Chapter 20

Computational Analysis of the Chaperone Interaction Networks

Ashwani Kumar, Kamran Rizzolo, Sandra Zilles, Mohan Babu, and Walid A. Houry

Abstract

We provide computational protocols to identify chaperone interacting proteins using a combination of both physical (protein–protein) and genetic (gene–gene or epistatic) interaction data derived from the published large-scale proteomic and genomic studies for the budding yeast *Saccharomyces cerevisiae*. Using these datasets, we discuss bioinformatic analyses that can be employed to build comprehensive high-fidelity chaperone interaction networks. Given that many proteins typically function as complexes in the cell, we highlight various step-wise approaches for combining both the genetic and physical interaction datasets to decipher intra- and inter-connections for distinct chaperone- and non-chaperone-containing complexes in the network. Together, these informatics procedures will aid in identifying protein complexes with distinctive functional specializations in the cell that yield a very broad and diverse set of interactions. The described procedures can also be leveraged to datasets from other eukaryotes, including humans.

Key words Chaperone network, Functional enrichment, Genetic interactions, Physical interactions, Protein complexes

1 Introduction

Molecular chaperones are key players of cellular protein folding and assembly [1, 2]. Chaperone proteins are found in all cellular compartments and are involved in numerous physiological processes. Typically, chaperones are grouped into families depending on sequence similarity and function. The major chaperone families in the budding yeast *Saccharomyces cerevisiae* are 2 Hsp90s, 14 Hsp70s, 22 Hsp40s, 8 CCTs, 1 Hsp60, 1 Hsp10, 6 prefoldins, 5 ATPases associated with diverse cellular activities (AAA+), 7 small heat-shock proteins (sHsps), and 1 calnexin (total of 67 chaperones). Additionally, the Hsp70 and Hsp90 chaperones function with 4 and 11

Ashwani Kumar and Kamran Rizzolo are Co-first authors.
Mohan Babu and Walid A. Houry are co-corresponding authors.

Stuart K. Calderwood and Thomas L. Prince (eds.), *Chaperones: Methods and Protocols*, Methods in Molecular Biology, vol. 1709, https://doi.org/10.1007/978-1-4939-7477-1_20,

partner proteins termed cochaperones, respectively [3]. Despite many mechanistic and functional studies on both chaperones and cochaperones (CCos), the spectrum of cellular substrates and cellular functions they mediate remains largely incomplete. Hence, to obtain a better view of the division of labor among molecular chaperones in the cell, it is necessary to study them at a global systems level.

The use of proteomic methods has become a key tool to study phenotypes in cells by mapping physical (protein–protein) and genetic (gene–gene or epistatic) interaction networks. Typically, experiments to map physical interactions involve three essential steps: (1) separation and isolation of proteins; (2) the acquisition of sequence information for protein identification; and (3) database utilization for downstream analysis [4]. While protein–protein interactions (PPIs) can be mapped using various proteomic approaches in many model organisms such as human and yeast [5–7], the most standard techniques used to perform large-scale, systematic measurements of PPIs involves precision-based mass-spectrometry (MS) methods [8]. For instance, PPIs can be obtained by affinity purifying the endogenously tagged bait protein and then identifying co-purifying interactors by tandem MS/MS.

On the other hand, large-scale genetic interaction (GI) data have also been used to unmask gene and protein organization in the cell. Most insights into genetic interaction networks have been gained from the work done in the budding yeast [9], Gram-negative bacteria [10, 11], Gram-positive bacteria [12, 13], and other species [14–16]. To study GIs in yeast, double mutant strains are systematically created by mating a resistance-marked "query" deletion mutant strain against an array of single-gene deletion mutants typically marked with kanamycin using synthetic genetic analysis (SGA) technology [9, 17]. Such methodology allows for a quantitative assessment of the relative fitness of a double-mutant meiotic progeny using the GI scores, which are further categorized into aggravating (negative or synthetic lethal) or alleviating (positive or buffering) GIs. Aggravating interactions occur when the double-mutant fitness is lower than the expected for the two single mutants and may reflect compensatory pathways. The most extreme type of aggravating GIs is referred to as "synthetic lethal" where the double-mutant (compared to single mutants) does not grow at all. In contrast, alleviating interactions occur when the double-mutant fitness is greater than that expected for the two single mutants. For instance, this scenario can occur when genes function in the same nonessential pathway or complex. Both types of GIs from the network can be organized in a two-dimensional hierarchical clustering, where clusters are formed from the query genes according to the overlap of their interactions with the array genes. Sets of genes either with similar GI scores (positive or negative) or those functioning within the same pathway

(or subunits within a complex) tend to cluster together. Furthermore, the GI profile similarity provides a potential biological function for an uncharacterized gene based on its GI profile similarity with known genes.

