ComPlexUs
ORIGINAL RESEARCH PAPER
ComPlexUs 2003;1:19–28; DOI:10.1159/000067640
Received: May 7, 2002
Accepted after revision: August 5, 2002

Prediction of Protein Essentiality Based on Genomic Data

Hawoong Jeong[a,c] Zoltán N. Oltvai[b]
Albert-László Barabási[a]

[a] Department of Physics, University of Notre Dame, Notre Dame, Ind.;
[b] Department of Pathology, Northwestern University Medical School, Chicago, Ill., USA;
[c] Department of Physics, Korea Advanced Institute of Science and Technology, Taejon, Korea

Key Words
Scale-free networks · Lethality · Microarray · Two-hybrid protein interaction · Drug target identification

Abstract
A major goal of pharmaceutical bioinformatics is to develop computational tools for systematic in silico molecular target identification. Here we demonstrate that in the yeast *Saccharomyces cerevisiae* the phenotypic effect of single gene deletions simultaneously correlates with fluctuations in mRNA expression profiles, the functional categorization of the gene products, and their connectivity in the yeast's protein-protein interaction network. Building on these quantitative correlations, we developed a computational method for predicting the phenotypic effect of a given gene's functional disabling or removal. Our subsequent analyses were in good agreement with the results of systematic gene deletion experiments, allowing us to predict the deletion phenotype of a number of untested yeast genes. The results underscore the utility of large genomic databases for in silico systematic drug target identification in the postgenomic era.

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com
KARGER

Accessible online at:
www.karger.com/cpu

Prof. Albert-László Barabási
Department of Physics, University of Notre Dame
213 Nieuwland Science, Notre Dame, IN 46556 (USA)
Tel. +1 219 631 5767, Fax +1 219 631 5952
E-Mail alb@nd.edu

Synopsis

One of the central practical aims of molecular biology lies in the discovery of new drugs to treat human diseases. Hence, it is natural to ask how the flood of new information produced by the genomics and proteomics 'revolutions' might be put to use in this regard. Consider antibiotics, for example. The genomes of numerous bacterial pathogens have been fully sequenced, and researchers possess vast databases concerning the properties of these organisms' proteins. This information should enable researchers to develop new and more effective antibiotics to counter bacterial strains that have evolved resistance to traditional compounds. And yet the data can be overwhelming. A central difficulty in drug development – not only for antibiotics but for new drugs to treat cancers, fungal infections and other disorders – lies in devising methods that are capable of drawing meaningful conclusions from masses of often confusing data.

On this issue, however, Jeong et al. offer some hope – at least on the project of identifying genes and proteins as potentially promising drug targets. In the context of the yeast *Saccharomyces cerevisiae*, they show how to integrate three distinct types of genomic and proteomic data so as to predict the 'essentiality' of a gene, i.e. to estimate how crucial it is to the organism's viability. Clearly, the more essential a gene (or its associated protein) is to a pathogen or to a cancerous cell, the more attractive it is as a drug target; hence, this algorithm illustrates one way in which emerging data may be put to use to identify potential new drug targets on a 'rational' basis. (Finding compounds to act against such targets is, of course, another problem altogether.)

How can one estimate the essentiality of a gene? Jeong et al. begin by considering the general functional character of a gene's protein product. Traditionally, a protein's function has been associated with its specific molecular function – as a catalyst for a certain reaction, perhaps, or as a structural component of the cell. But a more recent perspective views proteins as elements in biochemical networks that act as functional modules within the cell. One module might be a collection of proteins involved in pro-

Introduction

The development of novel antibiotics and antifungal drugs is emerging as a critical issue for global healthcare due to the rapid appearance of bacterial and fungal pathogens possessing natural or acquired resistance against available therapeutic agents [1]. Similarly, the toxic side effects and the limited spectrum of activity of conventional chemotherapeutics necessitate the development of new, rationally identified cancer therapeutics [2]. The sequential waves of emerging technologies in genomics, proteomics, and drug design hold the promise of new avenues for the identification of new medicines [3]. However, it is yet unclear if the currently available genomic databases, coupled with newly developed computational algorithms, can offer sufficient information for automated in silico drug target identification. Here we demonstrate that the available massively parallel, systematic and complementary datasets for the yeast *Saccharomyces cerevisiae* offer unprecedented potential for the development of quantitative predictions on the phenotypic effect of a single gene's functional disabling or removal. The good agreement between our in silico predictions and systematic gene deletion experiments allows us to take these tools from the *descriptive* to the *predictive* phase, inferring the essentiality of a substantial number of genes whose phenotypic profile is unknown.

Functional Classes

The phenotypic effect of a single gene deletion on the viability of an organism is traditionally thought to depend on the individual activity of its protein product. However, an emerging postgenomic view expands a protein's role into an element in a network of protein-protein interactions with a contextual or cellular function [4, 5]. This realization requires the grouping of gene products into functional classes based on their 'cellular role', rather than their individual function. Thus, following the categories established by the Yeast Protein Database (YPD) [6], we use the term 'function' or 'functional class' to describe the cellular role a protein is engaged in, and not its precise biochemical activity. While the available 43 categories are broad, we find that they offer a reasonable starting point for our study. Thus, we followed the categorization of YPD, in which a total of 3,765 proteins have been assigned to one or several of 43 functional classes, leaving 2,547 proteins with unknown function [6].

The bar plot shown in figure 1a compiles the result of systematic gene deletion studies [7] with the corresponding functional categorization of gene products (table A, Supplementary Material). It shows that there are wide differences between the percentage of proteins that are essential or nonessential in the different functional classes, variations that often correlate with biological expectations. For example, the functional class with the largest percentage (~60%) of essential genes is the one encompassing the proteins that are required for RNA splicing (class No. 37). Indeed, the proteins (together with additional RNA components) involved in this process form the splicesome [8], which can be viewed as a distinct module with a key molecular function [9]. Thus the absence of one of the protein subunits could easily jeopardize splicesome function and, consequently, cell viability. On the other hand, among the proteins responsible for small molecule transport (class No. 42), only a small fraction (4.9%) are essential, as the individual deletion of single transporters apparently can be mostly compensated by the activity of other transporters and/or activation of alternative metabolic pathways. As the black bars in figure 1a indicate, in each functional class the deletion phenotype of several genes remains unknown. Yet, assuming that within a functional class the phenotypically tested gene products do not represent a biased subset, we can assign a likelihood of essentiality to all genes based on the observed phenotypic ratios. Indeed, if functional class Y has N_Y proteins with known phenotypic effect, of which N_Y^{lethal} are known to be essential, then each unknown protein has an $f_Y = N_Y^{lethal}/N_Y$ probability to be essential. While f_Y differentiates between genes belonging to different functional classes, the prediction obtained this way has a low resolution, as it assigns to all tein synthesis and folding, while another might handle DNA packaging and nuclear transport. In this perspective, the function of a protein is the function of the module to which it belongs, much as one might see the function of an individual within an organization as that of the department to which he or she belongs – marketing, sales, production, etc.

