Minireview

The interaction map of yeast: *terra incognita?*

Joe Mellor and Charles DeLisi

Address: Program in Bioinformatics, 24 Cummington Street, Boston University, Boston, MA 02215, USA.

Correspondence: Joe Mellor. Email: mellor@bu.edu

Abstract

A systematic curation of the literature on *Saccharomyces cerevisiae* has yielded a comprehensive collection of experimentally observed interactions. This new resource augments current views of the topological structure of yeast's physical and genetic networks, but also reveals that existing studies cover only a fraction of the cell.

Biologists today find themselves in a situation not unlike that of 15th-century explorers. Roughly half a millennium ago, an era of exploration stemmed from a need for better information and more precise maps to facilitate new commerce. Novel technologies, including faster ships and improved navigation, facilitated exploration. The one-to-many communication made possible by the printing press accelerated the impact of these new discoveries, and our views of the planet and of ourselves were both revolutionized. In our own time, technology pushes biology towards equally revolutionary breakthroughs. The fundamental purpose - deeper understanding and improvement of life - remains the same now as then, although the details, methods and goals are of course vastly different. The sequencing of hundreds of genomes, the systematic measurements of genome activity, the large-scale assays of protein-protein and protein-DNA binding, and the use of computers to analyze information and facilitate many-to-many communication, collectively promise an unprecedented understanding of the workings of the cell, and a revolution in medicine.

The advent of high-throughput biology allows us for the first time in history to think concretely about a global representation of the cell. Unlike the cartographers of old, we are faced not merely with representing a static globe with fixed features; we must map a cellular universe with constantly interweaving themes, which alter as environments change. This enterprise is daunting, and so too is the less complex undertaking of specifying and representing the allowable interactions, which are selected by particular environments, without specifying the rules of selection. Data produced by current and yet-unforeseen technologies will eventually provide the interaction maps and the rules of environmental selection needed to fully understand the behavior of living cells. But at the moment, even the complexity of the problem remains unspecified. How many molecular connections make up a cell? How do these interactions combine to make functional cells, with a broad spectrum of phenotypes? A striking benefit of network mapping is not just what is revealed, but also what is not revealed and remains to be uncovered.

An important new paper by Reguly and Breitkreutz *et al.* [1] in *Journal of Biology* makes it clear that the landscape of even the best-studied eukaryote, the budding yeast *Saccharomyces cerevisiae*, remains significantly unexplored. The authors

used the extensive literature based on decades of research to curate a reference network of known interactions in yeast. This literature-curated collection corresponds to a network of some 33,000 high-confidence interactions between proteins or genes in yeast. Surprisingly, it shows little overlap with the published physical [2-6] and genetic [7] interaction networks reported in recent years by large-scale assays. Even with apparent similarities in topology or connectivity, only a fraction of the information in the curated network has been recovered by various high-throughput screening techniques such as systematic yeast two-hybrid analysis or synthetic genetic arrays (see Figure 1). Different views may exist on why this should be, for example in regard to levels and sources of false positives and false negatives in high-throughput datasets [8], but even the most optimistic assessment suggests that tens of thousands of interactions remain to be discovered in yeast. This in turn conveys the enormous scale of the problem of finding similar networks in higher organisms such as worm, mouse or human.

The curated network: a new benchmark

With an overlap of only 15% compared with previous high-throughput screening studies, the network of curated interactions reported by Reguly and Breitkreutz *et al.* [1] contains significant new information for use in the study of networks in yeast. Part of the curated information is in the form of a physical interaction network (LC-PI, 22,000 interactions)

Figure 1
Topological view of the curated protein-protein network of yeast interactions. Adapted from data in Reguly and Breitkreutz *et al.* [1]. Links are curated from thousands of literature articles referencing proteins in the *Saccharoymyces cerevisiae* genome. Links shown in black are interactions also recovered by any of five commonly used datasets derived from high-throughput yeast two-hybrid or mass spectrometric screening techniques. Visualization was performed with the VisANT analysis tool [19].

between proteins, as measured by various binding and affinity-based methods. Another network, of genetic interactions (LC-GI, approximately 11,000 interactions), consists of links between genes that manifest altered phenotypes, generally when a pair of genes is modified in tandem. Together, the literature-curated collection effectively doubles the amount of data now publicly available on interaction networks in yeast to some 50,000 nonredundant interactions. Whereas most previously available data has been delivered by large-scale and high-throughput assays such as comprehensive yeast two-hybrid screening (for protein-protein interactions) or synthetic genetic array (SGA) analysis and diploid-based synthetic lethality analysis on microarrays (dSLAM) (for genetic interactions) [7,9,10], the literature-curated network is almost entirely derived from smaller-scale experiments, with presumably higher average accuracy.