Using the PPI and GI frameworks, our group published in 2005 a comprehensive physical and genetic analysis of the Hsp90 chaperone interaction network, showing a broad role of the chaperones in many distinct cellular pathways [18, 19]. Subsequently, in 2009, we published a yeast chaperone physical interaction atlas for 63 chaperones [20], which allowed us to uncover a clear distinction between chaperones that are promiscuous and chaperones that are functionally specific. The analysis indicated the presence of cellular hot spots of chaperone interactions in the cell. Recent efforts by our groups have also concentrated on building a comprehensive chaperone and cochaperone (CCo) interaction network using a combination of PPI and GI data. The integration applicability of various data types in network biology can provide a multidimensional approach to the study of proteomics [21]. This is very useful for CCos given that they are typically promiscuous in their interactions. The use of both physical and genetic data types provides information on inter- and intra- CCo complex interactions that would otherwise be missed by using one single approach.

In this chapter, we provide detailed computational protocols and source codes to build a comprehensive interaction network based on PPI and GI data. Most of our work has concentrated on chaperones and their cochaperones, but the described algorithms can be applied to networks with proteins involved in any other functions.

2 Methods

The protocols provided demonstrate various computational methods to determine functional relationships among genes and proteins. We also describe approaches to integrating similar (e.g., protein interactions from different studies) as well as different (e.g., proteomic and genomic) biological data. Various sources of protein interactome and computational tools are provided below along with relevant analyses. The algorithms have been provided in R code (https://www.r-project.org), which is a language for statistical computing and graphics.

2.1 Construction of the Chaperone Genetic Interaction (GI) Network

SGA-based large-scale screening is the most widely used approach to identifying genetic interactions (i.e., epistatic relationships) between genes. Briefly, a GI between two genes is estimated by comparing the growth fitness defects of the strains having single-gene deletion mutants versus strains with both genes deleted. The process of construction and quantification of growth fitness

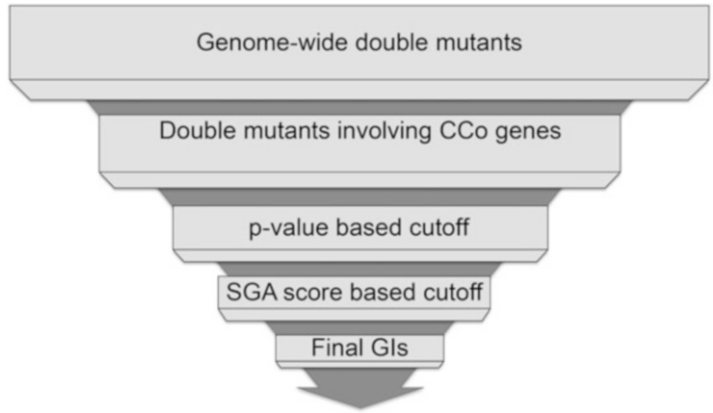

Fig. 1 A flowchart summarizing the construction of CCo GI network extracted from the genome-wide double mutant growth fitness data

of all three mutants (two single and one double) pertaining to two genes is described in [22]. The SGA score for a gene pair is calculated using the following multiplicative model [23].

$$\text{SGA Score} = W_{AQ} - (W_A \times W_Q).$$

where W_{AQ}, W_A, and W_Q are the double, single array, and single query mutant growth fitness values, respectively. A statistical confidence measure (*p*-value) is assigned to each interaction based on a combination of the observed variation of each double mutant across four experimental replicates and estimates of the background lognormal error distributions for the corresponding query and array mutants [22]. The two criteria of, for example, SGA scores $\geq |0.08|$ and *p*-value <0.05 are used to evaluate the strength of a GI. Such GIs are then combined to construct a GI network. In the case of multiple testing (gene pair tested multiple times in different batches) or reciprocal redundancy (gene pair tested as both array-query as well as query-array), the SGA score for that pair with best *p*-value is selected. A succinct schematic description of the GI network construction is shown in Fig. 1.

2.2 Quantifying CCo Interaction Densities

CCos help thousands of substrate proteins fold, assemble, and traffic appropriately. Consequently, CCos are expected to have, on average, a higher number of GIs in comparison to the rest of the yeast genes. To confirm this supposition, we can compare the GI density distribution of CCos with that of all other genes in the whole-genome network [22].

```
#R script to generate density distribution of the GIs of CCos
versus rest of the genes.
# Importing the input file
data <- read.table(file.choose(),header=T,sep="\t")
# Importing the required R libraries
```

```
library(ggplot2)
library(reshape2)
attach(data)
data.m <- melt(data)
p <- ggplot(aes(x=value, colour=variable), data=data.m)
p + geom_density() +
theme_bw() + theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank(), axis.text.x =
element_text(angle=0),legend.position = c(0.8, 0.8)) +
scale_x_continuous(breaks = round(seq(0, 1000, by = 100),1),
name="Number of interactions", expand = c(0, 0)) +
scale_y_continuous(breaks = seq(0,0.0035,0.0005), limits=c
(0,0.004), name="Density", expand = c(0, 0)) +
labs(colour = "Genes") + annotate(geom="text", x=400,
y=0.0015, label="italic(P) < 0.02", parse=TRUE, color-
="black") +
# To add lines to represent the average number of GIs of two
gene sets. For example, 160 and 80 for CCos and all other
genes, respectively.
geom_vline(xintercept = 80, size = 0.5, colour = "black",
linetype = "dashed") + geom_vline(xintercept = 160, size =
0.5, colour = "black",linetype = "dashed")
```