Does the function of a protein in this sense offer any information about the likely 'essentiality' of its associated gene? Jeong et al. show that it does by exploiting the results of systematic studies of the phenotypic consequences of gene deletions in *S. cerevisiae* as stored in the Saccharomyces Genome Database (SGD). One can consider a gene (or its associated protein) to be essential if its deletion leads to the loss of cell viability; otherwise it is inessential. The Yeast Protein Database has established 43 distinct functional classes for yeast proteins, and the gene deletion data reveal that the proteins within some of these classes have a significantly higher probability of being essential than those in others. Within the class of proteins involved in RNA splicing, for example, roughly 60% turn out to be essential. Of proteins responsible for small molecule transport, only 4.9% turn out to be essential. Some functional units are clearly more crucial than others.

Gene deletion data are not available for all yeast genes. But this analysis makes it possible to hazard a 'first-order' guess for the probability that any untested gene will be essential. Jeong et al.'s strategy is simple: for any untested gene, first identify the functional class into which its protein falls. Then, from the available gene-deletion data, calculate the fraction of genes from this class that are essential, and take this as the estimate that the untested gene will also be essential. This is a baseline method of target identification, and not terribly accurate, as it treats all genes from any family as equally essential. The aim of this paper is to show how one can improve on this estimate by bringing further, independent data to bear.

To begin with, consider data on gene expression, which are publicly available for a large number of yeast genes. One data set, for example, records mRNA expression levels

