Combining biological networks to predict genetic interactions

Sharyl L. Wong*, Lan V. Zhang*, Amy H. Y. Tong†, Zhijian Li†, Debra S. Goldberg*, Oliver D. King*, Guillaume Lesage‡, Marc Vidal§, Brenda Andrews†, Howard Bussey‡, Charles Boone†, and Frederick P. Roth*¶

*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115; †Banting and Best Department of Medical Research and Department of Medical Genetics and Microbiology, University of Toronto, Toronto, ON, Canada M5G 1L6; ‡Department of Biology, McGill University, Montreal, QC, Canada H3A 1B1; and §Dana–Farber Cancer Institute and Department of Genetics, Harvard Medical School, Smith 858, 1 Jimmy Fund Way, Boston, MA 02115

Communicated by Nancy Kleckner, Harvard University, Cambridge, MA, September 15, 2004 (received for review June 4, 2004)

Genetic interactions define overlapping functions and compensatory pathways. In particular, synthetic sick or lethal (SSL) genetic interactions are important for understanding how an organism tolerates random mutation, i.e., genetic robustness. Comprehensive identification of SSL relationships remains far from complete in any organism, because mapping these networks is highly labor intensive. The ability to predict SSL interactions, however, could efficiently guide further SSL discovery. Toward this end, we predicted pairs of SSL genes in *Saccharomyces cerevisiae* by using probabilistic decision trees to integrate multiple types of data, including localization, mRNA expression, physical interaction, protein function, and characteristics of network topology. Experimental evidence demonstrated the reliability of this strategy, which, when extended to human SSL interactions, may prove valuable in discovering drug targets for cancer therapy and in identifying genes responsible for multigenic diseases.

Mutations into two different genes sometimes confer a significantly more deleterious phenotype than either single mutation alone. Death or pronounced growth deficiency arising in such double mutants is referred to as synthetic lethality or synthetic sickness, respectively.

A comprehensive map of synthetic sick or lethal (SSL) interactions for an inbred laboratory organism may provide a valuable template for understanding the basic principles underlying genetic interaction networks (1–3) in both inbred and outbred populations (4, 5). In humans, genetic interactions are involved in many complex phenotypes and are the defining basis of multigenic genetic disease (6–8). SSL interactions can also be used to find effective drug combinations or to identify novel drug targets for tumor-specific therapy (4, 9). Finally, SSL interactions comprise a network that is far denser than, and largely nonoverlapping with, that of protein interactions (5). Thus, genetic and protein interaction networks provide complementary information.

Due to their combinatorial nature, mapping SSL networks is extremely labor intensive (5, 10), even in genetically amenable model organisms. For example, comprehensive assessment of SSL gene pairs in *Saccharomyces cerevisiae* (with ≈6,000 genes) requires constructing ≈18 million double mutants, including conditional mutations in essential genes. To date, Synthetic Genetic Array (SGA) analysis has been used to assess ≈4% of gene pairs in one growth condition (5, 11). However, full delineation of pairwise interactions requires assessment of mutant phenotypes in many growth conditions. Determining the SSL network for *Caenorhabditis elegans*, *Drosophila melanogaster*, or *Mus musculus* is even more daunting, because construction of double mutants is technically difficult and because these organisms have 10- to 25-fold more gene pairs than *S. cerevisiae*.

A reliable method for predicting SSL interactions, however, may alleviate this experimental bottleneck. The only previous attempt to predict genetic interactions relied on metabolic flux analysis, an approach applicable only to pairs of genes involved in central metabolism (12). Here, we integrate multiple data types to construct probabilistic decision trees with which we predict SSL gene pairs in *S. cerevisiae*. This study represents a rigorous demonstration that genetic interactions can be predicted. This approach should reduce the labor involved in identification of SSL interactions. Additionally, the nature of the method allows inferences as to which kinds of information are most useful in predicting such interactions and may thus sharpen our understanding of the fundamental basis for genetic interaction.