2.3 Analysis of GI Network

2.3.1 Clustering of GI Profiles

Two genes are considered to have similar GI profiles when their set of positive and negative GIs are significantly alike. Two genes with similar profiles can, therefore, be considered functionally associated [22]. A very powerful way to organize genes according to their GI profiles is by applying two-dimensional (2-D) hierarchical clustering. Conceivably, genes pertaining to same pathways and/or complexes are more likely to cluster together. 2-D hierarchical clustering can be performed by using a standalone tool called Cluster 3.0 which can be downloaded from http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm. A detailed manual explaining how to use Cluster 3.0 is also provided. Obtained images can then be visualized in the form of a heatmap using the Java TreeView tool (http://jtreeview.sourceforge.net/). The strength of this approach lies in the functional prediction of genes for which little or no information is available in the literature (also known as orphan genes). If a well-annotated gene is clustered together with an orphan gene through the guilt-by-association principle [24], it can be proposed that those two genes have similar molecular functions.

2.3.2 Bioprocess Enrichment in the GI Network

A bioprocess represents a group of genes that delineate a series of events achieved by one or more coordinated assemblies of molecular functions. The following analysis can be performed to determine

bioprocesses that are significantly enriched for GIs in the GI network indicating their importance. If NETn is the number of genes in the GI network whereas BPn is the number of genes in a bioprocess, we calculate four numbers:

1. The number of observed interactions for genes in a bioprocess G ($\mathrm{BPint}_{\mathrm{obs}}$)

$$|\{(u,v)\in E(G) \mid u\in V(BP)\vee \ v\in V(BP)\}|$$

here, u and v are two of the all (V) genes in the BP and E represents the edges (interactions) in the network.

2. The maximum possible number of interactions for genes in a bioprocess G ($\mathrm{BPint}_{\mathrm{max}}$)

$$BPn(NETn-BPn)+BPn(BPn-1)/2$$

3. The number of actually observed interactions in the GI network G ($\mathrm{NETint}_{\mathrm{obs}}$)

$$|\{(u,v)\in E(G) \mid u\in V(G)\vee \ v\in V(G)\}|$$

4. The maximum possible number of interactions in the GI network G ($\mathrm{NETint}_{\mathrm{max}}$)

$$NETn(NETn-1)/2$$

```
#R script to calculate enrichment
#Typically, large-scale GI studies involve genes from many
bioprocesses. Assuming that there is an input file containing
BPint_obs, BPint_max, NETint_obs and NETint_max values for each
bioprocess separated by tab delimiters.
# Importing the input file
data <- read.table(file.choose(),header=T,sep="\t")
# Function to calculate hypergeometric distribution based
p-values
data.P <- phyper(data$BPint_obs, data$BPint_max, data$NETint_obs -
data$BPint_max, data$NETint_max, lower.tail = FALSE)
# Function to calculate corrected p-values, i.e., false dis-
cover rate (FDR)
data.FDR <- p.adjust(data.P, "fdr")
# Joining the p- and FDR values to the input file
data.P.FDR <- cbind(data, data.P, data.FDR)
# Exporting the calculated values
write.table(data.P.FDR, "Data-P_FDR.txt", sep="\t")
```

In general, a bioprocess with p-value (or FDR) < 0.05 is accepted as enriched.

Table 1
contingency table

	GIs involving BP_1	GIs not involving BP_1
GIs involving BP_2	A	C
GIs not involving BP_2	B	D

Where,
A = Number of GIs between BP_1 and BP_2 genes
B = Number of GIs between BP_1 and non- BP_2 genes
C = Number of GIs between BP_2 and non- BP_1 genes
D = Number of GIs that do not involve BP_1 or BP_2 genes

2.3.3 Bioprocess Crosstalk in the GI Network

Significance of the observed GIs between two bioprocesses can be evaluated using Fisher's Exact test. For that, we make a contingency Table 1:

```
#Assuming that there is an input file containing BP1, BP2, A, B, C and D values for each bioprocess pair separated by tab delimiters, the R script to calculate bioprocess pair enrichment (p-value) is

data <- read.table(file.choose(),header=T,sep="\t")
get_fisher <- function(data){
  mat <- matrix(as.numeric(data[c(3:6)]),nrow=2, ncol=2)
  f <- fisher.test(as.table(mat), alternative="greater")
  return(c(df[1],  f$p.value))
}
P.values   <-  apply(df, 1,  get_fisher)
```

As described above, *p*-value corrections can be performed on the obtained *p*-values. Generally, a bioprocess pair with *p*-value (or FDR) < 0.05 is accepted as enriched. Similarly, we can compute crosstalk enrichment between CCo families.

