

Noël Malod-Dognin¹ / Nataša Pržulj¹

Omics Data Complementarity Underlines Functional Cross-Communication in Yeast

¹ University College London, Department of Computer Science, London, UK, E-mail: natasa@cs.ucl.ac.uk**Abstract:**

Mapping the complete functional layout of a cell and understanding the cross-talk between different processes are fundamental challenges. They elude us because of the incompleteness and noisiness of molecular data and because of the computational intractability of finding the exact answer. We perform a simple integration of three types of baker's yeast omics data to elucidate the functional organization and lines of cross-functional communication. We examine protein-protein interaction (PPI), co-expression (COEX) and genetic interaction (GI) data, and explore their relationship with the gold standard of functional organization, the Gene Ontology (GO). We utilize a simple framework that identifies functional cross-communication lines in each of the three data types, in GO, and collectively in the integrated model of the three omics data types; we present each of them in our new Functional Organization Map (FOM) model. We compare the FOMs of the three omics datasets with the FOM of GO and find that GI is in best agreement with GO, followed COEX and PPI. We integrate the three FOMs into a unified FOM and find that it is in better agreement with the FOM of GO than those of any omics dataset alone, demonstrating functional complementarity of different omics data.

Keywords: Molecular omics data, cell's functional organization, data-integration**DOI:** 10.1515/jib-2017-0018**Received:** March 22, 2017; **Revised:** April 18, 2017; **Accepted:** April 18, 2017

1 Introduction


1.1 Motivation

Genes and their corresponding protein products underlay the functioning of the cell. While dedicated studies allowed us to understand the chains of molecular interactions that lead to specific biological functions, such as biological pathways described in KEGG database [1], we lack models capturing the whole functional organization of a cell stemming collectively from all data types [2]. Among many model organisms, yeast is the most commonly used in systems biology, due to the ease of its experimental manipulation [3], [4]. Thanks to the advances in capturing technologies, large amounts of complex molecular data have been produced for yeast, including its complete genome [5], the complete set of its open reading frames [6], [7], and large scale omics data from genetic, genomic, proteomic, transcriptomics, and metabolomic studies [8], [9], [10], [11].

Among these data, we focus on protein-protein interactions, gene co-expressions and genetic interactions. Proteins perform their functions through binding to other molecules. Protein-protein interactions encode the pairwise bindings between proteins, as captured by methods such as yeast two hybrid [12], [13], or affinity capture coupled with mass spectrometry [14]. Also, the expression of a gene can be measured over time with experiments such as RNAseq [15]. When the expressions of two genes are significantly correlated over time, such genes are said to be co-expressed. Finally, two genes are said to genetically interact when a double mutant shows a significant deviation in fitness compared to the expected multiplicative effect of combining two single mutants [16], [17]. Negative interactions represent a more severe fitness defect than expected, with the extreme case being synthetic lethality, while positive interactions represent a less severe fitness defects than expected.

To understand functions of genes, these different omics data are modelled by networks, where nodes represent genes and two nodes are connected by an edge if the two genes are known to interact in some way, e.g. via a physical binding between the corresponding proteins. It has been shown that there exist relationships between the wiring patterns in networks of protein-protein interactions (PPIs) and the biological functions of proteins. Proteins with similar functions and cellular locations tend to cluster together in the PPI network of yeast [18] and several network-based approaches were proposed to predict the functions of proteins based on the "guilt by association" principle, where the functions of proteins can be transferred with statistical confidence to other

Nataša Pržulj is the corresponding author.

 ©2017, Nataša Pržulj et al., published by De Gruyter.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

proteins if they directly interact, or share common neighbourhoods [19], [20], [21], [22], [23]. Also, it has been shown that the functional similarities between proteins do not only depend on them being in the same local neighbourhood, but also on the similarities of their interaction patterns independent of the PPI network location [24] and that the number of functions of the proteins correlate with their degrees (numbers of neighbours) in the PPI network [25].

Nowadays, approximately 90 % of yeast's genes have some functional annotations [4], [26] and linking proteins to their cellular functions has possibly been the most important contribution of yeast as a model organism [27]. Yet, despite the abundance of biological and functional data, the functional organization of the cell remains largely unknown. Different molecular data are known to carry different information about a biological system [28], [29], offering multiple views into the system [30], [31]. Because a single data type is likely to reveal limited insight into the cell's functional organization, combining multiple data types has a potential to yield a more complete insight than those obtained from individual sources of data in isolation [31], [32]. Recently, data integration methods were applied to predict protein function [33], to identify common genetic phenotypes [34], and to predict the missing links in gene and disease ontologies [35], [36]. These motivated us not only to study the cell's functional organization from each molecular interaction dataset alone, but also to integrate them and propose a unified functional organization model of the yeast's cell.

