Assessing the role of multi-protein complexes in determining phenotype

Le Meur, Nolwenn (1st author)
EHESP, InfoBiostat Department
Avenue du Professeur Léon-Bernard
CS 74312 3504 Rennes Cedex, France
E-mail: Nolwenn.LeMeur-Rouillard@ehesp.fr

Gentleman, Robert (2nd author)
Genentech, Bioinformatics and Computational Biology Department
1 DNA Way
South San Francisco, CA 94080, USA
E-mail: gentleman.robert@gene.com

INTRODUCTION

Understanding regulatory mechanisms and sensitivity of cellular organizational units in complex biological systems is an important challenge. In medicine, in particular, it will lead to greater understanding of the processes involved in some diseases. In that context, we have demonstrated the importance of multi-protein complexes in synthetic lethality and characterized some of the biological mechanisms involved [1]. Other studies also suggest that some control of phenotype can be usefully attributed to multi-protein complexes rather than genes or pathways [2–6] and hence may help provide elucidation of the underlying roles or mechanisms that directly control changes in phenotype. In the long term, in the case of disease phenotype, knowledge of organizational units involved in the disease regulator mechanisms will enable us to identify biological targets for drug therapy and improve the specificity and efficacy of those drugs.

The challenge of understanding cellular regulatory mechanisms by cellular organizational units is difficult due to the size of the underlying biological network and the heterogeneous nature of the control mechanisms involved [7; 8]. Indeed, many genes are pleiotropic and their product play many roles in the cell. It may then not be clear which of those different functions is directly related to the change in phenotype [4; 9]. Moreover, epistasis can mask the phenotypic effect of a gene, obscuring the relationship between gene and phenotype [8]. Tools are therefore needed to identify which function of a gene relates to a disease phenotype. More generally, systems biology approaches are now required to understand the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of that system.

In this paper, we promote the concept that while phenotypic changes are often measured by the manipulation of single genes, such as gene deletion or up-regulation, interpreting the biological mechanisms that underly the change in phenotype will often depend on higher levels of organization, such as multi-protein complexes. We propose computational methods and present the use of R packages [10; 11] to disentangle the multi-protein complexe contribution to disease phenotype in *Saccharomyces cerevisiae.*

METHODS

Data sources

Multi-protein complex co-membership was determined from GO [12; 13], MIPS [14], protein-protein interactions data obtained from the IntAct database [15], and estimates from tandem affinity purification-mass spectrometry experiments (AP-MS) [2; 16–19]. This resulted in an estimated interactome of 1,803 unique genes and 947 multi-protein complexes: 398 curated multi-protein complexes from the online databases (GO, MIPS, IntAct) and 549 estimated and non-annotated multi-protein complexes from the AP-MS experiments. The multi-protein complexes estimated from the AP-MS experiment have a prefix *apComplex* followed by the author and year of the experiment and an arbitrary identification number [20].

We define as conditionally essential genes all genes that show fitness defect under some growth condition, that includes essential and haploinsufficient genes which deletion in rich media induces lethality. The list of *S. cerevisiae* rich media essential genes was obtained from the *Saccharomyces* Genome Database [21]. Among the 4,918 verified open reading frames (ORFs) believed to compose *S. cerevisiae* genome (source: `www.yeastgenome.org` - last updated April 2011) 1,101 are classified as essential genes [22]. One can access this dataset from the *SLGI* R package, available from the Bioconductor Project [11]. The list of haploinsufficient genes was extracted from Deutschbauer *et al.* [3] who found that 184 *S. cerevisiae* genes were haploinsufficient for growth in Yeast extract/Peptone/Dextrose (YPD). The fitness defect dataset was extracted from Dudley *et al.* [4] who created a collection of gene-deletion mutants to determine genes that contribute to a particular phenotype under 21 different experimental stress conditions. Both datasets, the haploinsufficient and fitness defect data, are included in the *PCpheno* R package, available from the Bioconductor Project [11].