2.3.4 Building the CCo GI Profile Correlation Similarity Network

The GI profile of a given gene is composed of the list of positive and negative GIs involving that gene across the whole genome. A strong correlation in the GI profile of two genes should indicate high similarity in the pattern of their genetic interactions with other genes in the genome, suggesting similar molecular function or pathway/complex [22]. This property can be used to assess the connectivity between CCos in the cell by building a GI profile correlation similarity network. The mathematical formula to calculate the Pearson correlation coefficient (r) is

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of pairs of data points in the GI profiles, x and y, of two genes. Assuming that we have a list of GIs as three column file (Gene1, Gene2, and SGA score), the following R script can be used to (1) generate the SGA score matrix and then (2) calculate Pearson correlation coefficient values for each gene pair in the matrix.

```
data <- read.table(file.choose(),header=T,sep="\t")
data.mat <- acast(data, Gene1~Gene2, value.var="Score")
 PCC <- cor(data.mat)
write.table(PCC, file="Output-matrix.txt")
```

Similar to GI scores, GI profile correlation can be used to generate the epistatic network. A threshold on the significance of the GI profile correlation scores can be set either by using statistical means such as null distribution-based p-values. The GI profile correlation similarity network can be visualized using the Spatial Analysis of Functional Enrichment (SAFE) tool which is described in detail elsewhere [25]. Briefly, SAFE highlights regions that are densely connected with a particular attribute such as Gene Ontology (GO) or cellular bioprocesses.

2.3.5 Finding Positive and Negative GI Hubs

Hub genes in the network are genes with high number of GIs [26], and are typically central to the network's architecture because of their essential role in the cellular processes. Their functions become even more vital in the differential (or dynamic) network when two static GI networks screened under two different conditions are compared. In a given static network, the number of direct connections a node i (gene or protein) has is referred to as its connectivity degree. When a network is represented as an adjacency matrix M, the degree of gene i is calculated by

$$\sum_{j=1}^{NETn} M(i,j)$$

where $M(i, j)$ is an index representing ith row and jth column of the matrix M. NETn is the number of genes in the network or number of columns in the matrix M. High number of interactions of a gene in the GI network is expected to have an important role. Consequently, CCos with wider role in the network are likely to be hubs. On the other hand, genes interacting with many CCos, especially a particular chaperone family, could be predicted to have close functional association with them. Furthermore, a

dominant type of GI (positive or negative) could further help in deciphering the nature of the potential associations.

Assuming that we have a list of GIs as three column files (Gene1, Gene2, and SGA score), the following R script can be used to calculate the number of positive and negative interactions of each gene in the network.

```
# Read GI file containing gene names and GI score (Gene1,
Gene2, and Score delimited by tab)
data <- read.table(file.choose(),header=T,sep="\t")
#Separating positive and negative GIs
data.neg <- subset(data, data$Score < 0, select = c(Gene1,
Gene2))
data.pos <- subset(data, data$Score > 0, select = c(Gene,
Gene2))


#Importing igraph library to do graph based calculations
library(igraph)
#Generating graph for positive and negative GIs, respectively
g.pos <- graph_from_data_frame(data.pos, directed=F)
g.neg <- graph_from_data_frame(data.neg, directed=F)
# Converting gene degree values into a dataframe
g.pos.degree <- as.data.frame(degree(g.pos))
g.neg.degree <- as.data.frame(degree(g.neg))
library(data.table)
#Use gene names as first column of the dataframe
setDT(g.pos.degree, keep.rownames = TRUE)
setDT(g.neg.degree, keep.rownames = TRUE)
#Assign desired names to the columns in the dataframe
setnames(g.pos.degree, 1, "Genes")
setnames(g.pos.degree, 2, "Degree")
setnames(g.neg.degree, 1, "Genes")
setnames(g.neg.degree, 2, "Degree")
#Merge 2 dataframes with respect to gene names
g.pos.neg.degree <- (merge(g.pos.degree, g.neg.degree, by.
x="Genes", by.y="Genes", sort=F, all =T))
#Assign desired names to the columns in the dataframe
setnames(g.pos.neg.degree, 2, "PositiveDegree")
setnames(g.pos.neg.degree, 3, "NegativeDegree")
#Replace "NA" value with 0
g.pos.neg.degree[is.na(g.pos.neg.degree)] <- 0
#Export the results
write.table(g.pos.neg.degree,"CHap-net-degree-alle-aggr.txt",
sep="\t", row.names = FALSE, quote = FALSE)
```

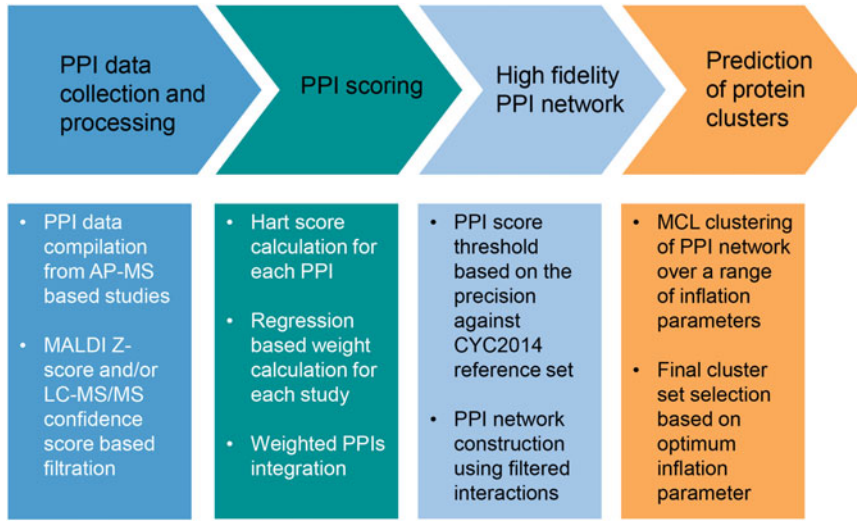

Fig. 2 A computational framework to integrate PPIs obtained from multiple high-throughput studies to construct a high-fidelity PPI network