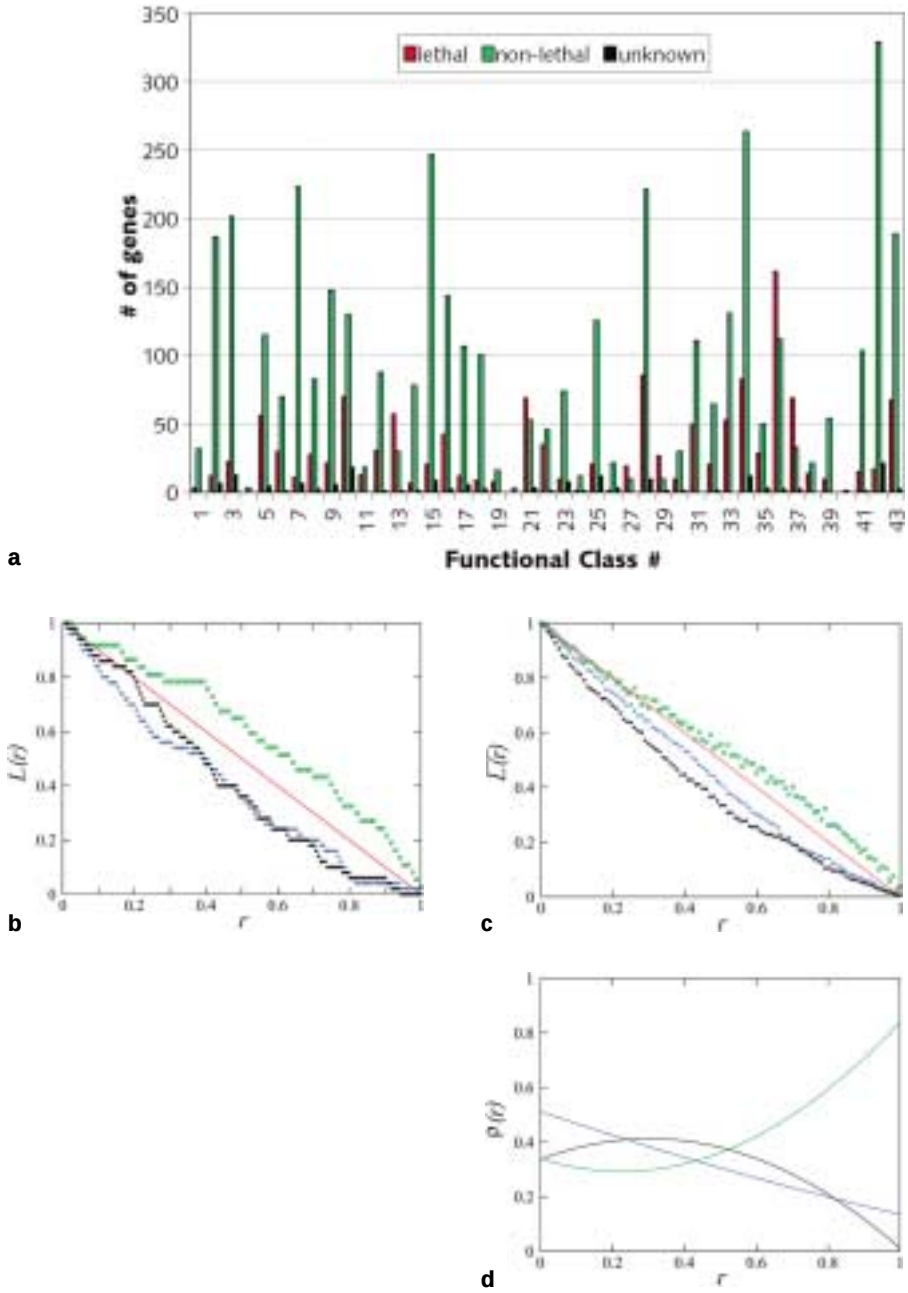


Fig. 1. Correlations between gene essentiality and various generic characteristics. **a** Essentiality vs. cellular function. The number of genes with essential (red bar), nonessential (green bar) or unknown (black bar) deletion phenotypes in each of the 43 functional classes is shown. The list of the functional classes is given in table A, Supplementary Material. **b** The lethality curve $L(r)$ for functional class 31 (protein degradation). The three symbols correspond to the lethality curves obtained from the 251 gene deletion (black), the 63 control microarray data (blue), and from the protein-protein interaction data (green). The horizontal axis denotes the normalized ranking $r = R/N_{31}$. The diagonal shown as a continuous red line represents the expected curve if gene essentiality is independent of gene expression fluctuations or the number of protein interactions. **c** The lethality curve $\overline{L(r)}$ averaged over all 43 modules, similar to that shown in figure 1b. **d** The lethality probability $\rho(r)$ as predicted from the data shown in figure 1b for the functional class 31 using the mathematical tools described in the Appendix. Note that the vertical axis represents a normalized probability. These predictions assign to each gene the probabilities $\rho_{\sigma 251}$, $\rho_{\sigma 63}$, and ρ_k, providing the expected phenotypic likelihood that a given gene is essential based on the three available datasets σ_{251}, σ_{63}, or k. Such predictions, issued separately for each gene, allow us to generate the ranking summarized in figure 3. For this we have first fit a third order polynomial to the lethality curves shown in figure 1c, and determined $\rho(r)$ using the methods described in the Appendix. The slightly non-monotonic nature of $\rho(r)$ as inferred from the 251 microarray dataset and the protein interaction data (black and green) is somewhat unexpected. While we could have used a lower order polynomial for the fit to eliminate this non-monotonicity, we were reluctant to bias the data.

genes within the same functional class an identical probability to be phenotypically essential. Note that in reality the set of phenotypically tested genes are somewhat biased towards biologically 'important' genes, of which one expects a larger essentiality ratio than random. Yet, as systematic gene deletion methods are close to cover all known yeast genes, this bias is not particularly relevant at this point. Extending our method to organisms for which the deletion phenotype profiling is incomplete would require addressing the effect of potential bias in gene deletion.

Expression Patterns

To refine our ability to predict deletion phenotypes we next examined the characteristics of large-scale mRNA expression patterns, as captured by two publicly available large DNA microarray datasets. The first set provides-steady state mRNA expression data in wild-type *S. cerevisiae* sampled 63 separate times ('63 control' set) [10]. A second set provides data on individual cDNA microarray measurements on 251 viable yeast derivatives with a single ORF deletion [10], see also Methods. Starting with the 251 gene deletion data, for each measured with DNA microarrays at 63 different times for wild-type *S. cerevisiae*. Another gives similar data for genes in 251 viable yeast derivatives, each derived by deleting one specific gene. The levels of expression for each gene vary (from one moment to the next in the first data set, or from one yeast derivative to the next in the second). For each data set, Jeong et al. calculated the standard deviation σ of the expression levels for each of the genes for which deletion data is available. This offers a statistical measure of how vigorously the expression of each gene fluctuates, and it turns out that the expression levels of some genes fluctuate far more than others. Do these fluctuations carry some information about how likely a gene is to be essential?

gene i we determined the standard deviation (σ_i) in their mRNA expression across the 251 different transcriptomes. Next, we selected a functional class Y, with N_Y genes with known deletion phenotype, and rank ordered the genes based on σ_i by assigning each gene i a rank R_i, so that the gene with the smallest σ has $R_1 = 1$, and the gene with the largest σ has $R_{N_Y} = N_Y$. Assigning a variable $\delta_i = 1$ (0) to each essential (nonessential) gene, we obtain the lethality function $L_\sigma(r)$ by summing δ_i from the large to small $r = R/N_Y$. An illustrative result for functional class 31 is shown in figure 1b. If phenotypic essentiality does not correlate with σ, i.e., if an essential gene can possess any σ value, $L_\sigma(r)$ should follow the straight diagonal line shown in the figure. On the other hand, we find that $L_\sigma(r)$ displays a systematic downward deviation from the random essentiality diagonal, an indication of correlations between gene essentiality and σ. In particular, the downward deviation demonstrates that the large r region (corresponding to high σ) has fewer essential proteins than the small r (i.e., small σ) region. This tendency is clearly seen in fig-ure 1c, where we show $\overline{L_\sigma(r)}$ averaged over all 43 functional classes, displaying a clear downward deviation from the diagonal, indicating that, on average, genes with small σ are more likely to be essential than genes with large σ. These results indicate the presence of robust feedback mechanism(s) within the genetic network of yeast cells which, upon perturbations, maintain the mRNA expression level of those genes that encode an essential protein, on average within a narrower range, while nonessential ones are allowed to fluctuate more widely. Using the same procedure, we have obtained similar results for the 63 control data set as well. As figure 1b, c shows, the obtained $L_\sigma(r)$ curve displays a downward deviation from the random essentiality diagonal, though the effect is somewhat less pronounced than that obtained for the 251 gene deletion set. The decreased fluctuations in the expression level of essential genes reflects the potential existence of feedback mechanisms that stabilize their expression level. Indeed, wide fluctuations in the expression level of the essential genes could lead to the death of the organism. Such constraints are less needed for non-essential genes, a fact reflected in the correlations between essentiality and σ in both the 251 gene deletion and the 63 control dataset.

Protein Interactions

Recent two-hybrid experiments in *S. cerevisiae* [11, 12], complemented with additional experimental data [13, 14], approximate the number of potential physical interactions a gene product possesses (although with a significant number of false positives and negatives) [15]. In a previous study [5] we demonstrated that the architecture of the resulting protein-protein interaction network is scale free [16], which implies that highly connected proteins play a more important role in guaranteeing the network's integrity than their less connected counterparts [5, 17]. Correlating this property with known deletion phenotypes we previously found that proteins with more interactions were more likely to be essential than less connected ones [5]. To further refine this observation, we selected in each functional class those proteins for which k and their deletion phenotype are known simultaneously, where k is the number of potential links a protein has with other proteins. We rank ordered the corresponding genes based on k, assigning $\delta_i = 1$ (0) to essential (nonessential) genes and summed δ_i starting from the most connected proteins towards the least connected one, obtaining for example the $L_k(r)$ lethality curve shown in figure 1b for functional class 31. If the likelihood that a gene is essential is independent of k, then $L_k(r)$ should follow the straight diagonal. The fact that $L_k(r)$ systemically deviates above this line indicates that highly connected nodes are more essential than their less connected counterparts, the degree of deviation offering a measure of the correlation between a protein's connectivity and its essentiality within a given functional class. We find that, to a varying extent, such correlation is present in all functional classes. Indeed, we determined $L_k^Y(r)$ for each of them separately, then calculated its average $\overline{L_k(r)}$ over all 43 functional classes.