1.2 Contribution

We make several steps toward describing and understanding the functional organization of the yeast's cell, using the three most comprehensive molecular interaction datasets: protein-protein interaction (PPI), co-expression (COEX) and genetic interaction (GI) networks. First, we explore the relationships between these interaction networks and the Gene Ontology (GO) annotations of the genes. We observe that PPI data, which capture direct physical bindings between proteins, COEX data, which capture the regulatory processes that are shared between genes, and GI data, which tend to capture indirect interactions often along alternative pathways [28], are only weakly functionally related with GO (Section 3.1). Second, we use a benchmark of fourteen GO-derived biological functions and apply receiver operating characteristics (ROC) curve analysis to quantify the "separations" of these biological functions in each of PPI, COEX and GI data. The low areas under the ROC curves that we obtained suggest that biological functions are not well separated in the "functional space", but are interlaced in each of the omics datasets, posing a question of identification of lines of functional cross-communication (see Section 3.2).

Hence, we build upon this observation and propose a new simple model of the cell's functional organization. For a given dataset, we measure the strength of the communication lines between regions containing different biological functions. This results in a square matrix encoding the cell's functional cross communication lines, which we term functional organization map (FOM). We compute the functional organization maps for PPI, COEX and GI datasets, and compare them to the reference functional organization map, that we compute from the semantic similarities between the GO annotations of genes. We observe that the functional map of GI best fits the reference GO-based one, followed by the COEX and by the PPI map (see Section 3.2).

Because each of the three omics datasets captures a different aspect of the cell's functional organization, we integrate the functional organization maps of COEX, GI and PPI network data into a single, unified, functional organization map of yeast, which correlates the best with the reference GO-based functional organization map, validating our data integration approach and indicating the complementarity of different omics datasets in capturing cellular functioning (Section 3.3).

2 Method

2.1 Data

In this study, we use the following three large scale molecular interaction datasets for yeast.

Protein-protein Interactions (PPIs). We use all the experimentally validated protein interactions from IID database [37] (downloaded in June, 2016). We model PPI data as a network where nodes are proteins and two nodes are connected by an edge if the two proteins can interact. The corresponding PPI network has 5723 nodes and 108,484 interaction edges. In the PPI network, we measure the *functional similarity* between two proteins with the overlap of their neighbourhoods, i.e., with Jaccard Similarity (JS), as was commonly done before [38]: $JS(i, j) = (n_i \cap n_j) / (n_i \cup n_j)$, where n_i is the set of the neighbours of protein i and n_j is the set of neighbours of protein j in the PPI network.

Gene Co-expression Data (COEX). We use the large data set from COXPRESdb [39] (downloaded in January, 2016), where co-expressions of genes are uniformly normalized and integrated. In COXPRESdb, the similarity between two genes is measured by the Pearson's correlation of their expression profiles. We use the absolute value of these correlation to measure the *functional similarity* in COEX dataset. To measure the overlap between PPI, COEX and GI datasets (Section 3.4), we need to represent COXPRESdb data as an unweighted, undirected network (as PPI data are in that format), which we do by connecting a gene to its top 1% most co-expressed genes. The resulting COEX network has 4432 nodes and 151,510 edges.

Genetic Interactions (GIs). We follow the approach of Costanzo et al. [11], in which each gene is characterized by a genetic profile, which is the vector of the gene's interaction values (based on the deviations of their fitness) with all other genes. GIs encode the similarity between the genetic profiles (e.g. if two genes have similar genetic interactions with all the other genes), as measured by the Pearson's correlation between their genetic profiles. Here, we use the latest, most complete set of GIs from Boone's lab [11]. We measure the *functional similarity* between two genes with the absolute value of the correlations between their genetic profiles, as we wish to capture both positive and negative correlations. To measure the overlap between PPI, COEX and GI datasets (Section 3.4), we model GI as an unweighted, undirected network by connecting two genes by an edge if the absolute value of the correlation between their profiles is ≥ 0.05 (the value suggested by Costanzo et al. [11]). The corresponding GI network has 4746 nodes and 62,320 interactions. We also tested the two other suggested thresholds, namely 0.15 and 0.2, but they produced smaller, disconnected networks, so we used the threshold of 0.05.