2.4 Construction of PPI Network from Multiple High-Throughput Studies

Ideally, the integration of PPIs from several high-throughput studies should be performed because it broadens the coverage of the chaperone interactome (Fig. 2). To illustrate the chaperone-based PPI network construction, we can use the following large-scale proteomic studies, where interactions can be restricted to a bait (or target) CCo protein: Gavin et al. [5], Krogan et al. [27], Wodak et al. [28], Babu et al. [29], and Gong et al. [20]. These studies utilized MALDI-TOF (Matrix-assisted Laser Desorption/Ionization Time of Flight) MS and/or tandem liquid chromatography (LC-MS/MS) based confidence probability scores [30] for protein identification. Briefly, the confidence scores are calculated as the probability of a prey protein (or peptide), suggesting the likelihood of its appearance in the purifications pertaining to a bait. Gold standard literature reference set of PPIs from BioGRID (https://thebiogrid.org/) can be used to calculate an optimum probability score as a threshold at a high precision value. Specific thresholds of these two (typically, Z-score ≥ 1 for MALDI-TOF/MS, LC-MS/MS and confidence score $> 70\%$) scores can be used to discard low confidence PPI detections [27, 31].

In order to make use of the compiled PPI datasets, an appropriate relative weighting must be performed as the different datasets may have different scoring methods, which can lead to a scoring bias. In the case of chaperone PPIs, the purification enrichment (PE) and the hypergeometric Hart interaction scores [32, 33] are computed and compared by selecting the method that yields the highest number of CCos. In the case of the Hart score, an integrated PPI score can be computed by summing the relative

weights from each dataset as follows: Gavin_Hart × γ1 + Babu_-Hart × γ2 + Krogan_Hart × γ3 + Gong_Hart × γ4, where γ is the weight of an individual dataset obtained by applying logistic regression. The γ value varies according to the influence of each dataset on the overall precision of the PPI scoring.

To determine an optimum threshold of the Hart score, PPIs can be compared to a high-confidence experimentally validated set of protein complexes from a CYC2014 gold standard reference set [34]. An individual dataset can increase or decrease the overall precision and, therefore, its γ value can be fine-tuned to obtain a score cutoff, where coverage of interactions is maximized while maintaining a high precision value. We find the Hart scoring method to be a better predictor of CCo PPIs.

2.5 Prediction of Protein Complexes Using Clustering Algorithm

Since densely connected regions of a PPI network suggest that associated proteins are likely to have a similar function [28], clustering methods have been used to identify and predict protein complexes and functional modules [35]. The Markov clustering method (MCL) can be used [36] to identify the macromolecular assemblies within the CCo PPI network. The resulting clusters can be benchmarked based on the overall cluster properties such as the number of clusters, average cluster size, intra- and inter-cluster functional diversity as measured by Shannon index of gene ontology (GO) terms in biological process and molecular function [37], as well as CYC2014 [34] complex coverage through precision and homogeneity metric [38]. Based on the minimization of intra-cluster average functional diversity and coverage of known CYC2014 complex members, an inflation parameter can be chosen to generate the finalized PPI clustering. An MCL clustering algorithm tool can be downloaded from http://micans.org/mcl/ and can be run from command line. An example of the application of this tool is shown below:

```
mcl <-|InputFile> --abc -o OutputFile
```

When running this program, the inflation value is usually chosen in the range of 1.2–5.0.

2.5.1 Quality Assessment of Predicted Protein Complexes

A substantial overlap of a predicted cluster with one or more high-confidence literature-curated protein complexes (CYC2014) is a measure of high quality [27]. Here, we provide a detailed explanation on how to do this analysis. Assuming that there are c predicted clusters ($C_1 \ldots C_c$) and m CYC2014 complexes ($CYC_1 \ldots CYC_m$), we construct a $m \times c$ matrix A (also called confusion matrix) where rows represent the number of common proteins in each of the CYC_i complexes in CYC2014 with the C_j clusters and columns representing the number of common proteins in each of the C_j clusters with that of CYC2014 complexes.

$$A = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mc} \end{pmatrix} = \left(p_{ij}\right) \in \mathbb{N}^{m \times c}$$

where p_{ij} represents an index of the matrix A.