To show that they do, Jeong et al. carry out a simple mathematical procedure within each functional class of proteins. Suppose that in some functional class Y, the consequences of gene deletion are known for N_Y genes. Arrange these genes in increasing order according to the standard deviation σ_i of their expression data, and then label each with an integer index – the rank R – starting at 1 on the left and increasing to N_Y on the right. Now, to find out if the level of fluctuations correlates in any way with gene essentiality, one can start at the right end of the list (where fluctuations are high) and see how one encounters essential genes when moving to the left (where fluctuations are low). Jeong et al. consider an informative 'lethality' function $L_\sigma(r)$ – written in terms of the scaled variable $r = R / N_Y$ – that one might compute along the way. To begin, at the far right, set $L_\sigma(r = 1) = 0$. Then, moving from right to left as r decreases from 1 to 0, define $L_\sigma(r)$ to remain constant except when one encounters an essential gene, in which case its value jumps by one.

Defined in this way, the points where $L_\sigma(r)$ suddenly jumps up in value mark the essential genes. Consequently, this lethality function has the form of an irregular set of steps going roughly upward from right to left. But the shape of this staircase reveals the presence or absence of correlations between gene essentiality and fluctuations in expression. If there is no correlation, then all genes – regardless of their place on the list – have the same chance to be essential. Hence, the function should follow roughly along some straight-line diagonal. Alternatively, if the chance of gene essentiality increases or decreases with decreasing levels of fluctuation in expression, then the function should begin to curve. If the likelihood of essentiality increases (or decreases) with decreasing fluctuations, then the function should curve upward (or downward).

The authors carried out this procedure for each functional class, and figure 1b shows representative results for functional class 31 (corresponding to proteins involved in protein degradation). In this figure, the black symbols represent calculations from the 251 viable yeast derivatives and the green from the 63 samples from wild-type

As shown in figure 1c, the obtained $\overline{L_k(r)}$ curve displays an upward deviation from the random essentiality diagonal, demonstrating a nonuniform essentiality within functional classes.

Predicting Lethality

To take full advantage of the observed correlations between essentiality and the ranking of the gene products, we developed a mathematical algorithm (Appendix) to predict the likelihood that a given gene is essential for the viability of *S. cerevisiae*. For each gene product with a known or unknown deletion phenotype we predict the probability, $\rho(r)$, that its ablation from the yeast proteome would be lethal using as an input the $L(r)$ curves fitted individually to each of the 43 functional classes and eq. (4) in the Appendix. A set of representative probabilities for functional class 31 – derived from the $L(r)$ curves in figure 1b – is shown in figure 1d. For example, the blue line based on the 63 control data in figure 1d predicts that a gene from class 31 with ranking 0.8 has a 20% chance of being essential. By contrast, the green curve indicates that when based on protein interaction a different gene with the same ranking has a ~60% chance to be essential, as protein interaction essentiality increases with their ranking.

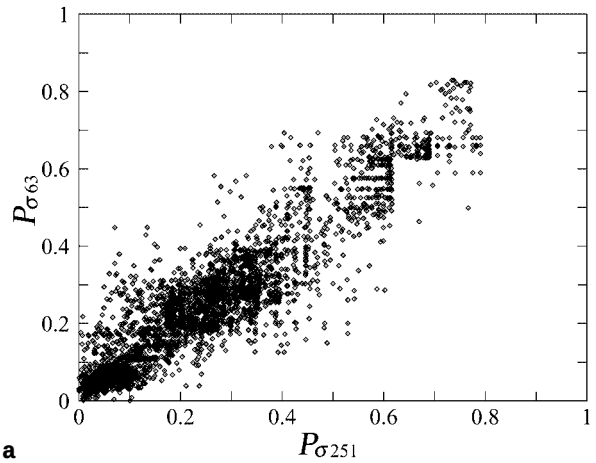
As the methods based on σ_{251}, σ_{63}, or k each offer a separate set of predictions for each gene product, denoted by the probabilities $\rho_{\sigma 251}$, $\rho_{\sigma 63}$, and ρ_k, we need to investigate to which degree they agree with each other. The scatter plot shown in figure 2 contains all of the 2,350 genes for which information is simultaneously available by all three methods. The clustering of the points around the linear 45° diagonal indicates a relatively strong correlation between the predictions based on the three separate databases. Yet, as the different datasets are complementary, it may be desirable to combine all three to improve our predictions. If all three predictions were independent from each other, the product $\rho_{\text{prod}} = \rho_{\sigma 251} \times \rho_{\sigma 63} \times \rho_k$ should be used as a combined lethality measure for each gene. An alternative approach would be to use the largest lethality probability offered by the three methods $\rho_{\max} = \text{Max}(\rho_{\sigma 251}, \rho_{\sigma 63}, \rho_k)$, with the assumption that a stronger signal is more likely to be relevant than a weaker one.

To test the validity of our method against experimental results, we first utilized data deposited in the *Saccharomyces* Genome Database (SGD) by the international deletion project consortium (http://www-deletion.stanford.edu) and compared them with our mathematical predictions. In this repository, each gene that has been deleted has its systematic deletion phenotype (viable/lethal/slow growth) identified based on identical growth conditions [7]. From the database, we rank ordered those 3,543 gene products for which both deletion phenotype and functional classification is known. Based on their predicted lethality probability ρ, we placed first those that have the highest probability to be essential. The quality of our predictions can be measured by the separation of the essential from the nonessential genes. In table 1 (see p. 28) we quantify this separation by showing the percentage of known essential genes in the first 20% (predicted most essential) and the last 20% (predicted least essential) portion of the list. As the table indicates, each of the five predictions was successful at segregating the essential genes. For example, the $\rho_{\max}$ method assembles 48.3% of all known essential genes into the top 20% of the lethality list. The method works even better at assigning low lethality probability to the nonessential genes: only 3% of the essential genes can be found at the bottom 20% on the $\rho_{\max}$ list, i.e., practically all genes assigned to the bottom fifth of the list are known to be nonessential. The fact that the confidence level of each of the five methods is comparable indicates that any of them can be used to individually predict lethality. Thus, the lack of availability of one or even two of the datasets for a given target organism would not jeopardize our ability to use the developed methodology. Nevertheless, in the following we focus on the $\rho_{\max}$ method to further validate our tools, as we find that it offers slightly better predictions.

