIN-like approach [68] in protein complex prediction.

3.2. Integrating structural information with PPI networks

Incorporating information from three-dimensional (3D) structures of interacting proteins can further aid in the identification of protein complexes. Structural information on interacting proteins has been previously used to identify the nature of the interactions [101]. Proteins using the same interaction interface to bind different partner proteins primarily participate in multiple transient interactions as in the case of several kinases. On the other hand, some proteins use multiple interfaces to bind distinct partners and are often seen as members of obligate complexes [101]. With increasing availability of protein structures, it is now possible to annotate PPI networks with known 3D structures or reasonably accurate homology models [102]. Docking is often used to predict an ensemble of possible macromolecular assemblies of proteins usually through the prediction of complementary binding surfaces on partner proteins [103]. Docking can also be used to identify interacting protein pairs through complete cross-docking, where each protein within a set is docked with all other proteins to identify its potential interaction partners [104–106]. Using the information of known interaction interfaces, or predicted binding sites obtained through evolutionary sequence analysis, can improve the accuracy of interaction partner prediction through cross-docking. The prediction of binding affinities of interacting proteins is of great interest not only for assessing the interactions obtained from high-throughput experiments for their reliability, but also for predicting novel interactions between proteins. However, it is difficult to predict binding affinities using docking scores obtained from current scoring algorithms [107]. Docking is further complicated by the conformational changes that proteins undergo as a result of binding to their cognate partners. These conformational changes include backbone flexibility and movements of amino acid side-chains, both of which can be addressed by flexible protein–protein docking methods [108]. Several algorithms and automated tools have been developed for this purpose [109–111]. However, flexible docking is much more difficult than rigid docking and docking protein pairs with large conformational changes is still a challenge.

3.2.1. Flexibility and intrinsic disorder in protein complexes

The conformational changes that take place in proteins upon binding correspond to their flexibility in the unbound state [112]. Flexibility is important for the formation of large complexes [113] as well as those containing a greater diversity of subunits [114]. Flexibility allows binding over larger distances and in the form of larger binding interfaces without the loss of entropy [115]. Such flexible proteins participate in dynamic complexes and often contain large regions of *intrinsic disorder*. Intrinsic disorder is an extreme form of flexibility in protein structure.

Intrinsically disordered regions in proteins lack stable 3D structure under physiological conditions and can take on an ensemble of conformations (reviewed in [116] and more recently [117]). The high flexibility of disordered regions allows them to reversibly bind to several partner proteins [118]. Indeed, the presence of intrinsic disorder has long been associated with the ability of proteins to bind to multiple partner proteins [119,120] allowing them to play an important role in cell signalling and many other aspects of cellular function [117].

3.2.2. Fuzzy complexes

There is increasing recognition for the importance of intrinsic disorder in protein complexes [113], also known as fuzzy complexes [121]. Some complexes show static fuzziness where the disordered region in a protein folds into an ordered conformation on binding by undergoing coupled folding upon binding. An example of this is the induced folding of the disordered pKID (phosphorylated kinase-inducible activation domain) of the transcription factor CREB (cAMP-responsive element-binding protein) binding the KIX domain of CBP (CREB-binding protein) to induce transcription of downstream genes [122] (Fig. 3a). Similar folding also takes place in the N-terminal disordered region of p53 on binding the E3 ubiquitin ligase MDM2 [120]. Binding of the disordered region may also take place through the selection of a preformed conformer. For example, the KID (kinase inhibitory domain) of p27$^{\mathrm{Kip1}}$, a cyclin-dependent inhibitor, has some preformed helical structure that is used to bind cyclin A and subsequently CDK2 to control cell cycle [123] (Fig. 3b). While this was previously proposed as an instance of the induced-fit mechanism of binding [124], the role of the preformed helical structure of p27 in effective binding has been recently identified [125].

Proteins with disordered regions also form dynamic fuzzy complexes where the disordered region stays disordered either partially or completely on binding [126]. Thus, the disordered regions may remain flexible during binding without folding into a fixed structure as in the case of the inhibitor I-2 when it binds PP1 (protein phosphatase 1) [127] (Fig. 3c). Disordered linkers between two ordered domains within a protein also form parts of dynamic fuzzy complexes. They are advantageous because they allow the two domains to sample a large number of orientations with respect to each other, as observed in the calcium-binding domains of calmodulin which adopt different relative orientations when binding to different proteins [128–130].

It has been proposed that in some dynamic complexes, the disordered partner does not bind the ordered partner in a single location but rapidly changes between several conformers binding with the help of a mean electrostatic field rather than through discrete charges [128]. This binding is further affected by post-translational modifications which can change the mean charge presented by the disordered binding interface [128].