Methods

Collecting and Organizing Gene-Pair Characteristic Data. To predict SSL gene pairs, we identified data types potentially helpful in characterizing SSL interactions. We then used multiple sources (see Table 1 for reference) to determine which yeast gene pairs possessed each characteristic. To construct our decision trees, we used only binary characteristics. Some characteristics, such as colocalization, were inherently binary, whereas continuous characteristics were mapped to several binary characteristics with alternative thresholds. For example, because homology between genes was measured (by BLAST) as a continuous E value, we created three binary characteristics by using BLAST E value thresholds of 10^{-3}, 10^{-6}, and 10^{-12} (13).

Constructing Decision Trees. Decision trees were constructed greedily, beginning with all gene pairs of the training set T in the root node. Gene pairs of each node N were recursively partitioned into two daughter nodes based on the characteristic, which yielded the highest conditional information gain with SSL interaction among gene pairs of node N. Let $Y_c(t)$ be a binary variable indicating whether gene pair t is annotated with characteristic c and X be the random variable indicating whether a gene pair is SSL. When gene pairs in node N were distributed between two nodes N_0 and N_1, where $N_a = \{t \in N, Y_c(t) = a\}$, the conditional information gain was calculated as

$$H_N(X) - \sum_{a=0,1} \frac{|N_a|}{|N|} H_{N_a}(X),$$

where $H_N(X)$ is the entropy of X at node N, defined as

$$-p_N \log(p_N) - (1 - p_N)\log(1 - p_N),$$

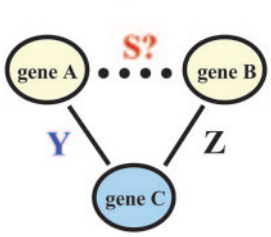
and p_N is the probability that a gene pair $t \in N$ is SSL. To compensate for small sample size, we added one pseudocount distributed in proportion to the fraction of SSL pairs in the entire training set T.

Abbreviations: SSL, synthetic sick or lethal; SGA, synthetic genetic array; MIPS, Munich Information Center for Protein Sequences.

¶To whom correspondence should be addressed. E-mail: fritz_roth@hms.harvard.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0406614101

Table 1. Categories of gene-pair characteristics

Major category		No. of characteristics	Refs.	Appears in trees 1	2	3
Common upstream regulator		3	38			
Gene cooccurrence		1	18–20			
Chromosomal distance		4				
Gene fusion		1	18			
Conserved gene neighborhood		1	16–18			
Physical interaction		15	39–42	x	x	x
mRNA coexpression		17	43, 44			
Same predicted physical complex		1	45			x
Same MIPS function		1	15	x	x	x
Same MIPS protein class		1	15			
Same subcellular localization		42	15	x	x	x
Same phenotype		1	15	x	x	x
Sequence homology		3	13	x	x	x
Mutual clustering coefficient in physical interaction network		16	21			
Posterior probability of physical interaction		4	21			
2hop Y-Z 	2hop H - S	1	2, 7, 13, 15	x	x	x
	2hop P - S	1	2, 7, 15, 39–42	x	x	x
	2hop S - S	1	2, 7, 15	x	x	x
	2hop S - X	1	2, 7, 15, 43, 44	x	x	x
	2hop H - H	1	13			
	2hop H - P	1	13, 39–42			x
	2hop H - R	1	13, 38			x
	2hop H - X	1	13, 43, 44	x		x
	2hop P - P	1	39–42			
	2hop P - R	1	38–42			
	2hop X - X	1	43, 44	x		

Presented are category description; the number of characteristics within each category; reference of data source; whether category is represented in decision tree of crossvalidation (1), predicting experimentally validated gene pairs (2), or predicting new SSL pairs (3). For 2hop characteristics: H, sequence homology; P, physical interaction; R, common upstream regulator; S, synthetic sick or lethal interaction; X, correlated mRNA expression. In the bottom left diagram, Y and Z represent characteristics, S represents synthetic sick or lethal interaction, and A, B, and C represent genes.

To avoid overfitting the training data, we enforced an "early-stopping" criterion based on the Bayesian Information Criterion (14) (asymptotically equivalent to Minimum Description Length). To split a node N into two daughters, we required that the maximal conditional information gain exceed $\log(|T|)/2|N|$. Here, $|N|$ is the number of gene pairs in node N, and $|T|$ is the number of pairs in the entire training set. If the maximal conditional information gain in a node fell below the criterion, the node was not split. Instead, the node became a leaf, or terminal node.