Gene Ontology (GO) [40]. GO is the gold standard that is used to study functional organization in biological networks [41], [42]. We use GO biological process annotations (from NCBI's Entrez web-portal, accessed in January, 2016) as a benchmark model of the cell's functional organization, against which we compare PPI, COEX and GI data. We measure the *functional similarity* between the GO annotations of two genes by using the semantic similarity of the following type. There exist two types of semantic similarities. Node-based semantic similarity defines the information content of a term as a function of its frequency of appearance in the annotated dataset and measures the similarity between two terms according to their most informative ancestor in the ontology [43], [44]. Edge-based semantic similarity only uses the ontology, directed acyclic graph, and measures the similarity between two terms based on the shortest path between them, or based on the depth in the ontology of their common ancestors [45]. We use Resnik (node-based) semantic similarity [44], because it achieves higher and more consistent correlations with molecular interaction data than other approaches [46], [47], [48].

2.2 Measuring the Agreements of Molecular Interactions with GO

To see how well the three above-described molecular interaction datasets correspond to the complete set of known GO annotations of genes and their relations, we compare the functional similarities of genes in omics data with those of GO. We measure the fit between functional similarities using Pearson's correlation coefficient, which measures the linear dependence between two functional similarity measures. i.e. it measures how well one functional similarity measure can be expressed as a linear combination of the other one. Also, we use Spearman's correlation coefficient, which measures less stringent monotonic dependence between the functional similarity measures. Since the common belief is that larger molecular interaction similarity between genes implies larger functional similarity between the genes, we are expecting molecular interaction similarities and semantic similarity to be positively correlated.

2.3 Measuring the Separation of Biological Functions

We hypothesise that all biological entities exist in a multi-dimensional functional space [49], which is only partially captured by the functional similarities of genes in a given molecular interaction data type. To measure how separated biological functions in this space are, we use Receiver Operating Characteristics (ROCs) analysis [50] between the *functional similarities* of genes in omics datasets (defined in Section 2.1 for each omics dataset) and the fourteen GO-derived functional groups, as defined by Costanzo et al. [28] (these fourteen functional groups are listed in Figure 1). The standard ROC analysis that we use is as follows. For each threshold $\theta \in [0,1]$ of omics functional similarity, we compute four values as follows.

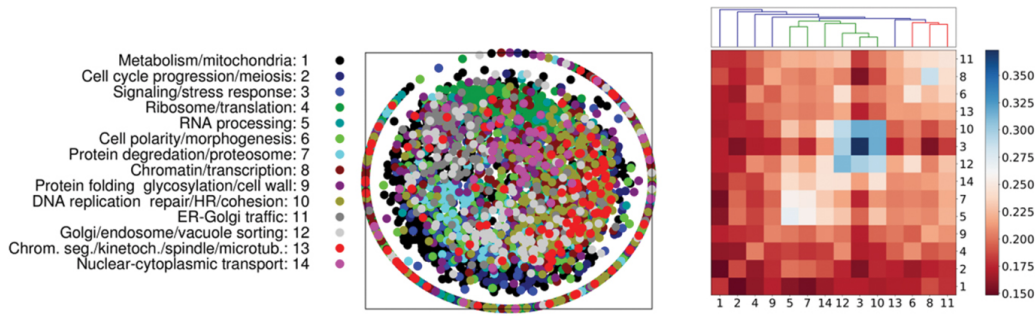


Figure 1: An illustration of our approach for constructing a functional organization map for COEX data. Left: the 14 colored functions that we use to group genes. Middle: a 2D embedding of the colour-coded genes (according to their function) based on COEX dataset (the larger is the absolute value of their co-expression, the closer are two genes in 2D space). While the plot shows some functionally coherent regions, functions are heavily interlaced. Right: the functional organization map (heat map) with hierarchically clustered functions. Blue means high omics functional similarity, i.e. low distance between functions. Due to heavy interlacing of biological functions illustrated in the middle panel, the average functional similarity of genes in one functional groups (measured as described in Section 2.1 for each of the omics data set) are not high, hence we do not have a blue diagonal. Also, note that if the diagonal was blue, that would mean that the biological functions are completely separated from each others, which is not the case in biology.

- The number of true positives, TP, is the number of gene pairs having functional similarity $\geq \theta$ and their two genes coming from the same functional group.
- The number of false positives, FP, is the number of gene pairs having functional similarity $\geq \theta$ and their two genes coming from different functional groups.
- The number of true negatives, TN, is the number of gene pairs having functional similarity $< \theta$ and their two genes coming from different functional groups.
- The number of false negatives, FN, is the number of gene pairs having functional similarity $< \theta$ and their two genes coming from the same functional group.