3.2.3. Binding interface and complex prediction

Intrinsically disordered regions in proteins frequently bind their interaction partners through the use of short linear motifs [129] which adopt different structures when binding different target proteins. On the other hand, proteins may also use molecular

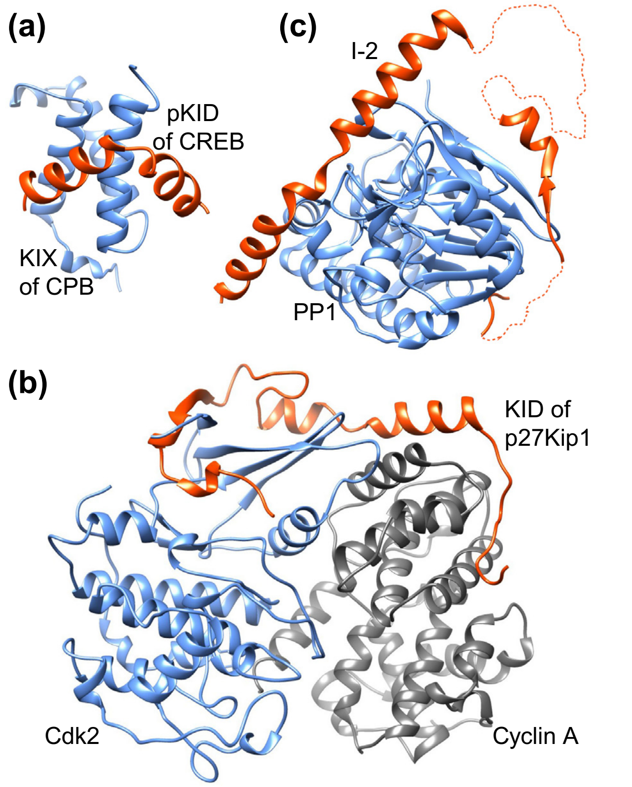

Fig. 3. Flexibility and intrinsic disorder in protein complexes. (a) Induced folding of the intrinsically disordered pKID domain of CREB (orange) on binding to the KIX domain of CBP (blue) (PDB id: 1KDX). (b) Binding of the intrinsically disordered KID domain (orange) of $p27^{Kip1}$ to the Cyclin A (grey) and Cdk2 (blue) (PDB id: 1JSU). (c) I-2 (orange) stays disordered when binding to PP1 (blue). Disordered regions are indicated by dotted lines and not visible in the X-ray crystal structure (PDB id: 2O8G).

recognition features (MoRFs) for binding their cognate partners [130]. MoRFs are short linear regions within disordered segments that participate in specific target recognition and undergo a disorder-to-order transition on binding. Given the flexibility of the binding partners in such complexes, it is difficult to predict their binding sites or protein assemblies. However, several tools have become available in recent years to predict the binding interfaces within the disordered regions. One such tool, SLiMPrints, predicts short linear motifs based on conserved regions that may participate in binding [131]. Other methods use machine learning techniques to identify binding interfaces using a host of features from the sequence of the disordered region including the amino acid propensity of known MoRFs and their flanking regions, physicochemical properties of the amino acids as well as evolutionary profiles [132–135]. It is now possible to use information from experimental techniques like Nuclear Magnetic Resonance (NMR) and Small Angle X-ray Scattering (SAXS) in combination with computational methods to model the ensemble of conformations that may be adopted by disordered regions within protein complexes [114,115]. The Protein Ensemble Database is a collection of such structural ensembles of proteins obtained from a combination of experimental and computational methods [136].

4. Identifying complexes in human diseases

Complexes are responsible for driving important mechanisms that maintain cellular homeostasis, but are also often the sites of dysregulation in diseases. The functional analysis of genes within complexes suggests that these complexes could be hotspots for perturbations due to genetic or environmental factors, thereby driving common and rare diseases [137,138]. Identifying complexes dysregulated in human diseases therefore forms an important extension of complex detection methods.

Vanunu et al. [139] employed a PPI network to associate complexes with diseases catalogued in the Online Mendelian Inheritance in Man (OMIM) database [140]. Using disease-to-gene associations from the OMIM database as a prior, the proposed method measures the association computed *via* network-propagation between causal genes and genetic diseases. The score computed for each gene is then used in combination with the PPI network to identify complexes involved in the disease. About 566 complexes were identified that were associated with hereditary or congenital diseases including ataxia-telangiectasia (AT), hereditary prostate cancer and microcephalic osteodysplastic primordial dwarfism (MOPD).

Similarly, Lage et al. [141] identified about 506 complexes that included disease-promoting genes implicated in disorders such as retinitis pigmentosa, epithelial ovarian cancer, inflammatory bowel disease, amyotrophic lateral sclerosis, Alzheimer disease, type-2 diabetes and coronary heart disease. To do this, the authors constructed a phenome-interactome network to identify candidate complexes, which were scored based on member proteins involved in these disorders, thus prioritising disease-associated complexes from the network.