Computational and Statistical Methods

Our null hypothesis was that there was no association between a collection of genes that induced a phenotypic change and some cellular organizational units (*e.g.*, multi-protein complexes, pathways). To test this hypothesis we considered a multi-faceted approach. First, we used a hypothesis test designed to determine whether there was an effect that could be attributed to that specific grouping of genes, without testing which cellular organizational units were involved. Then, if we rejected our null hypothesis of no association between a collection of genes that induced a phenotypic change and some cellular organizational units, the next step was to identify those specific organizational units. We thus looked for the cellular organizational units that had an over-representation of the genes that induced the phenotypic change (*e.g.*, conditionally essential genes).

Density Estimation For each cellular organizational unit, we computed the proportion of genes that affect the phenotype. We then computed the smoothed histogram of the proportions and compared it to a reference distribution. Our reference distribution was obtained by randomly permuting 1,000 times the gene labels for the interactome and computing together, for each permutation, the new (simulated) proportion of genes that affect the phenotype and the associated smoothed histograms.

Graph Theory The graph theory procedure is based on the permutation of graphs as proposed by Balasubramanian *et al.* [23]. Two distinct graphs, $G_i = (V, E_i)$ with $i = 1, 2$, were formed. The nodes, V, were the *S. cerevisiae* genes and they were common to both graphs. In one graph G_1 two proteins had an edge between them if, and only if, they were co-members of one, or more, cellular organizational units. In the second graph G_2 edges were created between all proteins that were associated with a phenotype of interest, so that if there were k genes associated with the phenotype of interest then we had a complete graph with $k(k-1)/2$ edges. We excluded self-loops in both graphs. We then computed the intersection of these two graphs and counted the edges in common. To test whether the number of edges in the resulting graph was unexpectedly large, a permutation analysis was performed. A reference distribution was obtained by permuting 1,000 times the labels on either G_1 or G_2 and counting the number of edges in common. A p-value was computed by comparing the observed counts to the estimated distributions of intersecting edges issued from the permutations.

Hypergeometric Test We used a hypergeometric test to assess whether a cellular organizational unit contains more genes that affect the phenotype than expected by chance. The hypergeometric test is the equivalent of Fisher's exact test for two-by-two tables. We adjusted the p-values for multiple comparisons by controlling the family wise error rate using the FDR method. We report both adjusted and raw p-values as it is not clear that the FDR method is the most appropriate adjustment for this analysis. Indeed, some work remains to be done to properly account for the fact that most genes are members of more than one multi-protein complex, hence there is a very complex dependency between the tests. We term the cellular organizational unit for which we reject this test (p-value ≤ 0.01) as *conditionally essential* since tightly related to the phenotype being studied.

Software Implementation and Availability The data used in the statistical analysis in this paper and the algorithms developed for the proposed computational methods are all available in the freely distributed open source R/Bioconductor packages [10; 11]. It is integrated into the R/Bioconductor environment for statistical computing and bioinformatics and run on multiple operating systems including Windows, Mac OS X and Unix.

Supplementary material is available at `http://nlmr.free.fr` under Science/Papers.

RESULTS AND DISCUSSION

Some phenotype can be attributed to multi-protein complexes In *S. cerevisiae*, out of the 4,918 verified ORFs believed to compose its genome, approximately 1,000 ORFs are said to be essential in rich media environment [22] and 184 are said to be haploinsufficient [3]. In addition among the numerous experiments to assess the fitness defect of gene deletion in various stressful environment, Dudley *et al.*[4] identified more than 800 genes that are sensitive in at least one stressful conditions (out of 21). In order to investigate whether those phenotypic changes can be usefully attributed to multi-protein complexes, we tested the null hypothesis that there is no association between a collection of genes that induce a phenotypic change and multi-protein complexes. If no relationship exists between the observed phenotype and multi-protein complex membership, we expect to see, for a given multi-protein complex, the proportion of genes associated with the observed phenotype close to the population proportion, *e.g.*, approximately one sixth for essential genes. If instead an association does exist, we expect that there will be some multi-protein complexes that have a large proportion of proteins associated to the observed phenotype and some with a small proportion, but fewer than expected (in the statistical sense) with moderate proportions. We tested our hypothesis using two omnibus tests and our current estimate of the *S. cerevisiae* multi-protein complex interactome [20]. The first test is based on density estimation [24]. The second approach is based on the permutation of graphs proposed by Balasubramanian *et al.* [23] (see *Materials and Methods* for details). Figure 1 shows the results of both approaches for the essential genes.