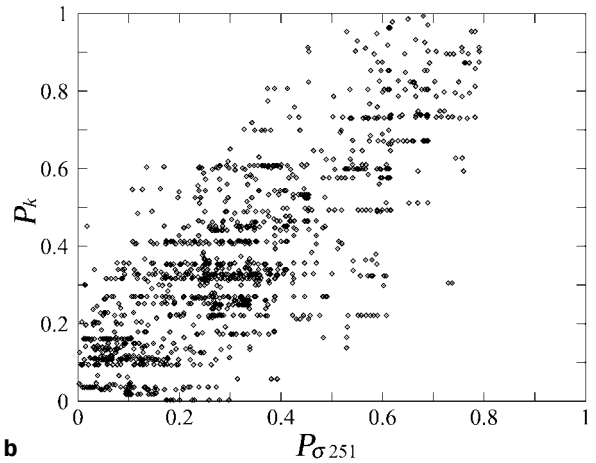
S. cerevisiae. As the data reveal, the chance of gene essentiality in this class becomes larger with decreasing σ, causing the lethality function to curve upward. The benefit of doing the analysis in terms of the scaled variable r is that it lets one combine the results from different functional classes in a natural way (since r always ranges from 0 to 1). This makes it easy to calculate lethality functions averaged over all functional classes (fig. 1c), which again show an upward curvature. So the fluctuations in gene expression levels do appear to reflect the essentiality of the gene, with lower levels of fluctuations being associated with a high chance of essentiality. One should be able to use this correlation to make better predictions for the likelihood of essentiality of untested genes than one could do using functional information alone.

Before turning to such predictions, however, Jeong et al. first note that another kind of data has the potential to be exploited in a similar way. Recent experiments have constructed an estimate of the topology of the protein-protein interaction network for *S. cerevisiae*. Suppose that each protein is a node in the network, and that two proteins are linked if they interact with one another. It turns out that within this network, not all proteins participate in the same number of links. In earlier work, these authors and others have shown that the network has a 'scale-free' character, i.e. the probability that a protein participates in k interactions follows a power-law function, $P(k) \sim k^{-\gamma}$, where γ is a constant. One consequence of this distribution is that a small number of proteins are very highly connected and play the role of 'hubs'. These proteins turn out to be crucial to guaranteeing the basic topological features of the network, such as its diameter – the number of links it takes, on average, to go between any two randomly selected proteins in the network. From a topological point of view, the loss of a node has consequences that grow with the node's 'degree' – its number of links to other proteins.

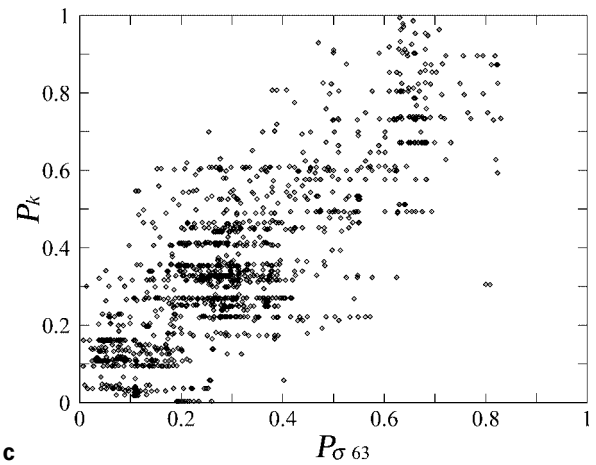
For this reason, it is plausible to suspect the connectivity of a protein might also carry information about its essentiality. To draw out this correlation, Jeong et al. again build a lethality function $L_k(r)$ within each



a



b



c

Fig. 2. Correlation between the predicted lethality probabilities. The three datasets, based on σ_{251}, σ_{63} and k offer independent predictions for the essentiality of the individual genes. The figure shows the correlations between these predictions. **a** Scatter plot showing the lethality probability as predicted by the σ_{251} and σ_{63} for each of the 3,543 genes for which microarray data and functional classification are available simultaneously. **b** Scatter plot showing the lethality probability as predicted by the σ_{251} and k. **c** The same for σ_{63} and k for each of the 1,425 genes for which functional classification and protein interaction information are simultaneously available. The clustering of the datapoints along the 45° diagonal indicates that the predictions offered by the three methods correlate with each other.

functional class. They follow precisely the same procedure as before, but now considering protein degree k rather than gene expression fluctuations σ. As the green data in figures 1b, c reveal, the lethality function for functional class 31 and also for the average over classes now curves downward, implying that proteins of higher degree tend to be essential more frequently than those of lower degree.

This brings us to the main point of the paper – how to integrate this information so as to make improved predictions concerning the likely essentiality of untested genes. The authors do this by using their lethality functions [either $L_\sigma(r)$ or $L_k(r)$] calculated for each of the 43 different functional classes. These functions were calculated based on the empirical data by summing over the various genes (in order of decreasing σ or k), adding 1 to the function for each essential gene and a 0 otherwise. In this way, each lethality function is proportional to a sum over the probability of essentiality for the genes in that class with different values of r, i.e. $L(r) \sim \int^r \rho(r')dr'$. Hence, an estimate of the probability $\rho(r)$ can be obtained by differentiation. This gives $\rho(r) = c\partial L(r)/\partial r$, where c is a constant that is determined by demanding that the sum of the probabilities within each functional class be equal to the number of lethal genes in that class (see eq. 3 in the Appendix). The mathematical formula of eq. 4 follows immediately.

How well does this algorithm perform? Since Jeong et al. assembled their lethality curves from several different data sets – two for gene expression profiles and one for protein centrality – they actually have three distinct procedures. It makes sense then to put this data together to produce one best estimate. One way to do this is simply to take the product of the various probabilities (suitably normalized). Another is to define the overall probability of a gene's likelihood of essentiality as the maximum of the probabilities offered by the independent data sets. Jeong et al. argue that ρ_{max} defined in this way should be more accurate than any one estimate alone, as stronger signals are more likely to be relevant than weakerones.