In diseases such as cancer, cellular dysregulation often involves complexes that regulate critical functions including genome-stability maintenance, cell-cycle checkpointing and control, growth signalling and metabolism, the disruptions of which lead to increased accumulation of genomic instability, cell proliferation and metabolic dysfunction. For example, dysregulation of the BRCA1-A, -B and -C complexes due to loss-of-function mutations or epigenetic silencing of the *BRCA1* gene results in decreased DNA double-strand break repair efficiency, thus contributing to genomic instability in breast cancer [142,143].

Recently, a number of works have attempted to identify complexes dysregulated in cancer by studying rewiring of complexes between cancer conditions. For example, Srihari et al. [39,144] constructed condition-specific PPI networks by integrating gene-expression profiles with human PPI network across normal and cancer conditions in breast and pancreatic cancers. Subsequently, complexes were identified using CMC [45] from each of these condition-specific networks separately and matched to detect rewiring or changes to protein composition within complexes between these conditions. Interestingly, several known cancer genes were involved in these rewiring events, and the affected complexes displayed significant differences in expression levels between these cancer conditions. Among the dysfunctional complexes were those involved in DNA-damage repair (*e.g.* BRCA1 complexes) and growth-factor signalling (*e.g.* EGFR signalling) and also proteasomes, signalosomes and ribosomal complexes.

Zhao et al. [145] estimated the differential abundance of protein complexes between normal and cancer conditions in the cancers of 39 human tissues by using gene expression profiles. The authors employed known human complexes from the CORUM database [33], and estimated complex abundance by computing the optimal number of proteins required to form each complex and the number of copies of proteins present in the cell. Complexes involved in DNA-damage repair (*e.g.* BLM-TOP3A), histone modification (*e.g.* HDAC complexes), minichromosome maintenance (*e.g.* MCM complexes) and protein translation (*e.g.* RNA polymerases) showed abnormal expression in several human cancers.

Chen et al. [146] modelled disease-complex prioritisation in a network as an optimisation problem by maximising the information flow between a query disease and a candidate complex through the network. For a queried disease, this approach identified the associated proteins and the complexes formed by these proteins in the network. Application of this approach to breast cancer yielded six complexes involved in DNA-damage repair (BRCA1 and MSH complexes), replication factor (RFC complexes) and chromatin remodelling (SWI/SNF complex).

Goh et al. [147,148] demonstrated that analysing proteomic profiles in the context of protein complexes significantly improved the reproducibility and sensibility of biomarker identification from proteomic data. They introduced the concept of proteomic signature profile (PSP), which is a vector of protein complexes and their "hit rates" (*i.e.*, the proportion of protein components detected in a patient sample for the respective complexes) irrespective of individual protein's quantitation level. Complexes that were significantly differential in their hit rates between cases and controls were reported. They uncovered in liver cancer, an interesting relationship between the purine metabolism pathway and two other complexes involved in DNA-damage repair, suggesting progression to poor-stage liver cancer requires additional mutations that disrupt DNA-damage repair enzymes.

5. Conclusion

With increasing availability of PPI and other functional datasets, prediction of complexes has come a long way over the last several years. Apart from widely studied model organisms such as yeast [1–7], fruit fly [8,9] and worm [10], it is now possible to predict complexes from more sophisticated organisms including human [15]. This has provided new opportunities to study complexes under different contexts and across species, thus tracing the functional and evolutionary conservation of complexes. Integrating diverse information including 3D structure and time or context-based gene-expression profiles has helped to map the dynamics of complex formation and also understand their roles in diseases. This forms an excellent example where a fundamental problem such as complex prediction has had far-reaching applications in understanding the organisation and functions of the cell. We hope that this review aptly commemorates these efforts and inspires further advancement of research in this exciting area.

Acknowledgement

We thank Dr. Alison Anderson (The University of Queensland) for critical reading of the manuscript. S.S. is supported by an Australian National Health and Medical Research Council (NHMRC) Grant# 1028742.

References

[1] Uetz, P., Giot, L., Cagney, G., Traci, A., Judson, R., Knight, S.R., Lokshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbeish, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403 (6770), 623–627.

[2] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. 98 (8), 4569–4574.

[3] Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415 (6868), 180–183.

[4] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., Remor, C., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, M., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R.-R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, G., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415 (6868), 141–147.

[5] Michinck, S.W. (2003) Protein fragment complementation strategies for biochemical network mapping. Curr. Opin. Biotechnol. 14 (6), 610–617.

[6] Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440 (7084), 631–636.

[7] Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. and Greenblatt, J.F. (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440 (7084), 637–643.

[8] Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley Jr., R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J. and Rothberg, J.M. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302 (5651), 1727–1736.

[9] Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., McKillip, E., Shah, S., Stapleton, M., Wan, K.H., Yu, C., Parsa, B., Carlson, J.W., Chen, X., Kapadia, B., VijayRaghavan, K., Gygi, S.P., Celniker, S.E., Obar, R.A. and Artavanis-Tsakonas, S. (2011) A protein complex network of *Drosophila melanogaster*. Cell 147 (3), 690–703.