**** Figure 1 *****

Figure 1 Panel (a) represents the outputs of the density estimation approach. This plot provides a heuristic tool for assessing whether the observed density (dark line) is similar to those generated under the null hypothesis of no association between the genes inducing a phenotype and the multi-protein complexes used (gray lines). The observed data show that there is not only an over-abundance of multi-protein complexes with values near 0 but also an over-abundance of multi-protein complexes with proportions of essential genes near 1. The curves representing the smoothed histograms for the permuted data are clearly very different from the observed data, with larger values near the center (proportions between 0.4 and 0.6) and lower values near 0 and 1. We note that while the observed proportions must be between zero and one, this constraint is not imposed on the smoothed histograms, and they do extend beyond 0 and 1. This is not particularly problematic as all estimates (both the observed and the permutations) are subject to the same procedure. We could use ordinary histograms, but they simply could not be plotted one on top of the other, so we could not easily visualize the difference between the observed and permutation data (also for visualization purposes only 100 out of the 1,000 permutations performed are shown in Fig. 1). We also remark that all curves show a number of peaks. These arise due to the discrete nature of the multi-protein complexes. There are many complexes composed only of 2, 3 and 4 proteins. For these the observed proportions are similarly limited (*e.g.,* a cluster of size 3 can have proportions 0, 1/3, 2/3 or 1). Figure 1 Panel (b) presents the results of the graph theory approach. The histogram represents the distribution of the number of edges observed using the permutation model, under the null hypothesis, and the red line indicates the number of edges in the observed data. The observed number of edges is far larger than any value from the permutations and hence the permutational p-value is less than 1 in 1,000. Theses results provide strong evidence against the null hypothesis and indicate that some association exists between the genes related to those

particular phenotypes and the multi-protein complexes used in our analysis.

For the haploinsufficient genes and stress conditions the effect appears to be much less substantial (Fig. S1 for haploinsufficient dataset). One possible explanation is that this is due to the small number of those genes represented in our interactome (Tab. S1). For the haploinsufficient genes for instance, the outputs are different by an order of magnitude compare to the essential genes; about 30% of all genes in yeast are essential under the conditions tested while only 3% are haploinsufficient and only 152 out of the 183 haploinsufficient genes are represented in our interactome. However the results of the graph theory approach provide evidence against the null hypothesis for most conditions and indicate that some association exists between the genes associated with those particular phenotypes and the multi-protein complexes used in our analysis.

Multi-protein complexes contributing to phenotype Since the overall tests provided strong evidence against the null hypothesis and demonstrated that conditional essentiality can be usefully attributed to multi-protein complexes, we looked for the multi-protein complexes that have an over-representation of the genes inducing those phenotypic changes. To this aim we propose to apply a hypergeometric test approach with false discovery rate (FDR) adjustment and a p-value threshold of ≤ 0.01 (see *Materials and Methods*). Using the hypergeometric test approach, we identified several complexes composed of a significant number of conditionally essential genes (Tab. 1 and 2). In an attempt to circumvent multiple testing correction, we also tested the use of a model-based clustering algorithm proposed by the R package *mclust* [25]. We defined *a priori* 2 groups: conditionally essential complexes and non-essential complexes. The clustering approach then allowed computing the probabilities (along with a measure of certainty) that a multi-protein complex belongs to the different groups. The results were in accordance to our current approach (data not shown) however the discrete nature of our data and the $k = 2$ clustering imposed by our hypothesis complicate their interpretation.

For essentiality (Tab. 1) and haploinsufficiency phenotypes(Tab. S2), the annotated complexes are mostly involved in the replication and transcription machineries (*e.g.*, the *pre-replication* and *replication complexes*, the *small nucleolar ribonucleoprotein complexes*). This result is not so surprising as DNA replication and protein transcription are very critical processes. Additionally, the molecules and mechanisms that ensure a faithful DNA replication and protein transcription have been highly conserved throughout evolution [26].