Each literature-curated interaction recorded by Reguly and Breitkreutz *et al.* [1] is associated with a publication, or publications, of origin, allowing more precise understanding of its experimental origins, or level of confidence, depending on the method or the number of confirming observations. The availability of this type of refined data, downloadable through the BioGRID [11] and *Saccharomyces* Genome Database (SGD) [12] projects, is a significant contribution to the network and systems biology community.

This is not the first project to curate interaction data; current projects such as the Biomolecular Interaction Network Database (BIND) [13], the Molecular Interaction Database (MINT) [14], the Munich Center for Information on Protein Sequences (MIPS) [15], the Database of Interacting Proteins and IntANT [16] and the Human Protein Reference Database (HPRD) [17] have already laid significant groundwork in creating resources of published interaction data. Reguly and Breitkreutz *et al.* [1] have gone further by expanding the coverage to all electronically available publications, representing nearly 10,000 research articles. This coverage is not exhaustive or saturating, but a useful framework is now in place for continued curation of similar data from the remaining literature. A large number of published articles pre-date electronic publication, and much would probably be gained by curating articles that are older, albeit harder to find.

At present, the most valuable application of this curated interaction data may be for benchmarking the quality and coverage of current and future high-throughput data. As more and more analyses of biological systems use information from large-scale experiments, the accuracy and coverage of these datasets will become more important as well. Computational analyses of the modular structure and function of systems encoded by various types of interactions clearly
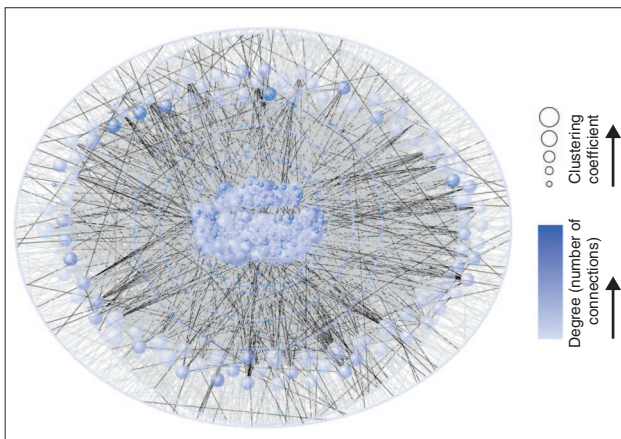
depend on the underlying quality of the data to hand. Reguly and Breitkreutz *et al.* [1] show that the higher-quality literature-curated interaction data can in fact provide more accurate predictions of the integrated network - for example in the prediction of protein complexes from physical interactions, or the Bayesian integration of multiple sources - than those obtained from high-throughput data alone. They also show that among the different methods of assessing interactions between genes and proteins, the literature-curated data appear to be best predictors of shared Gene Ontology (GO) function or pathway, transcriptional co-regulation, and tendency towards evolutionary conservation.

Comparisons of high-throughput versus literature-curated networks

Reguly, Breitkreutz and colleagues [1] also make comparisons of the function and structure of interaction networks obtained from the literature versus high-throughput screening. Here, some compelling results suggest that the information gathered from curation has subtle trends that are absent from high-throughput studies. First, certain GO functions [18] are enriched in the LC-PI and LC-GI networks compared with corresponding high-throughput datasets. This is probably due to the nature of small-scale studies, which often focus on particular cellular functions and systems of interest, compared with the 'dragnet' approach of many large-scale studies. A speculative consequence of this might be that large-scale studies are more likely to find 'new' information, because they effectively look at many more possibilities. Indeed, direct comparison of interaction enrichment in LC-PI versus high-throughput physical interaction (HTP-PI) datasets shows that while the high-throughput interactions are enriched for literature-curated interactions, the converse is apparently not true. This may be due to the known high rate of false positives in high-throughput datasets, especially in two-hybrid approaches, as mass spectrometric screens appear to perform better in this comparison.