The ROC curve plots the true positive rate, $TPR = TP / (TP + FN)$, against the false positive rate, $FPR = FP / (FP + TN)$, over all θ . The Area Under the ROC Curve (AUC) represents the probability that two randomly chosen genes belonging to the same functional group will have a higher molecular interaction similarity than two randomly chosen genes belonging to different functional groups. For a given dataset, a large AUC means that the functional groups are well separated in functional space defined by the omics *functional similarities* (Section 2.1), while a low AUC means that functional groups are interlaced ($AUC = 0.5$ is the expected result of a random classifier).

2.4 Characterizing the Cell's Functional Organization

We present a method to untangle the interlacing of biological functions in the cell's functional space, which enables us to map the functional layouts of various interaction data (illustrated in the right panel of Figure 1). Within a given omics dataset, we measure the proximity P_{ij} of two functional groups (out of the 14 described in Section 2.3), i and j , with the average of the omics functional similarity of their genes. The larger P_{ij} , the stronger the communication line between functions i and j . Also, for the fourteen GO-derived functional groups described in Section 2.3 (listed in Figure 1), we construct a 14×14 matrix of average Resnik similarities of GO annotations of genes belonging to the functional groups. Then, we cluster the matrix of pairwise proximities of each of the 14×14 heat maps to reveal the patterns of communication lines between biological functions in the omics data and in the GO-derived gold-standard of functional organization. Deciding which clustering method is best for a given data set with no a priori knowledge of what the answer should be is a hard problem with no right or wrong answers. We chose hierarchical clustering (WPGMA [51]) as it yields groups of inter-functional links in descending order of their communication strength, from strongest to weakest (the heat map and dendrogram in Figure 1, right panel).

3 Results and Discussion

3.1 Agreements of Molecular Interactions with GO

First, we measure how the three molecular datasets relate to GO, using the fitting strategy presented in Section 2.2. The results, presented in Figure 2, show that the three interaction datasets only weakly correlate with GO: COEX achieves the highest Pearson's correlation coefficient (0.355), followed by GI (0.106) and PPI (0.048), with all correlations being statistically significant, as measured by using the F-test (p -values $\leq 10^{-20}$). When measuring the agreement with GO using Spearman's correlation coefficient, COEX achieves again the highest correlation (0.440), followed by PPI (0.199) and GI (0.170). Similar low correlations have been observed by Sevilla et al. [46] and are expected because the data are both noisy and incomplete. These low agreements between molecular interaction similarities and functional similarities motivate us to study the separation of the biological functions in the molecular interaction space.

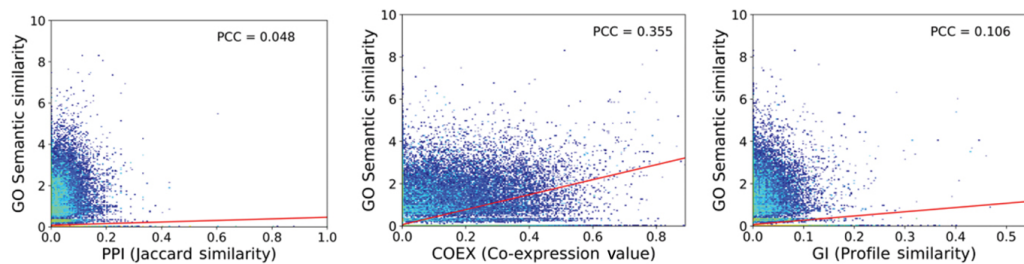


Figure 2: Correlations with GO. Correlations between molecular interaction and semantic similarity based gene similarities. COEX data (middle) best fits to GO with Pearson's correlation coefficient (PCC) of 0.355, followed by GI data (right, with PCC of 0.106) and PPI data (left, with PCC of 0.048). The heat maps are based on the number of gene pairs in the corresponding region: yellow/green areas contain more gene pairs than blue areas; white areas contain no pairs.

3.2 Functional Organization of the Interaction Data

We measure the spatial separation of 14 biological functions in our molecular interaction datasets using the ROC curve analysis detailed in Section "Measuring the Separation of Biological Functions". AUCs detect some functional separation, with AUC = 0.577 for COEX, AUC = 0.552 for PPI, and AUC = 0.535 for GI. While these AUCs are statistically significant according to Mann-Whitney U -test (p -values $\leq 10^{-20}$), the fact that they are all close to 0.5 indicates a lack of functional separation in the interaction space of all three data sets (Figure 3). This means that biological functions are interlaced in these omic data, and that the existing clustering methods based solely on distances between genes cannot discern the cell's functional organization. This interlacing of the biological functions also suggests the presence of strong communication lines between the functions and supports our approach to characterize the functional organization of the cell.