***TABLE1 ***

In the stress condition experiments [4], the results should be taken with caution due to the small number of tested genes represented in our interactome (Tab. S1). Nevertheless in some experiments, the observed phenotypes seem to be related to fewer multi-protein complexes (Tab. 2). It is not that surprising as most media used by Dudley *et al.* [4] are drugs that should have specific targets. For instance the cycloheximide, an inhibitor of protein biosynthesis, act on chromatin remodeling complexes (*GO:0000508*) and transcription co-activator complexes (*GO:0016593,* *MIPS-230.20.10*). It is also interesting that some of the multi-protein complexes involved in phenotypic changes induced by the anti-fungal drug Paraquat are similar to the one found for the anti-fungal drug Nystatin tested by Giaever *et al.* (2002) (Tab. S5). In addition we note that similar multi-protein complexes relate to different conditions. For instance, the H+-transporting ATPase, vacuolar appears especially critical

*** TABLE2 ***

While analyzing the list of multi-protein complexes related to phenotype, we note some redundancy in term of multi-protein complex definition. Indeed some caution is needed as some multi-protein complexes overlap substantially (within and between databases) and have similar descriptions. As an example, the *MIPS-510.120* complex (*RNA polymerase III*) is entirely contained in the *GO:0005666* complex (*DNA-directed RNA polymerase III complex*). In fact, it is well known that multi-protein complexes can have several functional isoforms but it is virtually impossible to distinguish them *via* AP-MS or pull-down technologies if all variants are present [16]. It is also difficult to accurately represent this behavior in the data structures used to model these data. Furthermore, Lichentenberg *et al.* [27] have shown that many cell-cycle related complexes use a 'just-in-time' assembly mechanism before being active. Therefore, some complex definitions do not correspond directly to functional complexes as their different functional isoforms are not necessarily well separated in the time and space. High throughput interaction protein experiments are also not error free [28]: difficulties to identifying and annotating complexes, technological problem detecting small complexes, *etc.*, all lead to errors. Nevertheless those experimental limitations and data representation issues are likely to be overcome with the improvement or development of technologies. And our methods will be directly applicable to such improved predictions when they become available.

Finally, as suggested by other [9; 29], we also considered the role of pathways in explaining phenotypic changes at the system level. We thus used the well-known KEGG database [30] to test whether for each KEGG pathway we observed a higher proportion of genes associated with the observed phenotypes than expected by chance. The smoothed density approach indicated substantial discrepancies (Fig. S4). Many more pathways than expected by chance have no genes associated with the observed phenotypes, suggesting that the null hypothesis is not tenable for either the haploinsufficient genes or the essential genes. In fact, out of the 99 known KEGG pathways, 28 are known to have no essential gene and 83 have no haploinsufficient gene.

CONCLUSION

In this paper we have confirmed the hypothesis that some phenotypes can usefully be attributed to multi-protein complexes. Our results support and supplement the observations by Yu *et al.* [9] that protein-protein interactions are condition-specific and relate to the pleiotropic properties of genes. We showed that genes that did not exhibit essential phenotype under rich medium condition as used by Giaever *et al.* [22] could be critical under other conditions, *i.e.* conditionally essential. As a proof of principle, we used the essential genes characterized by Giaever *et al.* [22], the haploinsufficient genes identified by Deutschbauer *et al.* [3] and the fitness genes characterized by Dudley *et al* [4] under stressful condition. However the method applies to virtually any system where phenotypic outputs are measured for single gene perturbations defined *a priori*. For instance one could make use of phenotypic datasets from the YeastMiner database (`http://yeastmine.yeastgenome.org`) which to our knowledge is currently the richest source of phenotypic data. [31]. In addition, to assess these relationships, we provide open source computational and statistical tools as R packages available on the Bioconductor website [10; 11].