To test the accuracy of the approach, the team first used the SGD database covering

Figure 3a summarizes our main results by predicting the deletion phenotype of a major portion of the *S. cerevisiae* genome, listing all 3,656 genes with known cellular functions in decreasing order of predicted lethality probability, ρ_{max}(see also tables C and D, Supplementary Material). Note that the first line, which contains the predicted most essential genes of *S. cerevisiae*, is indeed dominated by genes known to be required for their viability, demonstrating that the algorithm efficiently differentiates between essential and nonessential genes (see also table 1, p. 28). Indeed, we find that in the first 20% there are 2.41 times more essential genes than expected, if the method would assign essentiality randomly, collecting more than 48% of all known essential genes into this line. Conversely, in the last line only 23 essential genes out of a total of 728 genes are found, indicating that genes with a predicted low likelihood of essentiality are indeed, with very high confidence, nonessential. Also, note that while not incorporated into our model explicitly, the predictions assign the highest level of essentiality to those proteins that participate in several functional classes. By contrast, the end of the list, with few predicted essential proteins, is dominated by those that participate in one or at most two functional classes (figure A, Supplementary Material).

As of April, 2001, the deletion phenotype of 440 *S. cerevisiae* genes haves not yet been deposited in the SGD database. To further investigate the validity of our method and to predict the deletion phenotype of experimentally untested gene products, we have selected those 113 gene products for which functional class assignment is available. The deletion phenotype of some of these genes was previously determined under various (nonsystematic) experimental conditions, and collated by the YPD/Proteome database [6]. We sorted all 113 genes from the most to the least essential based on their computationally predicted ρ_{max} and compared the results to the available experimental data. As shown in figure 3b (see also table E, Supplementary Material), there was a general agreement between the available nonsystematic experimental data and our predictions. Indeed, we find that of the first 12 genes listed, 9 are either lethal or have growth defects, based on the predicted ρ_{max}, and only 3 known lethal genes can be found outside of this domain. In addition, figure 3b indicates that the untested gene products, shown as white boxes, are not likely to represent essential proteins of the *S. cerevisiae* proteome. This is supported by the predicted probabilities (ρ_{max}) as well: we find that ρ_{max} drops rapidly from 70% chance of being essential, assigned to the first gene shown in figure 3b, to less than 40% starting from gene 12 on (and decaying approximately exponentially after that).

Over the last decades the search for antimicrobial and antifungal agents has been largely restricted to well-known compound classes active against a standard set of drug targets. Recent advances in genomics provide an opportunity to expand the range of potential targets. These include identification of genes associated with pathogenic processes [18, 19] or using systematic gene expression profiling to collect information about cellular response to treatment with various drug candidates [10]. However, the full power of genomics can be exploited only with the introduction of powerful algorithms that use all available genome-derived information for novel drug target identification.

Based on the results obtained in *S. cerevisiae*, here we propose a general scheme for putative drug target identification based on genomic data. Following the identification of most – or all – open reading frames (ORFs), which is now accomplished for hundreds of microorganisms, there are two prerequisites for the utility of our computational method. The first requirement is the identification of functional class lethality factors, for which both the assortment of ORFs into functional classes and representative deletion phenotyping within these groups are needed. Although by no means trivial, an increasing number of computational methods for assigning protein function are available [4, 20–22], while single gene deletion phenotyping can now be relatively quickly accomplished by RNA interference analysis [23, 24]. Also, the functional class lethality factor values may be highly similar among closely related microorganisms. Indeed, we find that of the 3,543 yeast genes for which both deletion phenotype and functional classification are known. Using the predicted probabilities from their algorithm (both independently and in the two combined methods), Jeong et al. listed these genes in decreasing order of likelihood to be essential. They then compared these probabilities with reality. The ρ_{max} method places 48.3% of all the known essential genes into the top 20% of the lethality list. In other words, half of those genes that really were essential showed up in the highest fifth of the list of likely candidates. Meanwhile, only 3% of the essential genes were incorrectly placed in the bottom 20% of the list. Jeong et al. find that ρ_{max} offers the best predictions, but the individual data sets all do roughly as well as one another (table 1, p. 28); hence this method should be useful even if just one dataset could be obtained.

As another test, they used the method to predict the essentiality of 113 yeast genes that were not yet (as of April 2001) deposited in the SGD database. The deletion phenotypes of these genes were, however, determined in other non-systematic experiments and this data was collected by the YPD/Proteome database. Again, the predicted ρ_{max} for these 113 genes compares well with the experimental data. Of the first twelve genes – those with the highest ρ_{max} values – nine turned out to be either lethal or to induce serious growth defects. Only three lethal genes appeared elsewhere on the list.

All of this suggests that this manner of predicting gene essentiality may offer a practical means for putting some of the emerging genomics and proteomics data to work in identifying drug targets. As Jeong et al. point out, the search for antibiotic or antifungal drugs over the past two decades has focused on 'well-known compound classes that are active against a standard set of drug targets'. The advent of genomics offers means to expand the range of targets, especially if one can produce mathematical algorithms that can identify targets automatically by bringing together a wide variety of genomic information.