[10] Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Den, Van., Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E. and Vidal, M. (2004) A map of the interactome network of the metazoan *C. elegans*. Science 303 (5657), 540–543.

[11] Spirin, V. and Mirny, L. (2003) Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. 100 (21), 12123–12128.

[12] Bader, G.D. and Hogue, C.W.V. (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2.

[13] Li, X.L., Wu, M., Kwoh, C.C. and Ng, S.K. (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genomics 11 (Suppl. 1), S3.

[14] Srihari, S. and Leong, H.W. (2013) A survey of computational methods for protein complex prediction from protein interaction networks. J. Comput. Biol. Bioinf. 11 (2), 1230002.

[15] Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S., Babu, M., Craig, S.A., Hu, P., Wan, C., Vlasblom, J., Dar, V.U., Bezginov, A., Clark, G.W., Wu, G.C., Wodak, S.J., Tillier, E.R., Paccanaro, A., Marcotte, E.M. and Emili, A. (2012) A census of human soluble protein complexes. Cell 150 (5), 1068–1081.

[16] Hartwell, L.H., Hopfield, J.J., Leiber, S. and Murray, A.W. (1999) From molecular to modular cell biology. Nature 402 (Suppl. 6761), C47–C52.

[17] Zhang, B., Park, B.H., Karpinets, T. and Samatova, N. (2008) From pull-down data to protein interaction networks and complexes with biological relevance. Bioinformatics 24 (7), 979–986.

[18] Bader, G.D. and Hogue, C.W.V. (2002) Analysing yeast protein–protein interaction data obtained from different sources. Nat. Biotechnol. 20 (10), 991–997.

[19] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale datasets of protein–protein interactions. Nature 417 (6887), 399–403.

[20] Chua, H.N. and Wong, L. (2008) Increasing the reliability of protein interactomes. Drug Discov. Today 13 (15–16), 652–658.

[21] Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.-F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D.E., Ecker, J.R., Roth, F.P. and Vidal, M. (2009) Literature-curated protein interaction datasets. Nat. Methods 6 (1), 39–46.

[22] Patil, A., Nakai, K. and Nakamura, H. (2011) HitPredict: a database of quality assessed protein–protein interactions in nine species. Nucleic Acids Res. 39 (Database issue), D744–D749.

[23] Brohee, S. and van Helden, J. (2006) Evaluation of clustering algorithms for protein–protein interaction networks. BMC Bioinformatics 7, 488.

[24] Vlasblom, J. and Wodak, S. (2009) Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics 10, 99.

[25] Collins, S.R., Kemmeren, P., Zhao, X.C., Greenbalt, J.F., Spencer, F., Holstege, F., Weissman, J.S. and Krogan, N.J. (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. Mol. Cell. Proteomics 6 (3), 439–450.

[26] Yong, C.H. and Wong, L. (2015) From the static interactome to dynamic protein complexes: three challenges. J. Bioinfom. Comput. Biol. 13 (2), 1571001.

[27] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34 (Database), D535–D539.

[28] Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. and Pandey, A. (2009) Human Protein Reference Database–2009 update. Nucleic Acids Res. 37 (Database), D767–D772.

[29] Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., Reguly, T., Breitkreutz, A., Sellam, A., Chen, D., Chang, C., Rust, J., Livstone, M., Oughtred, R., Dolinski, K. and Tyers, M. (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res. 41 (Database), D816–D823.

[30] Rolland, T., Tasan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S.D., Yang, X., Ghamsari, L., Balcha, D., Begg, B.E., Braun, P., Brehme, M., Broly, M.P., Carvunis, A.R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B.J., Hardy, M.F., Jin, M., Kang, S., Kiros, R., Lin, G.N., Luck, K., MacWilliams, A., Menche, J., Murray, R.R., Palagi, A., Poulin, M.M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J.M., Scholz, A., Shah, A.A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A.O., Trigg, S.A., Twizere, J.C., Vega, K., Walsh, J., Cusick, M.E., Xia, Y., Barabási, A.L., Iakoucheva, L.M., Aloy, P., De Las, Rivas J., Tavernier, J., Calderwood, M.A., Hill, D.E., Hao, T., Roth, F.P. and Vidal, M. (2014) A proteome-scale map of the human interactome network. Cell 159 (5), 1212–1226.

[31] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., Münsterkötter, M., Pagel, P., Strack, N., Stümpflen, V., Warfsmann, J. and Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 32 (Database), D41–D44.

[32] Pu, S., Wong, J., Turner, B., Cho, E. and Wodak, S.J. (2009) Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res. 37 (3), 825–831.

[33] Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V. and Mewes, H.W. (2008) CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 36 (Database), D646–D650.