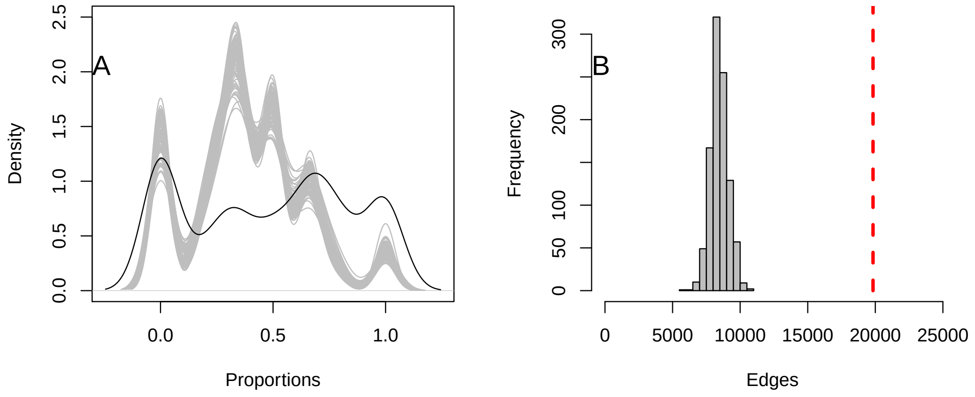

Figure 1: Essential genes are not randomly distributed among multi-protein complexes. Panel A. Smoothed histograms of the proportion of genes per multi-protein complexes that are associated with a phenotype. The dark line represents the observed data and the light curves represent the permuted data. Only the first 100 simulated density estimates out of 1,000 permutations are displayed for visualization efficiency. Panel B. Distribution of the number of edges, under the null distribution (1,000 permutations) of genes randomly distributed in multi-protein complexes (grey histogram)compared to the number of observed edges, dashed line.

	Observed	Expected	Size	Odds	P-value (adj)	P-value	Description
GO:0005732	42	21.59	56	5.03	1.87e-05	1.98e-08	small nucleolar ribo...
GO:0005666	17	6.55	17	Inf	3.84e-05	8.11e-08	DNA-directed RNA pol...
MIPS-410.30	16	6.17	16	Inf	6.74e-05	2.13e-07	Pre-replication comp...
apCompGavin2002: 228	18	7.32	19	29.43	8.87e-05	3.75e-07	-
GO:0005656	15	5.78	15	Inf	1.06e-04	5.61e-07	pre-replicative comp...
MIPS-360	28	13.88	36	5.77	2.37e-04	1.50e-06	Proteasome
MIPS-410.35	18	7.71	20	14.70	2.84e-04	2.40e-06	Replication complex
apCompGavin2002: 231	18	7.71	20	14.70	2.84e-04	2.40e-06	-
MIPS-510.120	13	5.01	13	Inf	4.07e-04	3.87e-06	RNA polymerase III
apCompGavin2002: 224	14	5.78	15	22.76	1.35e-03	1.43e-05	-
GO:0046540	22	10.79	28	6.00	1.40e-03	1.63e-05	U4/U6 x U5 tri-snRNP...
apCompGavin2002: 203	11	4.24	11	Inf	2.10e-03	2.66e-05	-
apCompGavin2002: 50	19	9.25	24	6.20	3.72e-03	5.36e-05	-
apCompGavin2002: 12	16	7.32	19	8.68	3.72e-03	5.50e-05	-
GO:0000172	10	3.85	10	Inf	4.39e-03	6.96e-05	ribonuclease MRP com...
GO:0005847	13	5.78	15	10.54	7.83e-03	1.70e-04	mRNA cleavage and po...
GO:0005669	13	5.78	15	10.54	7.83e-03	1.70e-04	transcription factor...
MIPS-360.10.10	13	5.78	15	10.54	7.83e-03	1.70e-04	20S proteasome
apCompGavin2002: 43	13	5.78	15	10.54	7.83e-03	1.70e-04	-
GO:0005849	9	3.47	9	Inf	7.83e-03	1.82e-04	mRNA cleavage factor...
GO:0005655	9	3.47	9	Inf	7.83e-03	1.82e-04	nucleolar ribonuclea...
apCompGavin2002: 205	9	3.47	9	Inf	7.83e-03	1.82e-04	-
GO:0005681	22	11.95	31	3.99	9.41e-03	2.29e-04	spliceosomal complex

Table 1: Multi-protein complexes associated with Essentiality (P-value<0.01). Observed: number of essential genes in the complex; Expected: expected number of essential genes in the complex; Size: total number of genes in the complex; Odds: odds ratios; P-value (adj): adjusted P-value of the Hypergeometric test (bonferroni correction); P-value: P-value of the Hypergeometric test; Description: annotation of for the given protein complex. Note that when the multi-protein complex is entirely composed of essential genes (Observed = Size) the odds ratio are infinite (Inf).