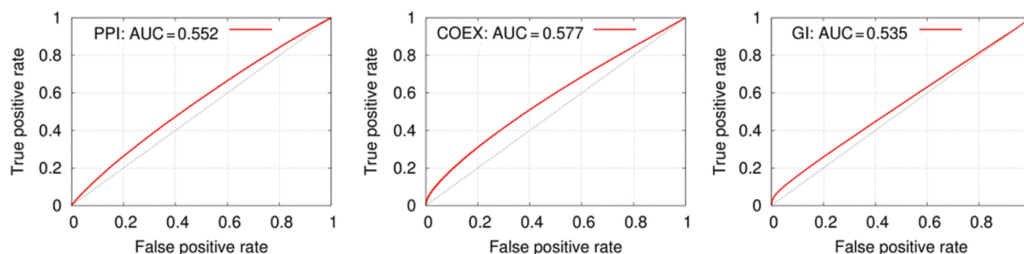


Figure 3: ROC curves indicate a weak separation of the 14 functional groups in GI, COEX and PPI data.

We generate the functional organization maps of the three omics dataset as detailed in Section 2.4 (illustrated in the right panel of Figure 1). The functional communication lines that we obtain in each omics data have differently structured organization, as shown by different clusters in the functional organization maps of PPI, COEX and GI data (blue clusters in the first three heat maps in Figure 4). We find that the strongest functional links in the COEX data are between Signalling/stress response, DNA replication repair/HR/cohesion and Golgi/endosome/vacuole sorting, while GI data shows the strongest links between Cell polarity/morphogenesis and Protein folding glycosylation/cell wall. In the PPI data, the strongest connections are between Cell polarity/morphogenesis, Protein degradation/proteasome and Chromatin/transcription. All these are in

agreement with known biology. For example, Protein degradation, through the ubiquitin pathway, plays important roles in a broad array of basic cellular processes, including cell differentiation and development [52]. Also, the proteasome directly regulates the structure and function of chromatin and chromatin regulatory proteins, and influences gene transcription [53]. All these are highly complex, temporally controlled and tightly regulated process, and these dynamic changes require rapid exchange of information, which explains why they are best seen in the PPI network.

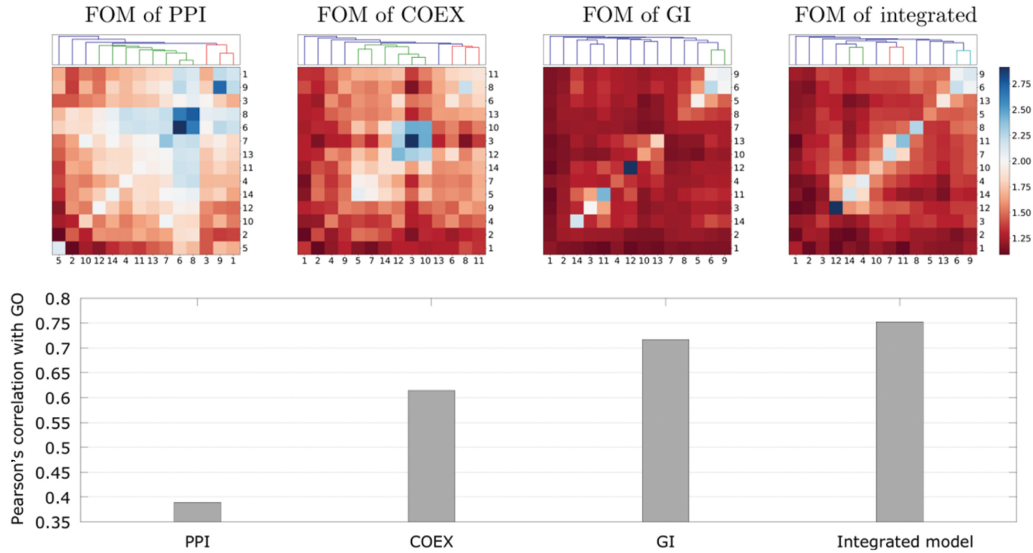


Figure 4: The fit of the functional organization maps obtained from COEX, GI and PPI data to the reference model of functional organization obtained from GO. Y-axis is the value of the Pearson's correlation coefficient between the functional organization maps of the interaction data (x-axis) and GO. Above each box plot is the corresponding functional organization map. The correlations for the three data sets are (from left to right): 0.389 for PPI (p -value 4.02×10^{-5}), 0.614 for COEX (p -value 3.21×10^{-12}) and 0.717 for GI (p -value 7.88×10^{-18}). The last point on x-axis shows the fit of the functional organization map of the integrated model of the interaction data with the functional organization map of GO (0.752, p -value $\approx 2.19 \times 10^{-20}$). The functions are numbered as in Figure 1.