[34] Csermely, P., Korcsmáros, T., Kiss, H.J., London, G. and Nussinov, R. (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. Pharmacol. Ther. 138 (3), 333–408.

[35] Kuzmanov, U. and Emili, A. (2013) Protein–protein interaction networks: probing disease mechanisms using model systems. Genome Med. 5 (4), 37.

[36] Nguyen, P.V., Srihari, S. and Leong, H.W. (2013) Identifying conserved protein complexes between species by constructing interolog networks. BMC Bioinformatics 14 (Suppl. 16), S8.

[37] Boone, C., Bussey, H. and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. Nat. Rev. Genet. 8 (6), 437–449.

[38] Kaelin Jr., W.G. (2005) The concept of synthetic lethality in the context of anticancer therapy. Nat. Rev. Cancer 5 (9), 689–698.

[39] Srihari, S. and Ragan, M.A. (2013) Systematic tracking of dysregulated modules identifies novel genes in cancer. Bioinformatics 29 (12), 1553–1561.

[40] Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13 (11), 2498–2504.

[41] van Dongen S. (2000). Graph clustering by flow simulation (Ph.D. thesis), University of Utrecht.

[42] Enright, A.J., van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30 (7), 1575–1584.

[43] Pu, S., Vlasblom, J., Emili, A., Greenbalt, J. and Wodak, S.J. (2007) Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. Proteomics 7 (6), 944–960.

[44] Friedel, C., Krumsiek, J. and Zimmer, R. (2009) Bootstrapping the interactome: unsupervised identification of protein complexes in yeast. J. Comput. Biol. 16 (8), 971–987.

[45] Liu, G., Wong, L. and Chua, H.N. (2009) Complex discovery from weighted PPI networks. Bioinformatics 25 (15), 1891–1897.

[46] Adamcsek, B., Palla, G., Farkas, I., Derenyi, I. and Vicsek, T. (2006) CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22 (8), 1021–1023.

[47] Li, X.L., Tan, S.H., Foo, C.S. and Ng, S.K. (2005) Interaction graph mining for protein complexes using local clique merging. Genome Inform. 16 (2), 260–269.

[48] Tomita, E., Tanaka, A. and Takahashi, H. (2006) The worst-case time complexity for generating all maximal cliques and computational experiments. J. Theor. Comput. Sci. 363 (1), 28–42.

[49] Nepusz, T., Yu, H. and Paccanaro, A. (2012) Detecting overlapping protein complexes in protein–protein interaction networks. Nat. Meth. 9, 471–472.

[50] Wang, H., Kakaradov, B., Collins, S.R., Karotki, L., Fiedler, D., Shales, M., Shokat, K.M., Walter, T., Krogan, N.J. and Koller, D. (2009) A complex-based reconstruction of the *Saccharomyces cerevisiae* interactome. Mol. Cell. Proteomics 8, 1361–1377.

[51] Kaufman, L. and Rousseeuw, P.J. (2009) Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience, New York.

[52] Sokal, R.R. and Michener, C.D. (1958) A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38, 2.

[53] Yong, C.H., Liu, G., Chua, H.N. and Wong, L. (2012) Supervised maximum-likelihood weighting of composite protein networks for complex prediction. BMC Syst. Biol. 6 (Suppl. 2), S13.

[54] Leung, H., Xiang, Q., Yiu, S.M. and Chin, F.Y. (2009) Predicting protein complexes from PPI data: a core-attachment approach. J. Comput. Biol. 16 (2), 133–144.

[55] Wu, M., Li, X. and Ng, S.K. (2009) A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinformatics 10, 169.

[56] Srihari, S., Ning, K. and Leong, H.W. (2009) Refining Markov clustering for complex detection by incorporating core-attachment structure. Genome Inform. 23 (1), 159–168.

[57] Srihari, S., Ning, K. and Leong, H.W. (2010) MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. BMC Bioinformatics 11, 504.

[58] Wu, M., Li, X., Kwoh, C.K., Ng, S.K. and Wong, L. (2012) Discovery of protein complexes with core-attachment structures from TAP data. J. Comput. Biol. 19 (9), 1027–1042.

[59] King, A.D., Przulj, N. and Jurisca, I. (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20 (17), 3013–3020.

[60] Li, X.L., Foo, C.S. and Ng, S.K. (2007) Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. Proc. Comput. Syst. Bioinform. Conf. 6, 157–168.

[61] Chua, H.N., Ning, K., Sung, W.K., Leong, H.W. and Wong, L. (2008) Using indirect protein–protein interactions for protein complex prediction. J. Bioinform. Comput. Biol. 6 (3), 435–466.

[62] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: a tool for the unification of biology. Nat. Genet. 25 (1), 25–29.

[63] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004) IntAct: an open source molecular interaction database. Nucleic Acids Res. 32 (Database), D452–D455.