	Dudley et al (2005)	Interactome	p.value	nb.C 0.01	nb.C 0.05
cyclohex	164	79	0	0	6
FeLim	35	17	0	3	3
MPA	11	6	0.001	0	2
Paraq	36	22	0.001	3	5
YPGal	41	20	0.002	0	1
YPRaff	31	16	0.002	2	4
HU	87	52	0.003	0	5
CaCl2	180	88	0.007	2	7
YPGly	206	76	0.008	3	4
UV	32	22	0.009	1	1
EtOH	75	51	0.012	-	-
YPLac	159	52	0.014	-	-
CAD	83	45	0.027	-	-
lowPO4	34	10	0.037	-	-
pH3	16	8	0.052	-	-
rap	119	51	0.071	-	-
HygroB	264	109	0.145	-	-
Caff	208	105	0.192	-	-
NaCl	57	29	0.244	-	-
benomyl	34	19	0.594	-	-
DTT	5	-	-	-	-
Sorb	8	-	-	-	-

Table 2: Dudley et al. (2005) environmental stress conditions. Each row corresponds an environmental stress condition. The first column indicates the number of mutants with growth defect in Dudley's experiment. The second column indicates the number of those deleted genes in the interactome. The third column presents the p-value obtained by the graph theory test. A p-value <= 0.01 indicates that those deleted genes are not randomly distributed in the multi-protein complexes of the interactome. The fourth and fifth columns indicate the number of multi-protein complexes involved at a FDR adjusted pvalue <= 0.01 and 0.05. The 22 environmental conditions listed are: benomyl: 15ug/ml benomyl,microtubule function; CaCl2: 0.7M calcium chloride, divalent cation; CAD: 55uM Cadmium, heavy metal; Caff: 2mg/ml Caffeine; cyclohex: 0.18ug/ml cycloheximide, protein synthesis; DTT: unknown; EtOH YPD + 6% Ethanol; FeLim: irion limited,nutrient limited condition; HU: 11.4mg/ml Hudroxyurea, DNA replication and repair; HygroB: 50ug/ml hygromycin B, aminoglycosides; lowPO4: low phosphate, nutrient limited condition; MPA: 20ug/ml mycophenolic acid, transcriptional elongation; NaCl: 1.2M sodium chloride, general stress condition; Paraq: 1mM paraquat, oxidative stress; pH3: low pH, general stress condition; rap: 0.1ug/ml rapamycin, protein synthesis; Sorb: 1.2M sorbitol, general stress condition; UV: 100J/m2 ultra-violet, DNA replication and repair; YPGal 2% galactose, carbon source; YPGly 3% glycerol, carbon source; YPLac 2% lactate, carbon source; YPRaff 2% raffinose, carbon source.

ACKNOWLEDGEMENT

We thank Denise Scholtens and Avner Bar-Hen for helpful discussions on the statistical and graph theoretic aspects of this problem. We would also like to thank Muneesh Tewari and Dan Gottschling for providing assistance in our understanding of different aspects of this investigation.

REFERENCES (RÉFERENCES)

[1] N. Le Meur and R. Gentleman. Modeling synthetic lethality. *Genome Biology*, 9(9):R135, 2008.

[2] AC. Gavin, M. Boesche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, JM. Rick, AM. Michon, CM. Cruciat, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[3] AM. Deutschbauer, DF. Jaramillo, M. Proctor, et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics*, 169:1915–1925, 2005.

[4] AM. Dudley, DM. Janse, A. Tanay, R. Shamir, and G. McDonald Church. A global view of pleiotropy and phenotypically derived gene function in yeast. *Molecular Systems Biology*, 1:E1–E11, 2005.

[5] V. Spirin, MS. Gelfand, AA. Mironov, and LA. Mirny. A metabolic network in the evolutionary context: multiscale structure and modularity. *PNAS*, 23:8774–8779, 2006.

[6] Magali Michaut, Anastasia Baryshnikova, Michael Costanzo, Chad L Myers, Brenda J Andrews, Charles Boone, and Gary D Bader. Protein complexes are central in the yeast genetic landscape. *PLoS Computational Biology*, 7(2):e1001092, February 2011. PMID: 21390331.

[7] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101(9):2981–2986, 2004.

[8] M. Oti and HG Brunner. The modular nature of genetic diseases. *Clin Genet*, 71(1):1–11, 2007.