Mark Buchanan

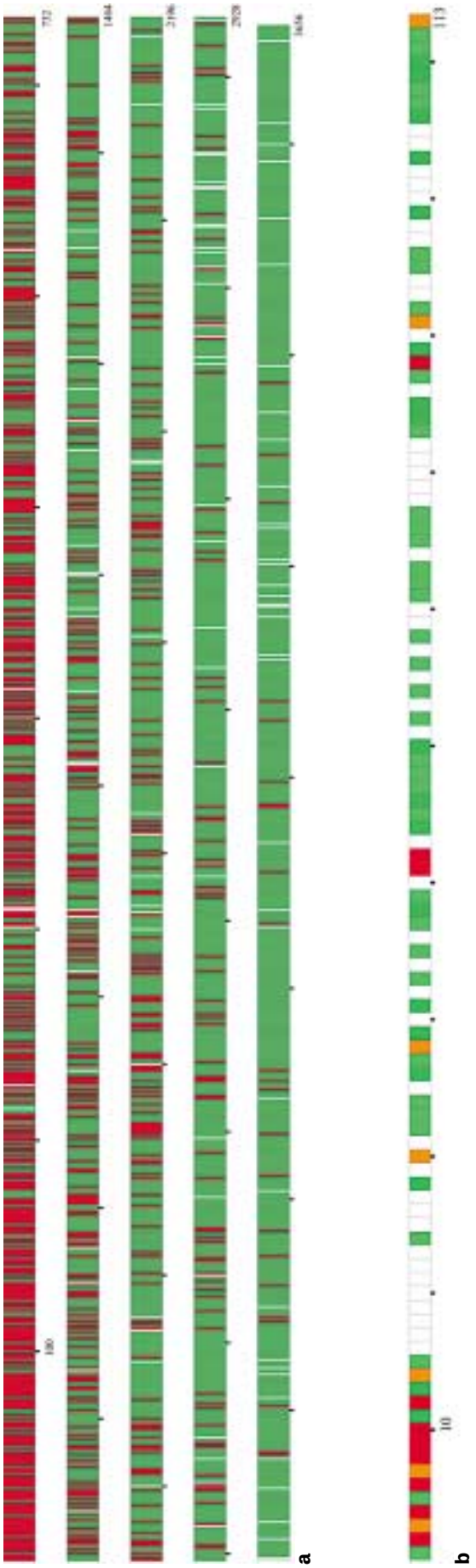


Fig. 3. Computational predictions versus experimental results. Summary of the predictions offered by ρ_{max}, integrating the data from all three databases, based on σ_{251}, σ_{63}, and k, for those 3,656 genes from the SGD database (**a**) and for those 113 genes out the 442 that were not tested by the yeast deletion consortium (**b**), for which functional classification is available. Each box in (**a**) and (**b**) corresponds to a separate gene, the first from the left having the largest probability of being essential, while the last gene on the right end of the fifth line has been assigned the lowest probability to be essential. The colors correspond to essential (red), nonessential (green) and unknown (white) in (**a**) and essential (red), nonessential with growth defect (orange), nonessential with normal growth (green), and unknown phenotype (white) in (**b**).

isms, potentially allowing the establishment of a value range for each functional class. The second requirement is the generation of a series of global gene expression measurements on wild-type cells or on viable deletion mutants, possibly in conjunction with systematic protein interaction data.

While the method proposed here is by nature probabilistic, i.e. it offers the likelihood of essentiality of a given gene product, nevertheless it clearly indicates the utility of inferring functionally relevant correlations from the available genomic databases for systematic drug target identification. The further improvement of computational algorithms, the increasing availability of systematically collected biologic data, and a better understanding of dynamic [25, 26] and biologic noise effects [27, 28] are likely to significantly enhance the role of such methods in drug discovery.

Methods

Microarray analysis

We used the 'control_expts1-63_ratios. txt' and 'data_expts1-300_ratios.txt' files from the data package that is publicly available at http://www.rii.com/tech/pubs /cell_hughes.htm [10]. Of the second (300) set, 13 expression profiles represent treatment of wild-type yeast cells with various chemical compounds, the other 287 represent single gene deletion mutants grown under the same steady-state conditions as wild-type yeast cells. Two hundred and seventy-six of these are from mutants that are viable in the absence of the deleted gene product, but a number of these also possess secondary genetic aberrations [29], resulting in 251 mutants with no known additional genetic change ('251 gene deletion' set). The relative changes in gene expression level of gene i under experiment j is defined as $\Delta e_{ij} \equiv (e_{ij} - e_i^0)/e_i^0 = e_{ij}/e_i^0 - 1$, where e_i^0 is unperturbed gene level. From this we calculated the standard deviation

$$\sigma_i \equiv \sqrt{\frac{1}{N_j}\sum_j (\Delta e_{ij} - \overline{\Delta e_i})^2}$$

of Δe_{ij} over $N_j = 251$ and 63 experiments, where

$$\overline{\Delta e_i} \equiv \frac{1}{N_j}\sum_j \Delta e_{ij}$$

is the average change in the expression for gene i.

Protein Interaction Map Analysis

Using the results of the two hybrid method of Uetz et al. [11], we assigned to each gene i the number of interactions k^i its protein product is known to participate in.

Determining $L(R)$ Curves

For each module $Y = 1,\ldots,43$, we determine the $L(R)$ curves following the same 5 steps: (a) rank the genes based on σ_{251}, σ_{63} or k; (b) assign a $\delta_i = 0,1$ variable to each gene i whose lethality is known; (c) determine the $L^Y(R)$ curve by summing δ_i starting from the gene with the highest ranking $R = N_Y$, moving towards $R = 1$. The obtained $L^Y(R)$ curves starts at the $(N_Y, 0)$ point and ends at $(0, f_Y N_Y)$ coordinates; (d) normalize the $L^Y(R)$ curves by dividing the x-axis with N_Y and the y-axis with $f_Y N_Y$ for each functional class, and (e) after normalization, fit a third-order polynomial $L_x(r) = a_x r^3 + b_x r^2 + c_x r + 1$ to the obtained curve, where the subscript x stands for either of σ_{251}, σ_{63}, or k. The coefficients for each functional class are shown in table B, Supplementary Material. The functional classes 1, 4, 14, 19, 20, 24, 26 and 40 contain fewer than 8 lethal genes, thus individual fitting was not reliable. For these we used the average lethality curves, shown in figure 1c. Consequently, for the genes belonging to these classes the confidence level of the predictions is lower than for the genes in the remaining 35 classes. As the dominant contribution to ρ is given by the functional classes, i.e. f_Y, the difference is not substantial.

Predicting Lethality

For each functional class Y we rank ordered *all* genes based on σ_{251}, σ_{63}, or k (if known), whether their phenotypic effect is known or not. Using eq. 4 we calculate ρ, the lethality probability for each gene i with rank R. The results are summarised in table C, Supplementary Material.

Supplementary Material

This material is available on our designated website (http://www.nd.edu/~networks/cell).

Acknowledgements

We would like to acknowledge all members of the various projects for making their database publicly available for the scientific community. We also thank Anna-Lisa Somera (Northwestern University) for data collection for figure 3b. Research at the University of Notre Dame was supported by the National Science Foundation, Department of Energy and National Institute of Health and at Korea Advanced Institute of Science of Technology by Ministry of Information and Communication (IMT2000-B3-2) and at Northwestern University by grants from the National Cancer Institute.