Next, we create a functional organization map for GO, as described in Section 2.4. We compare each of the FOMS of the three omics datasets (first three panels in Figure 4) with the one of GO by Pearson's correlation test. The functional organization of GI data correlates best with that of GO, with Pearson's correlation coefficient (PCC) of 0.717 (p -value $\approx 7.88 \times 10^{-18}$), followed by that of COEX, which has PCC of 0.614 (p -value $\approx 3.21 \times 10^{-12}$), and then followed by that of PPI data, which has PCC of 0.389 (p -value $\approx 4.02 \times 10^{-5}$) (bottom panel of Figure 4).

3.3 Complementarity of Biological Data

To test whether combining multiple data sources produces a functional map of the yeast's cell that is in higher agreement with GO than those offered by each data set individually, we propose an integration model that can easily scale to accommodate an arbitrary number of input data types. It works by solving the multiple linear regression, $GO \approx \alpha \times PPI + \beta \times COEX + \gamma \times GI$, and finding (α, β, γ) coefficients for the quadruplet of functional organization maps derived from GO, PPI, COEX and GI data. This approach is useful as it allows for easy inclusion or omission of data sets, enabling us to determine the exact contribution of individual data sources and their combinations to the integrated model.

We find that integrating the functional maps obtained from PPI, COEX and GI data results in the model that is in best functional agreement with GO, achieving the Pearson's correlation coefficient of 0.752 (and p -value $\approx 2.19 \times 10^{-20}$), a 5% increase over the highest-agreeing single data source (the FOM of this integrated model is on the right of Figure 4). That is, the best model is obtained by using all molecular data, showing that each dataset captures a complementary functional aspect of the cell. The integrated map is obtained as $9.3 \times PPI + 2.4 \times COEX + 30.5 \times GI$, which means that GI and PPI data contribute the most.

In the integrated map, the strongest functional link is between Cell polarity/morphogenesis and Protein folding glycosylation/cell wall, which are also strongly linked in the functional organization maps of GI and of PPI, but not in the one of COEX, and between Nuclear-cytoplasmic transport and Ribosome/translation, which are not linked in the FOMs of either of PPI, COEX and GI when considering them in isolation. This is an example of new functional relations emerging from integration of omics data. These associations are

biologically meaningful. For example, bi-directional traffic occurs between nucleus and cytosol through nuclear pore complexes; many proteins that function in the nucleus (e.g. histones, DNA and RNA polymerases and RNA-processing proteins) are made in the Ribosome (through translation of mRNAs) and are imported into the nucleus from the cytosol. At the same time, tRNAs and mRNAs are synthesized in the nuclear compartment and then exported to the cytosol [54].

3.4 Complementarity of the Omics Datasets

As highlighted above, the three molecular interaction datasets capture different, but complementary functional aspects of the cell. Here, we seek to better understand how the datasets differ. A straightforward measure of similarity between two interaction data sets is by measuring the number of genes and interactions they have in common in their network representation. Using our three interaction datasets represented by their corresponding networks, we find that PPI and GI share only 1.79 % of interactions, while COEX shares 2.01 % of interactions with PPI and 5.01 % of interactions with GI (Figure 5). The larger overlap between COEX and GI interactions may explain the larger similarity of their functional organization maps ($PCC \approx 0.63$, versus $PCC \approx 0.34$ between the FOMs of PPI and COEX, and $PCC \approx 0.48$ between the FOMs of PPI and GI). As only a portion of the interactions is shared between the data sets, this means that there is a unique informational value within each data set, which explains why combining the datasets achieves a better functional agreement with GO than any single data set alone (as also observed in previous studies [28], [29], [31], [35], [36] and in our results described in Section 3.3).