[9] Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, Hao T, Rual JF, Dricot A, Vazquez A, Murray RR, Simon C, Tardivo L, Tam S, Svrzikapa N, Fan C, de Smet AS, Motyl A, Hudson ME, Park J, Xin X, Cusick ME, Moore T, Boone C, Snyder M, Roth FP, BarabÃąsi AL, Tavernier J, Hill DE, and Vidal M. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10, 10 2008.

[10] CRAN. The comprehensive r archive network. `http://www.R-project.org`.

[11] Bioconductor. Open source software for bioinformatics. `http://www.bioconductor.org/`.

[12] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[13] E. Camon, M. Magrane, D. Barrell, et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 32(Database issue):D262–D266, Jan 2004.

[14] U. Guldener, M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, SJ Wodak, J. Garcia-Martinez, JE Perez-Ortin, et al. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research*, 33(Database Issue):D364, 2005.

[15] EBI. Intact database. `http://www.ebi.ac.uk/intact/site`.

[16] A-C. Gavin, P. Aloy, P. Grandi, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, Jan 2006.

[17] NJ. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, AP. Tikuisis, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440:637–643, 2006.

[18] NJ. Krogan, WT. Peng, G. Cagney, MD. Robinson, R. Haw, G. Zhong, X. Guo, X. Zhang, V. Canadien, DP. Richards, et al. High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell*, 13:225–239, 2004.

[19] Y. Ho, A. Gruhler, A. Heilbut, GD. Bader, L. Moore, SL. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, 2002.

[20] D. Scholtens and R. Gentleman. Making Sense of High throughput Protein-Protein Interaction Data. *Statistical Applications in Genetics and Molecular Biology*, 3(1):39, 2004.

[21] R. Balakrishnan, K. R. Christie, M. C. Costanzo, et al. Saccharomyces genome database. 2006.

[22] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, et al. Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391, Jul 2002.

[23] R. Balasubramanian, T. LaFramboise, D. Scholtens, and R. Gentleman. A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics*, 20(18):3353–3362, Dec 2004.

[24] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[25] Chris Fraley and Adrian Raftery. Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, 18(6):1–13, 1 2007.

[26] Park K and Kim D. Localized network centrality and essentiality in the yeast-protein interaction network. *Proteomics*, 9(22):5143–54, 11 2009.

[27] U. de Lichtenberg, L.J. Jensen, S. Brunak, and P. Bork. Dynamic Complex Formation During the Yeast Cell Cycle. *Science*, 307(5710):724–727, 2005.

[28] T Chiang, D Scholtens, D Sarkar, R Gentleman, and W Huber. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biology*, 8(9):R186, sep 2007.

[29] R. Kelley and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, 23(5):561–566, May 2005.

[30] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.

[31] Stacia R Engel, Rama Balakrishnan, Gail Binkley, Karen R Christie, Maria C Costanzo, Selina S Dwight, Dianna G Fisk, Jodi E Hirschman, Benjamin C Hitz, Eurie L Hong, Cynthia J Krieger, Michael S Livstone, Stuart R Miyasato, Robert Nash, Rose Oughtred, Julie Park, Marek S Skrzypek, Shuai Weng, Edith D Wong, Kara Dolinski, David Botstein, and J Michael Cherry. Saccharomyces genome database provides mutant phenotype data. *Nucleic Acids Research*, 38(Database issue):D433–436, January 2010. PMID: 19906697.

RÉSUMÉ (ABSTRACT) — optional

In Saccharomyces cerevisiae, we and others showed that molecular interactions within and between multi-protein complexes are critical for cell fate. Recent studies also suggest that some control of phenotype can be usefully attributed to multi-protein complexes rather than genes or pathways. Indeed, while phenotypic changes are often measured by the manipulation of single genes (deletion, up-regulation, etc.), the biological mechanisms that underly the change in phenotype might depend on higher levels of organization, such as multi-protein complexes. In this work we thus attempted to assess the role of multi-protein complexes in determining phenotype. We tested whether gene products known to be associated with a phenotype are randomly distributed in the interactome or cluster in specific multi-protein complexes. In addition, since the expression of phenotype highly depends on the environmental conditions, we investigated different datasets to evaluate if similar phenomena (random distribution or cluster) could be observed and thus associate multi-protein complex activity (fitness) to environmental conditions.