The following four quantities are then computed:

1. S_i is the sensitivity that quantifies the extent by which a CYC2014 complex CYC_i aggregated within the same predicted cluster.

$$S_i = \max_j\left(p_{ij}\right) / \sum_{j=1}^{c} p_{ij}$$

2. H_i^{CYC} is the homogeneity that quantifies the extent by which a CYC2014 complex is distributed among predicted clusters.

$$H_i^{\text{CYC}} = \sum_{j=1}^{c} \left(p_{ij} / \sum_{j=1}^{c} p_{ij}\right) \cdot \left(p_{ij} / \sum_{i=1}^{m} p_{ij}\right)$$

3. Positive predicted value, PPV_j, of a predicted cluster determines the maximum portion of it being part of a CYC2014 complex

$$\text{PPV}_j = \max_i\left(p_{ij}\right) / \sum_{i=1}^{m} p_{ij}$$

4. H_j^C is the homogeneity of a predicted cluster calculating the extent to which it is distributed among CYC2014 complexes.

$$H_j^C = \sum_{i=1}^{m} \left(p_{ij} / \sum_{i=1}^{m} p_{ij}\right) \cdot \left(p_{ij} / \sum_{j=1}^{c} p_{ij}\right)$$

These four quantities are then used to calculate overall agreement between CYC2014 complexes and predicted clusters represented by $\text{Precision}_{\text{total}}$ and $\text{Homogeneity}_{\text{total}}$ which are defined as:

$$\text{Precision}_{\text{total}} = \text{sqrt}(S_{\text{mean}} \times PPV_{\text{mean}})$$
$$\text{Homogeneity}_{\text{total}} = \text{sqrt}\left(H_{\text{mean}}^{\text{CYC}} \times H_{\text{mean}}^{C}\right)$$

Here, S_{mean} and PPV_{mean} are the averages of S_i and PPV_j values across the columns and rows, respectively. $H_{\text{mean}}^{\text{CYC}}$ and H_{mean}^{C} are the averages of all the H_i^{CYC} and H_j^C values, respectively. Below is an R script to calculate the overall precision and homogeneity as described above.

```
#Read an input file containing a matrix in which each cell
represents the overlapping proteins between a literature
complex (rows) and predicted MCL cluster (columns)

data <- (read.table("mat.txt",sep="\t", header=T, row.names=1))

#Function to calculate Precision_total and Homogeneity_total

Preci.homog <- function (D){
M <- as.matrix(D, nrow = 1, ncol = 1, byrow = FALSE, dimnames
= NULL, row.names = 1)

HCmean <- mean(colSums(t(apply(M, 1, function(i) i/sum(i)))
*apply(M, 2, function(i) i/sum(i))))

HMmean <- mean(rowSums(t(apply(M, 1, function(i) i/sum(i)))
*apply(M, 2, function(i) i/sum(i))))

Smean <- mean(apply(M,1,max)/(rowSums(M)))

PPVmean <- mean(apply(M,2,max)/(colSums(M)))

Precision.tot <- sqrt(Smean*PPVmean)

Homogeneity.tot <- sqrt(HMmean*HCmean)

newlist <- list("Precision_total" = Precision.tot, "Homoge-
neity_total" = Homogeneity.tot)

return(newlist)
}

#Calling the function for input matrix
Preci.homog(data)
```

2.6 Building a Combined Physical-Genetic Interaction Network

Deciphering the functional relationships among proteins is essential to comprehend all facets of cell biology. Typically, proteins in the cellular environment work as complexes and are part of a multidimensional proteome [21]. In Subheading 2.3.1, we described an elegant approach to predict protein complexes or pathways by applying hierarchical clustering to the GI profiles. Experimentally identified or predicted protein complexes on their own would not exhibit connections with each other. An auxiliary course to fill in this knowledge gap is to use GI information to investigate association between already known or predicted protein complexes [39]. Furthermore, these techniques help in the prediction of new members of complexes since a gene holding high GI profile correlation with most of the complex members is likely to be

another component of the complex. To establish functional connections between protein complexes, the GI profile similarity network can be overlaid onto the predicted MCL clusters obtained from the physical PPIs. This data integration is essential as the PPI network cannot establish genetic (or epistatic) connections between complexes. Hence, integrating GI with PPI provides an additional layer of information to the network.

By using the average GI profile similarity scores between genes (proteins) of two clusters as a connectivity score, a quantitative inter-complex connection can be established (Fig. 3). At the intra-complex level, proteins are connected using their respective GI profile similarity metrics. Only positive correlations among genes within the same complex are considered to be meaningful in this analysis.

3 Concluding Remarks

Our stepwise strategy provides a computational framework for capturing PPI and GI data on CCos. Given their promiscuous and typically transient folding functions in the cell, most CCo interactions tend to be difficult to obtain using affinity purification