[64] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., Jandrasits, C., Jimenez, R.C., Khadake, J., Mahadevan, U., Masson, P., Pedruzzi, I., Pfeiffenberger, E., Porras, P., Raghunath, A., Roechert, B., Orchard, S. and Hermjakob, H. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res. 40 (Database), D841–D846.

[65] Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res. 35 (Database), D572–D574.

[66] Srihari, S. and Leong, H.W. (2012) Employing functional interactions for characterisation and detection of sparse complexes from yeast PPI networks. Int. J. Bioinf. Res. Appl. 8 (3), 286–304.

[67] Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M. and Wiuf, C. (2008) Estimating the size of the human interactome. Proc. Natl. Acad. Sci. USA 105 (19), 6959–6964.

[68] Jung, S.H., Hyun, B., Jang, W.H., Hur, H.Y. and Han, D. (2009) Protein complex prediction based on simultaneous protein interaction network. Bioinformatics 26 (3), 385–391.

[69] Ozawa, Y., Saito, R., Fujimori, S., Kashima, H., Ishizaka, M., Yanagawa, H., Miyamoto-Sato, E. and Tomita, M. (2010) Protein complex prediction via verifying and reconstructing the topology of domain–domain interactions. BMC Bioinformatics 11, 350.

[70] Will, T. and Helms, V. (2014) Identifying transcription factor complexes and their roles. Bioinformatics 30 (17), 415–421.

[71] Liu, G., Yong, C.H., Chua, H.N. and Wong, L. (2011) Decomposing PPI networks for complex discovery. Proteome Sci. 9 (S1), S15.

[72] Tatsuke, D. and Maruyama, O. (2012) Sampling strategy for protein complex prediction using cluster size frequency. Gene 518, 152–158.

[73] Brooks, S., Gelmen, A., Jones, G.L. and Meng, X.L. (2011) Handbook of Markov Chain Monte Carlo, Chapman & Hall/CRC.

[74] Kobiki, S. and Maruyama, O. (2014) ReSAPP: predicting overlapping protein complexes by merging multiple-sampled partitions of proteins. J. Bioinform. Comput. Biol. 12 (6), 1442004.

[75] Yong, C.H., Maruyama, O. and Wong, L. (2014) Discovery of small protein complexes from PPI networks with size-specific supervised weighting. BMC Syst. Biol. 8 (Suppl. 5), S3.

[76] Ruan, P., Hayashida, M., Maruyama, O. and Akutsu, T. (2013) Prediction of heterodimeric protein complexes from weighted protein–protein interaction networks using novel features and kernel functions. PLoS One 8 (6), e65265.

[77] Ruan, P., Hayashida, M., Maruyama, O. and Akutsu, T. (2014) Prediction of heterotrimeric protein complexes by two-phase learning using neighbouring kernels. BMC Bioinformatics 15 (Suppl. 2), S6.

[78] Zaki, N. and Mora, A. (2014) A comparative analysis of computational approaches and algorithms for protein subcomplex identification. Sci. Rep. 4, 4262.

[79] Kelley, B.P., Sharan, R., Karp, R.M., Sittler, T., Root, D.E., Stockwell, B.R. and Ideker, T. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. USA 100, 11394–11399.

[80] Sharan, R., Ideker, T., Kelley, B. and Shamir, R. (2005) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. J. Comput. Biol. 12 (6), 835–846.

[81] van Dam, J.P. and Snel, B. (2008) Protein complex evolution does not involve extensive network rewiring. PLoS Comp. Biol. 4 (7), e1000132.

[82] Hirsh, E. and Sharan, R. (2007) Identification of conserved protein complexes based on a model of protein network evolution. Bioinformatics 23 (2), e170–e176.

[83] Marsh, J.A., Hernandenz, H., Hall, Z., Ahnhert, S.E., Perica, T., Robinson, C.V. and Teichmann, S.A. (2013) Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell 153 (2), 461–470.

[84] Zhang, J., Zheng, F. and Grigoryan, G. (2014) Design and designability of protein-based assemblies. Curr. Opin. Struct. Biol. 27, 79–86.

[85] Bork, P., Hoffman, K., Bucher, P., Neuwald, A.F., Alstchul, S.F. and Koonin, E.V. (1997) A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. FASEB J. 11 (1), 68–76.

[86] Lovell, S.C. and Robertson, D.L. (2010) An integrated view of molecular coevolution in protein–protein interactions. Mol. Biol. Evol. 27 (11), 2567–2575.

[87] de Juan, D., Pazos, F. and Valencia, A. (2013) Emerging methods in protein co-evolution. Nat. Rev. Genet. 14, 249–261.

[88] Andreani, J. and Guerois, R. (2014) Evolution of protein interactions: from interactomes to interfaces. Arc. Biochem. Biophys. 554, 65–75.

[89] Nurse PM. (2001). Cyclin dependent kinases and cell cycle control. Nobel Lecture, December 9.