Finally, the intrinsic biases in different methods may play a direct role in how interactions are reported. Reguly and Breitkreutz *et al.* [1] found that persistently cited genes were more connected on average in the new literature-curated network than in the high-throughput network. Thus, smaller-scale studies, in their focus on particular genes or proteins, are perhaps more efficient in finding new interactions for particular genes or proteins than large-scale studies. Fundamental differences in method explain how genetic interactions, as well, are often different when studied on large and small scales. Large-scale genetic screens such as SGA and dSLAM are effective where neither gene in a pair is essential, but more subtle growth effects can be

examined in small-scale studies even between conditional alleles of essential genes. More nuanced views of interactions gained by smaller-scale studies can potentially explain the increased overlap that Reguly and Breitkreutz *et al.* [1] observe among physical and genetic networks in literature-curated versus high-throughput data. In this sense, high-throughput data may be a decent 'first-pass' view of yeast's network structure, but as more types of interactions are included in a network, and its density increases, correlations between physical and genetic evidence become more apparent, and the full complexity of the network emerges.

In order to gain a clear picture of what is needed to fully map the networks that underlie biology, it will be important to establish the amount of interaction information needed to assemble accurate representations of these networks. Each mapping endeavor contributes to a larger understanding of the puzzle, and the new work of Reguly and Breitkreutz *et al.* [1] represents a useful benchmark by which to judge these mapping endeavors. A recent, rapid expansion in our knowledge of cellular interaction networks has been largely due to the development of large-scale techniques in molecular biology, not only the experimental technology needed to assess interaction data but also the computational innovations needed to filter it and infer function. The curation effort of Reguly and Breitkreutz *et al.* shows that the inference problem is far from saturated, and that significant numbers, and types, of interactions in the cell are unexplored.

References

1. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskya OG, Ideker T, Dolinski K, Batada NN, Tyers M: **Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*.** *J Biol* 2006, **5:**11.
2. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415:**180-183.
4. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
5. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: **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 USA* 2000, **97:**1143-1147.
6. Uetz P, Hughes RE: **Systematic and large-scale two-hybrid screens.** *Curr Opin Microbiol* 2000, **3:**303-308.
7. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, *et al.*: **Systematic genetic analysis with ordered arrays of yeast deletion mutants.** *Science* 2001, **294:**2364-2368.

8.   Bader GD, Hogue CW: **Analyzing yeast protein-protein interaction data obtained from different sources.** *Nat Biotechnol* 2002, **20:**991-997.

9.   Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD: **A robust toolkit for functional profiling of the yeast genome.** *Mol Cell* 2004, **16:**487-496.

10. Tong AH, Boone C: **Synthetic genetic array analysis in** *Saccharomyces cerevisiae.* *Methods Mol Biol* 2006, **313:**171-192.

11. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34(Database issue):**D535-D539.

12. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE, *et al.*: *Saccharomyces* **Genome (SGD) provides tools to identify and analyze sequences from** *Saccharomyces cerevisiae* **and related sequences from other organisms.** *Nucleic Acids Res* 2004, **32(Database issue):**D311-D314.

13. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31:**248-250.

14. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513:**135-140.

15. Guldener U, Munsterkotter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stumpflen V: **MPact: the MIPS protein interaction resource on yeast.** *Nucleic Acids Res* 2006, **34(Database issue):**D436-D441.

16. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, *et al.*: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2004, **32(Database issue):**D452-D455.

17. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, *et al.*: **Human protein reference database - 2006 update.** *Nucleic Acids Res* 2006, **34(Database issue):**D411-D414.

18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.

19. Hu Z, Mellor J, Wu J, Yamada T, Holloway D, Delisi C: **VisANT: data-integrating visual framework for biological networks and modules.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W352-W357.