References

1 Rosamond J, Allsop A: Harnessing the power of the genome in the search for new antibiotics. Science 2000;287:1973–1976.
2 Gibbs JB: Mechanism-based target identification and drug discovery in cancer research. Science 2000;287: 1969–1973.
3 Bailey D, Zanders E, Dean P: The end of the beginning for genomic medicine. Nat Biotechnol 2001;19: 207–209.
4 Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: Protein function in the post-genomic era. Nature 2000; 405:823–826.
5 Jeong H, Mason SP, Barabási A-L, Oltvai ZN: Lethality and centrality in protein networks. Nature 2001;411: 41–42.
6 Costanzo MC, et al: The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res 2000;28:73–76.
7 Winzeler EA, et al: Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 1999;285:901–906.
8 Staley JP, Guthrie C: Mechanical devices of the spliceosome: Motors, clocks, springs, and things. Cell 1998;92:315–326.
9 Hartwell LH, Hopfield JJ, Leibler S, Murray AW: From molecular to modular cell biology. Nature 1999;402 (suppl):C47–C52.
10 Hughes TR, et al: Functional discovery via a compendium of expression profiles. Cell 2000;102:109–126.
11 Uetz P, et al: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 2000;403:623–627.
12 Ito T, et al: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001;98:4569–4574.
13 Xenarios I, et al: DIP: The database of interacting proteins. Nucleic Acids Res 2000;28:289–291.
14 Mewes HW, et al: MIPS: A database for genomes and protein sequences. Nucleic Acids Res 2000;28:37–40.
15 von Mering G, et al: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002;417:399–403.
16 Barabási A-L, Albert R: Emergence of scaling in random networks. Science 1999;286:509–512.
17 Albert R, Jeong H, Barabási A-L: Error and attack tolerance of complex networks. Nature 2000;406:378–382.
18 Camilli A, Mekalanos JJ: Use of recombinase gene fusions to identify *Vibrio cholerae* genes induced during infection. Mol Microbiol 1995;18:671–683.
19 Heithoff DM, Conner CP, Mahan MJ: Dissecting the biology of a pathogen during infection. Trends Microbiol 1997;5:509–513.
20 Brown MP, et al: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000;97:262–267.
21 Pavlidis P, Cai J, Weston J, Grundy WN: Gene classification from heterogenous data. Recomb 2001 http://www.cs.columbia.edu/compbio/exp-phylo/.
22 Wu LF, et al: Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. Nat Genet 2002;31: 255–265.
23 De Backer MD, et al: An antisense-based functional genomics approach for identification of genes critical for growth of *Candida albicans*. Nat Biotechnol 2001;19:235–241.
24 Elbashir SM, et al: Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature 2001;411:494–498.
25 Holter NS, et al: Fundamental patterns underlying gene expression profiles: Simplicity from complexity. Proc Natl Acad Sci USA 2000;97:8409–8414.
26 Alter O, Brown PO, Botstein D: Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000;97: 10101–10106.
27 McAdams HH, Arkin A: Stochastic mechanisms in gene expression. Proc Natl Acad Sci USA 1997;94:814–819.
28 Hasty J, Pradines J, Dolnik M, Collins JJ: Noise-based switches and amplifiers for gene expression. Proc Natl Acad Sci USA 2000;97:2075–2080.
29 Hughes TR, et al: Widespread aneuploidy revealed by DNA microarray expression profiling. Nat Genet 2000;25:333–337.

Appendix

Algorithm for predicting gene deletion phenotypes

Consider an arbitrary functional class Y, which has N_Y genes with known phenotypic effect, and of which a fraction $f_Y = N_{lethal}/N_Y$ are essential. The genes in this class can be ranked based on σ_{251}, σ_{63}, or k (fig. 2). While we use σ as the ranking variable to introduce the key expressions, the results can be readily generalized for k as well. Let us denote by $P_Y(\sigma)$ the probability that a gene within functional class Y has standard deviation equal to σ. The rank R of a gene with standard deviation σ is defined as

$$R = N_Y \sum_{\sigma' < \sigma} P_Y(\sigma') , \quad (1)$$

This σ based ordering assigns to the gene with the lowest (largest) σ the ranking $R = 1$ ($R = N_Y$). The $L_\sigma(r)$ curve shown in figure 1b is obtained by assigning $\delta = 0,1$ (essential, nonessential) to each gene within Y based on their known phenotypic effect, plotted in function of the reduced ranking variable $r = R/N_Y$. Denoting with $\rho(r)$ the probability that a gene with ranking r is essential, the $L_\sigma(r)$ curve can be written as $L_\sigma(r) \sim \int \rho(r)$, which implies that the probability

$$\rho(r) = c \left. \frac{\partial L_\sigma(r)}{\partial r} \right|_r , \quad (2)$$

where c is a normalization constant that can be obtained from the equation

$$\frac{1}{N_Y} \sum_r \rho(r) = f_Y , \quad (3)$$

providing the probability for phenotypic essentiality as

$$\rho(r) = \frac{L_\sigma(0)}{\sum_{r'} \left. \frac{\partial L_\sigma(r)}{\partial r} \right|_{r'}} \left. \frac{\partial L_\sigma(r)}{\partial r} \right|_r . \quad (4)$$

Note that by using f_Y in the normalization (3), we automatically include the functional class' overall phenotypic fraction given by figure 1a. If $L_\sigma(r)$ is known, (4) can be used to determine the probability that a given gene product is essential. For this we fit with a third-order polynomial the $L_\sigma^Y(r)$ curves obtained for each functional class, as shown in figure 1c for the averaged data. The coefficient of the polynomial fit for each functional class is given in Supplementary Material.

Table 1. Quantitative comparison between the predictions offered by the three datasets and the combined ρ_{max} and $\rho_{product}$ methods

	Lethal genes relative to total lethal numbers in the first 20%	Lethal genes relative to total lethal numbers in the last 20%
ρ_{max}	48.3	3.0
$\rho_{\sigma_{251}}$	46.5	2.4
$\rho_{\sigma_{63}}$	42.7	3.5
ρ_k	42.5	4.5
$\rho_{product}$	45.5	2.3

The first column shows the percentage of all the genes that are known to be essential that are in the top fifth of our lethality list, predicted by the five different methods. For example, of all 824 genes known to be essential, the ρ_{max} method selects 48.3%, i.e., 398 of them in the first 20% of the lethality list. The second column shows the same quantity in the bottom fifth of the list, indicating, for example, that only 3% (n = 25) of the known essential genes are among the predicted least essential ones. All figures are percentages.