[90] Przytycka, T., Singh, M. and Slonim, D.K. (2010) Toward the dynamic interactome: it's about time. Brief Bioinf. 11 (1), 15–29.

[91] Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature 430 (6995), 88–93.

[92] Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D. and Tyers, M. (2006) Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol. 4 (10), e317.

[93] Agarwal, S., Deane, C.M., Porter, M.A. and Jones, N.S. (2010) Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. PLoS Comput. Biol. 6 (6), e1000817.

[94] Ning, K., Ng, H.K., Srihari, S., Leong, H.W. and Nesvizhskii, A.I. (2010) Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology. BMC Bioinformatics 11, 505.

[95] Komurov, K. and White, M. (2007) Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. Mol. Syst. Biol. 3, 110.

[96] de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. Science 307 (5710). 727-7.

[97] Srihari, S. and Leong, H.W. (2012) Temporal dynamics of protein complexes in PPI networks: a case study using yeast cell-cycle complexes. BMC Bioinformatics 13 (Suppl. 17), S16.

[98] Li, M., Chen, W., Wang, J., Wu, F.-X. and Pan, Y. (2014) Identifying dynamic protein complexes based on gene expression profiles and PPI networks. BioMed. Res. Inter. 22, 375262.

[99] Ou-Yang, L., Dai, D.Q., Li, X.L., Wu, M., Zhang, X.F. and Yang, P. (2014) Detecting temporal protein complexes from dynamic protein–protein interaction networks. BMC Bioinformatics 15, 335.

[100] Goh, W., Oikawa, H., Sng, J., Sergot, M. and Wong, L. (2012) The role of miRNAs in complex formation and control. Bioinformatics 28 (4), 453–456.

[101] Kim, P.M., Lu, L.J., Xia, Y. and Gerstein, M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314, 1938–1941.

[102] Mosca, R., Ceol, A. and Aloy, P. (2013) Interactome3D: adding structural details to protein networks. Nat. Methods 10, 47–53.

[103] Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B. and Sali, A. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 10, e1001244.

[104] Sacquin-Mora, S., Carbone, A. and Lavery, R. (2008) Identification of protein interaction partners and protein–protein interaction sites. J. Mol. Biol. 382, 1276–1289.

[105] Yoshikawa, T., Tsukamoto, K., Hourai, Y. and Fukui, K. (2009) Improving the accuracy of an affinity prediction method by using statistics on shape complementarity between proteins. J. Chem. Inf. Model. 49, 693–703.

[106] Lopes, A., Sacquin-Mora, S., Dimitrova, V., Laine, E., Ponty, Y. and Carbone, A. (2013) Protein–protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput. Biol. 9, e1003369.

[107] Kastritis, P.L. and Bonvin, A.M. (2010) Are scoring functions in protein–protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J. Proteome Res. 9, 2216–2225.

[108] Andrusier, N., Mashiach, E., Nussinov, R. and Wolfson, H.J. (2008) Principles of flexible protein–protein docking. Proteins 73, 271–289.

[109] de Vries, S. and Zacharias, M. (2013) Flexible docking and refinement with a coarse-grained protein model using ATTRACT. Proteins 81, 2167–2174.

[110] Torchala, M., Moal, I.H., Chaleil, R.A., Fernandez-Recio, J. and Bates, P.A. (2013) SwarmDock: a server for flexible protein–protein docking. Bioinformatics 29, 807–809.

[111] Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S.E., Xia, B., Hall, D.R. and Vajda, S. (2013) How good is automated protein docking? Proteins 81, 2159–2166.

[112] Marsh, J.A., Teichmann, S.A. and Forman-Kay, J.D. (2012) Probing the diverse landscape of protein flexibility and binding. Curr. Opin. Struct. Biol. 22, 643–650.

[113] Hegyi, H., Schad, E. and Tompa, P. (2007) Structural disorder promotes assembly of protein complexes. BMC Struct. Biol. 7, 65.

[114] Marsh, J.A. and Teichmann, S.A. (2014) Protein flexibility facilitates quaternary structure assembly and evolution. PLoS Biol. 12, e1001870.

[115] Mittag, T., Kay, L.E. and Forman-Kay, J.D. (2010) Protein dynamics and conformational disorder in molecular recognition. J. Mol. Recognit. 23, 105–116.

[116] Wright, P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293, 321–331.

[117] Wright, P.E. and Dyson, H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. 16, 18–29.

[118] Patil, A., Kinoshita, K. and Nakamura, H. (2010) Hub promiscuity in protein–protein interaction networks. Int. J. Mol. Sci. 11, 1930–1943.

[119] Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M. and Uversky, V.N. (2005) Flexible nets. FEBS J. 272, 5129–5148.

[120] Patil, A. and Nakamura, H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett. 580, 2041–2045.

[121] Tompa, P. and Fuxreiter, M. (2008) Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. Trends Biochem. Sci. 33, 2–8.