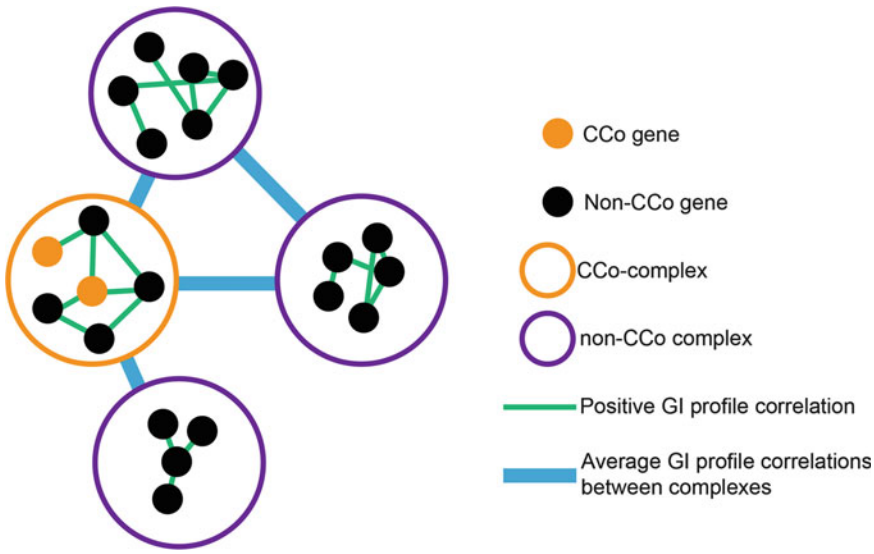

Fig. 3 Schematic overview of mapping GI profile correlations onto predicted protein clusters. Functional relationships between predicted protein clusters are obtained by averaging the GI profile correlation scores among their (gene) members. Similarly, positive GI profile correlation scores are used to enhance the envisaged functional connections within a cluster

(AP)/MS methods. Furthermore, some CCo genes, like Hsp90, are pleiotropic and are considered prototypical capacitors of genetic variation in many organisms [40–42]. This complicates their interaction study using GI data since pleiotropic genes tend to have many GIs with different genes from multiple pathways. Therefore, even though pleiotropic genes often are hubs in the GI network, they seldom show functional enrichment with specific pathways/processes. Hence, by combining both the GI and the PPI data, a global CCo functional pattern can be elucidated where complexes containing specialized CCos have more inter-complex connections compared to complexes containing functionally general CCos.

The computational approaches described here can generate a comprehensive and high-fidelity CCo network that exposes the global functional role of CCos in protein homeostasis. This can serve as a powerful resource for anyone studying CCo interactions from any organism as we recently did for yeast CCos [43].

Acknowledgements

K.R. was supported by a Canadian Institutes of Health Research (CIHR) Training Program in Protein Folding and Interaction Dynamics: Principles and Diseases fellowship and by a University of Toronto Fellowship in the Department of Biochemistry. M.B. holds a CIHR New Investigator award (MSH-130178). This work was funded by CIHR grants MOP-125952, RSN-124512, 132191, and FDN-154318 and MOP-132191 to M.B. and by MOP-93778, MOP-81256, and MOP-130374 to W.A.H.

References

1. Saibil H (2013) Chaperone machines for protein folding, unfolding and disaggregation. Nat Rev Mol Cell Biol 14:630–642
2. Balchin D, Hayer-Hartl M, Hartl FU (2016) In vivo aspects of protein folding and quality control. Science 353:aac4354
3. Finka A, Mattoo RU, Goloubinoff P (2016) Experimental milestones in the discovery of molecular chaperones as polypeptide unfolding enzymes. Annu Rev Biochem 85:715–742
4. Graves PR, Haystead TA (2002) Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev 66:39–63
5. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147
6. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaino JA (2015) Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics 15:930–949
7. Martens L, Vizcaino JA (2017) A golden age for working with public proteomics data. Trends Biochem Sci 42:333–341
8. Aebersold R, Mann M (2016) Mass-spectrometric exploration of proteome structure and function. Nature 537:347–355

9. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. Nat Rev Genet 8:437–449
10. Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, Stewart G, Samanfar B, Aoki H, Wagih O, Vlasblom J, Phanse S, Lad K, Yeou Hsiung YA, Graham C, Jin K, Brown E, Golshani A, Kim P, Moreno-Hagelsieb G, Greenblatt J, Houry WA, Parkinson J, Emili A (2014) Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in Escherichia Coli. PLoS Genet 10:e1004120
11. Kumar A, Beloglazova N, Bundalovic-Torma C, Phanse S, Deineko V, Gagarinova A, Musso G, Vlasblom J, Lemak S, Hooshyar M, Minic Z, Wagih O, Mosca R, Aloy P, Golshani A, Parkinson J, Emili A, Yakunin AF, Babu M (2016) Conditional epistatic interaction maps reveal global functional rewiring of genome integrity pathways in Escherichia Coli. Cell Rep 14:648–661
12. Peters JM, Colavin A, Shi H, Czarny TL, Larson MH, Wong S, Hawkins JS, CH L, Koo BM, Marta E, Shiver AL, Whitehead EH, Weissman JS, Brown ED, Qi LS, Huang KC, Gross CA (2016) A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell 165:1493–1506
13. Koo BM, Kritikos G, Farelli JD, Todor H, Tong K, Kimsey H, Wapinski I, Galardini M, Cabal A, Peters JM, Hachmann AB, Rudner DZ, Allen KN, Typas A, Gross CA (2017) Construction and analysis of two genome-scale deletion libraries for Bacillus subtilis. Cell Syst 4:291–305. e297
14. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet 38:896–903
15. Yu J, Pacifico S, Liu G, Finley RL Jr (2008) DroID: the drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics 9:461
16. Fischer B, Sandmann T, Horn T, Billmann M, Chaudhary V, Huber W, Boutros M (2015) A map of directional genetic interactions in a metazoan cell. eLife 4:e05464
17. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. Science 303:808–813
18. Zhao R, Davey M, Hsu YC, Kaplanek P, Tong A, Parsons AB, Krogan N, Cagney G, Mai D, Greenblatt J, Boone C, Emili A, Houry WA (2005) Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. Cell 120:715–727
19. Zhao R, Houry WA (2007) Molecular interaction network of the Hsp90 chaperone system. Adv Exp Med Biol 594:27–36
20. Gong Y, Kakihara Y, Krogan N, Greenblatt J, Emili A, Zhang Z, Houry WA (2009) An atlas of chaperone-protein interactions in Saccharomyces cerevisiae: implications to protein folding pathways in the cell. Mol Syst Biol 5:275
21. Larance M, Lamond AI (2015) Multidimensional proteomics for cell biology. Nat Rev Mol Cell Biol 16:269–280
22. Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, Pelechano V, Styles EB, Billmann M, van Leeuwen J, van Dyk N, Lin ZY, Kuzmin E, Nelson J, Piotrowski JS, Srikumar T, Bahr S, Chen Y, Deshpande R, Kurat CF, Li SC, Li Z, Usaj MM, Okada H, Pascoe N, San Luis BJ, Sharifpoor S, Shuteriqi E, Simpkins SW, Snider J, Suresh HG, Tan Y, Zhu H, Malod-Dognin N, Janjic V, Przulj N, Troyanskaya OG, Stagljar I, Xia T, Ohya Y, Gingras AC, Raught B, Boutros M, Steinmetz LM, Moore CL, Rosebrock AP, Caudy AA, Myers CL, Andrews B, Boone C (2016) A global genetic interaction network maps a wiring diagram of cellular function. Science 353:aaf1420
23. Baryshnikova A, Costanzo M, Kim Y, Ding H, Koh J, Toufighi K, Youn JY, Ou J, San Luis BJ, Bandyopadhyay S, Hibbs M, Hess D, Gingras AC, Bader GD, Troyanskaya OG, Brown GW, Andrews B, Boone C, Myers CL (2010) Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. Nat Methods 7:1017–1024
24. Oliver S (2000) Guilt-by-association goes global. Nature 403:601–603
25. Baryshnikova A (2016) Systematic functional annotation and visualization of biological networks. Cell Syst 2:412–421