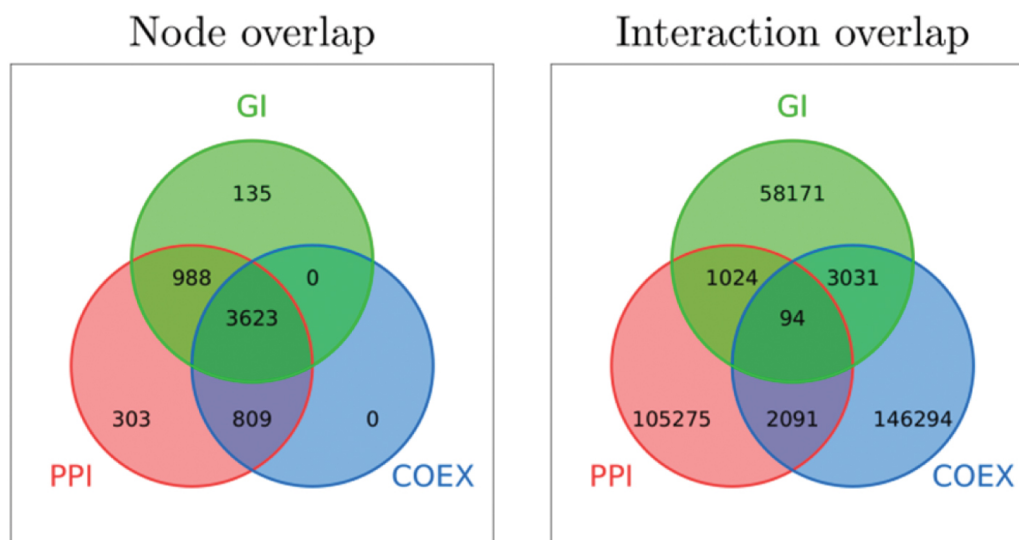


Figure 5: Overlaps between molecular interaction datasets. While molecular interaction datasets share most of their genes (nodes), most of the interactions (edges) are unique to each dataset.

4 Concluding Remarks

We provide a simple method to untangle the functional space of the yeast cell captured by the three omics datasets, protein-protein interactions, genetic interactions and co-expression data. We show that biological functions are differently organized in each of these omics datasets with varying agreement with the cell's functional organization of GO. By combining the functional organization maps of the three datasets into a unifying functional organization framework that we demonstrate is in higher agreement with GO than those of each of the datasets in isolation, we demonstrate complementarity of the functional information carried in the three omics datasets.

We use GO as the gold standard of the cell's functional organization, although GO is continuously being improved and re-evaluated. However, the methods that we introduce are generic and can be applied to any descriptors of the cell's functional organization (e.g. for studying how cancer related pathways communicates with each other in healthy and cancer cells). Also, while our integration framework that elucidates the cell's cross-functional communication lines produces functional organization maps that are in high agreement with the functional organization of GO (Figure 4), we are still in search of the model that can fully capture the functional organization of the cell and explain the lines of inter-functional cross-communication at a finer level of

functional granularity. Nevertheless, our results demonstrate that it is possible to tackle these biological challenges even with currently available noisy and incomplete omics data and even with very simple computational methods.

Acknowledgements

We thank Dr. Charles Boone (Donnelly Centre, Department of Molecular Genetics, University of Toronto, Canada), Prof. Chad L. Myers (Computational Biology and Functional Genomics Lab, University of Minnesota, USA) and Dr. Anastasia Baryshnikova (Lewis-Sigler Institute for Integrative Genomics, Princeton University, USA) for helpful suggestions and comments, and Dr. Vuk Janjic for his help on preliminary work.

This work was supported by the European Research Council (ERC) Starting Independent Researcher Grant 278212, the Serbian Ministry of Education and Science Project III44006, and the awards to establish the Farr Institute of Health Informatics Research, London, from the Medical Research Council, Arthritis Research UK, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Economic and Social Research Council, Engineering and Physical Sciences Research Council, National Institute for Health Research, National Institute for Social Care and Health Research, and Wellcome Trust (grant MR/K006584/1).

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

- [1] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Res.* 2014;42:D199–205.
- [2] Ryan CJ, Cimermančič P, Szpiech ZA, Sali A, Hernandez RD, Krogan NJ. High-resolution network biology: connecting sequence with function. *Nat Rev Genet.* 2013;14:865–79.
- [3] Botstein D, Fink G. Yeast: an experimental organism for modern biology. *Science.* 1988;240:1439–43.
- [4] Botstein D, Fink GR. Yeast: an experimental organism for 21st century biology. *Genetics.* 2011;189:695–704.
- [5] Goffeau A, Barrell BG, Bussey H, Davis RW. Life with 6000 genes. *Science.* 1996;274:546–67.
- [6] Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science.* 1999;285:901–6.
- [7] Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature.* 2002;418:387–91.
- [8] Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet.* 2013;14:333–46.
- [9] Kahn SD. On the future of genomic data. *Science.* 2011;331:728–9.
- [10] Gross M. Riding the wave of biological data. *Curr Biol.* 2011;21:R204–6.
- [11] Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science.* 2016;353:aaf1420.
- [12] Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, et al. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci.* 2000;97:1143–1147.
- [13] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. A comprehensive analysis of protein–protein interactions in *saccharomyces cerevisiae*. *Nature.* 2000;403:623–7.
- [14] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, et al. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature.* 2002;415:180–3.
- [15] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science.* 2008;320:1344–9.
- [16] Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science.* 2001;294:2364–8.
- [17] Mani R, Onge RP, Hartman JL, Giaever G, Roth FP. Defining genetic interaction. *Proc Natl Acad Sci.* 2008;105:3461–6.
- [18] Chua HN, Sung W-K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. *Bioinformatics.* 2006;22:1623–30.
- [19] Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. *Nat Biotechnol.* 2000;18:1257–61.
- [20] Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast.* 2001;18:523–31.
- [21] Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol.* 2003;21:697–700.