[122] Sugase, K., Dyson, H.J. and Wright, P.E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447, 1021–1025.

[123] Russo, A.A., Jeffrey, P.D., Patten, A.K., Massague, J. and Pavletich, N.P. (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature 382, 325–331.

[124] Lacy, E.R., Filippov, I., Lewis, W.S., Otieno, S., Xiao, L., Weiss, S., Hengst, L. and Kriwacki, R.W. (2004) P27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. Nat. Struct. Mol. Biol. 11, 358–364.

[125] Otieno, S. and Kriwacki, R. (2012) Probing the role of nascent helicity in p27 function as a cell cycle regulator. PLoS One 7, e47177.

[126] Fuxreiter, M. and Tompa, P. (2012) Fuzzy complexes: a more stochastic view of protein function. Adv. Exp. Med. Biol. 725, 1–14.

[127] Marsh, J.A., Dancheck, B., Ragusa, M.J., Allaire, M., Forman-Kay, J.D. and Peti, W. (2010) Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. Structure 18, 1094–1103.

[128] Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., Tyers, M. and Forman-Kay, J.D. (2010) Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. Structure 18, 494–506.

[129] Nguyen Ba, A.N., Yeh, B.J., van Dyk, D., Davidson, A.R., Andrews, B.J., Weiss, E.L. and Moses, A.M. (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci. Signal. 5, rs1.

[130] Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K. and Uversky, V.N. (2006) Analysis of molecular recognition features (MoRFs). J. Mol. Biol. 362, 1043–1059.

[131] Davey, N.E., Cowan, J.L., Shields, D.C., Gibson, T.J., Coldwell, M.J. and Edwards, R.J. (2012) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res. 40, 10628–10641.

[132] Meszaros, B., Simon, I. and Dosztanyi, Z. (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. 5, e1000376.

[133] Disfani, F.M., Hsu, W.L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N. and Kurgan, L. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28, i75–i83.

[134] Jones, D.T. and Cozzetto, D. (2014) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31 (6), 857–863.

[135] Malhis, N. and Gsponer, J. (2015) Computational Identification of MoRFs in Protein Sequences. Bioinformatics, http://dx.doi.org/10.1093/bioinformatics/btv060.

[136] Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., Sussman, J., Svergun, D.I., Uversky, V.N., Vendruscolo, M., Wishart, D., Wright, P.E. and Tompa, P.

(2014) PE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res. 42, D326–D335.

[137] Barabási, A.L., Gulbahce, N. and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. Nat. Rev. Genet. 12 (1), 56–68.

[138] Lage, K. (2014) Protein–protein interactions and genetic diseases: the interactome. Biochim. Biophys. Acta (BBA) 1842 (10), 1971–1980.

[139] Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. and Sharan, R. (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. 6 (1), e1000641.

[140] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. and McKusick, V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33 (Database), D514–D517.

[141] Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y. and Brunak, S. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. 25 (3), 309–316.

[142] Paolo, A.D., Racca, C., Calsou, P. and Larminat, F. (2014) Loss of BRCA1 impairs centromeric cohesion and triggers chromosomal instability. FASEB J. 28 (12), 5250–5261.

[143] Liu, C., Srihari, S., Lê Cao, K.-A., Chevenix-Trench, G., Simpson, P.T., Ragan, M.A. and Khanna, K.K. (2014) A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. Nucleic Acids Res. 42 (10), 6106–6127.

[144] Srihari, S., Madhamshettiwar, P.B., Song, S., Liu, C., Simpson, P.T., Khanna, K. and Ragan, M.A. (2014) Complex-based analysis of dysregulated cellular processes in cancer. BMC Syst. Biol. 8 (Suppl. 4), S1.

[145] Zhao, J., Lee, S.H., Huss, M. and Holme, P. (2013) The network organization of cancer-associated protein complexes in human tissues. Sci. Rep. 3, 1583.

[146] Chen, Y., Jacquemin, T., Zhang, S. and Jiang, R. (2014) Prioritizing protein complexes implicated in human diseases by network optimization. BMC Syst. Biol. 8 (Suppl. 1), S2.

[147] Goh, W., Lee, Y.H., Ramdzan, Z., Sergot, M., Chung, M. and Wong, L. (2012) Proteomics Signature Profiling (PSP): a novel contextualization approach for cancer proteomics. J. Proteome Res. 11 (3), 1571–1581.

[148] Goh, W., Fan, M., Low, H.S., Segot, M. and Wong, L. (2013) Enhancing the utility of Proteomics Signature Profiling (PSP) with Pathway Derived Subnets (PDSs), performance analysis and specialised ontologies. BMC Genomics 14, 35.

[149] Kalaev, M., Smoot, M., Ideker, T. and Sharan, R. (2008) NetworkBLAST: comparative analysis of protein networks. Bioinformatics 24 (4), 594–596.