26. Boucher B, Jenna S (2013) Genetic interaction networks: better understand to better predict. Front Genet 4:290
27. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:637–643
28. Wodak SJ, Pu S, Vlasblom J, Seraphin B (2009) Challenges and rewards of interaction proteomics. Mol Cell Proteomics 8:3–18
29. Babu M, Vlasblom J, Pu S, Guo X, Graham C, Bean BD, Burston HE, Vizeacoumar FJ, Snider J, Phanse S, Fong V, Tam YY, Davey M, Hnatshak O, Bajaj N, Chandran S, Punna T, Christopolous C, Wong V, Yu A, Zhong G, Li J, Stagljar I, Conibear E, Wodak SJ, Emili A, Greenblatt JF (2012) Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature 489:585–589
30. Babu M, Krogan NJ, Awrey DE, Emili A, Greenblatt JF (2009) Systematic characterization of the protein interaction network and protein complexes in Saccharomyces cerevisiae using tandem affinity purification and mass spectrometry. Methods Mol Biol 548:187–207
31. Babu M, Kagan O, Guo H, Greenblatt J, Emili A (2012) Identification of protein complexes in Escherichia coli using sequential peptide affinity purification in combination with tandem mass spectrometry. J Vis Exp 69:e4057
32. Hart GT, Lee I, Marcotte ER (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. BMC Bioinformatics 8:236
33. Zhang B, Park BH, Karpinets T, Samatova NF (2008) From pull-down data to protein interaction networks and complexes with biological relevance. Bioinformatics 24:979–986
34. Pu S, Wong J, Turner B, Cho E, Wodak SJ (2009) Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 37:825–831
35. Wang J, Li M, Deng Y, Pan Y (2010) Recent advances in clustering methods for protein interaction networks. BMC Genomics 11 (Suppl 3):S10
36. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584
37. Loganantharaj R, Cheepala S, Clifford J (2006) Metric for measuring the effectiveness of clustering of DNA microarray expression. BMC Bioinformatics 7(Suppl 2):S5
38. Pu S, Vlasblom J, Emili A, Greenblatt J, Wodak SJ (2007) Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. Proteomics 7:944–960
39. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature 446:806–810
40. Rutherford SL, Lindquist S (1998) Hsp90 as a capacitor for morphological evolution. Nature 396:336–342
41. Queitsch C, Sangster TA, Lindquist S (2002) Hsp90 as a capacitor of phenotypic variation. Nature 417:618–624
42. Karras GI, Yi S, Sahni N, Fischer M, Xie J, Vidal M, D'Andrea AD, Whitesell L, Lindquist S (2017) HSP90 shapes the consequences of human genetic variation. Cell 168:856–866
43. Rizzolo K, Huen J, Kumar A, Phanse S, Vlasblom J, Kakihara Y, Zeineddine HA, Minic Z, Snider J, Wang W, Pons C, Seraphim TV, Boczek EE, Alberti S, Costanzo M, Myers CL, Stagljar I, Boone C, Babu M, Houry WA (2017) Features of the chaperone cellular network revealed through systematic interaction mapping. Cell Rep 20:2735–2748