- [22] Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol*. 2007;3:88.
- [23] Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, et al. Genemania prediction server 2013 update. *Nucleic Acids Res*. 2013;41:W115–22.
- [24] Milenkovic T, Przulj N. Uncovering biological network function via graphlet degree signatures. *Cancer Inform*. 2008;6:257–73.
- [25] Gillis J, Pavlidis P. The impact of multifunctional genes on “guilt by association” analysis. *PLoS One*. 2011;e172586.
- [26] Christie KR, Hong EL, Cherry JM. Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns. *Trends Microbiol*. 2009;17:286–94.
- [27] Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TF, Stemple DL. The future of model organisms in human disease research. *Nat Rev Genet*. 2011;12:575–82.
- [28] Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The genetic landscape of a cell. *Science*. 2010;327:425–31.
- [29] Baryshnikova A, Costanzo M, Myers CL, Andrews B, Boone C. Genetic interaction networks: toward an understanding of heritability. *Annu Rev Genomics Hum Genet*. 2013;14:111–33.
- [30] Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat Biotechnol*. 2010;28:149–56.
- [31] Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. *Trends Plant Sci*. 2014;19:212–21.
- [32] Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. *Nat Methods*. 2012;9:796–804.
- [33] Van Landeghem S, De Bodd S, Drebert Z, Inze D, Van de Peer Y. The potential of text mining in data integration and network biology for plant research: a case study on *Arabidopsis*. *Plant Cell*. 2013;25:794–807.
- [34] Simeone A, Marsico G, Collinet C, Galvez T, Kalaidzidis Y, Zerial M, et al. Revealing molecular mechanisms by integrating high-dimensional functional screens with protein interaction data. *PLoS Comput Biol*. 2014;e100380110.
- [35] Zitnik M, Janjic V, Larminie C, Zupan B, Przulj N. Discovering disease-disease associations by fusing systems-level molecular data. *Sci Rep*. 2013;3:2023.
- [36] Gligorijevic V, Janjic V, Przulj N. Integration of molecular network data reconstructs Gene Ontology. *Bioinformatics*. 2014;30:i594–600.
- [37] Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res*. 2016;44:D536–41.
- [38] Bass JI, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJ. Using networks to measure similarity between genes: association index selection. *Nat Methods*. 2013;10:1169–76.
- [39] Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S, Narise T, et al. COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res*. 2015;43:D82–6.
- [40] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;25:25–29.
- [41] Cannistraci CV, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci Rep*. 2013;3:1613.
- [42] Cannistraci CV, Alanis-Lobato G, Ravasi T. Minimum curvilinearity to enhance topological prediction of protein interactions by network embedding. *Bioinformatics*. 2013;29:i199–209.
- [43] Lin D. An information-theoretic definition of similarity. *Proceedings of the fifteenth international conference on machine learning*, Burlington, MA; Morgan Kaufmann Publishers Inc, 1998;296–304.
- [44] Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res*. 1999;11:95–130.
- [45] Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, et al. A knowledge-based clustering algorithm driven by gene ontology. *J Biopharm Stat*. 2004;14:687–700.
- [46] Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, et al. Correlation between gene expression and go semantic similarity. *IEEE/ACM Trans Comput Biol Bioinf*. 2005;2:330–8.
- [47] Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*. 2009;e10004435.
- [48] Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief Bioinform*. 2012;13:569–85.
- [49] Kuchaiev O, Przulj N. Learning the structure of protein-protein interaction networks. *Pacific symposium on biocomputing Vol. 14* 2009:39–50.
- [50] Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006;27:861–74.
- [51] Sokal R, Michener C. A statistical method for evaluating systematic relationships. *Univ Kansas Sci Bull*. 1958;38:1409–38.
- [52] Ciechanover A. The ubiquitin–proteasome pathway: on protein death and cell life. *EMBO J*. 1998;17:7151–60.
- [53] McCann TS, Tansey WP. Functions of the proteasome on chromatin. *Biomolecules*. 2014;4:1026–44.
- [54] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *The transport of molecules between the nucleus and the cytosol*. New York, NY: Garland Science, 2002.