Scoring Leaves and Gene Pairs. Each leaf of a decision tree received a score equal to the fraction of its gene pairs (from the training set) that were SSL. To compensate for small sample size, a total of one pseudocount was added to the number of SSL and non-SSL pairs, distributed in proportion to the fractions of SSL and non-SSL gene pairs, respectively, in the entire training set.

To score our predictions, we mapped each gene pair from the test set to a leaf based on its characteristics. Beginning at the root node, each gene pair was successively assigned to a left or right daughter node based on whether the pair possessed the characteristic used to split the node. Once a gene pair reached a leaf, its mapping was complete, and the pair acquired the score of the leaf. The highest-scoring pairs became our predicted SSL pairs.

Results

Gene-Pair Characteristics. To predict SSL gene pairs, we identified gene-pair characteristics potentially helpful in characterizing SSL interactions. For example, protein products of SSL partners that belong to redundant pathways may share sequence homology, be localized in the same subcellular compartment, and/or belong to the same functional category according to the Munich Information Center for Protein Sequences (MIPS) (15). We also assembled several measures of functional relatedness: conserved gene neighborhood (16, 17), whether pairs of orthologs are chromosomal neighbors in at least two different species; gene fusion (18), whether pairs of orthologs are fused in another genome; gene cooccurrence (19, 20), whether pairs of orthologs have correlated appearance across genomes; and chromosomal distance, whether two genes are located near one another in the *S. cerevisiae* genome. In addition, we included characteristics that describe local network topology around a gene pair, such as mutual clustering coefficient in the physical interaction network (21) and 11 characteristics prefixed by "2hop." Each 2hop characteristic captures specified relationships between a given pair, A–B, and a third gene, C. For example, if protein A physically interacts with protein C, and gene B is SSL with gene C, then the gene pair A–B possesses the characteristic "2hop physical–SSL" (Table 1). As its name implies, 2hop describes a two-step path from A to B through C. We can then ask whether a 2hop physical–SSL relationship is predictive of two genes being SSL, as may be true in compensating pathways (discussed later). We compiled a list of 123 hierarchically organized gene-pair characteristics, falling into 26 major categories (Table 1; for a complete list, see Table 2, which is published as supporting information on the PNAS web site, and for descriptions, see

Supporting Text, which is published as supporting information on the PNAS web site), and then used multiple sources (Table 1) to determine which gene pairs possess each characteristic.

Probabilistic Decision Trees. Probabilistic decision trees are powerful tools for classifying objects and modeling probabilities (22). Here, we use them to model the conditional probability that a gene pair is SSL given a combination of its non-SSL characteristics. Unlike alternative "black box" methods such as neural nets or support vector machines, decision trees can explicitly reveal the characteristics that determine a gene pair's prediction score, and, collectively, these characteristics can suggest biological rationales for the prediction. Furthermore, decision trees do not assume independence between predictive characteristics, as do other methods such as naïve Bayes. Finally, decision trees produce scores that serve to rank predictions according to confidence and have a useful probabilistic interpretation.

To build a decision tree, we first assigned a training set of gene pairs to the root node. Beginning with the root node, we then successively sorted gene pairs in each node into two daughter nodes based on the characteristic deemed most informative of SSL interaction (see *Methods*). If no characteristic was sufficiently informative, a given node was not divided into daughters and further branching was terminated. Thus, each gene pair in the training set was assigned to a single terminal node, or leaf, of the tree. Each leaf then received a score based on its fraction of SSL pairs. To predict the SSL status of a gene pair outside the training set, we mapped the pair to a leaf by its known characteristics, and the pair received the score of that leaf. The highest-scoring pairs became our top predictions. (See *Methods* for further details.) Ultimately, the decision tree served to determine rules that segregated gene pairs by their non-SSL characteristics into subsets enriched in or depleted of SSL pairs.

Assessing Method Performance by Cross-Validation. To assess the performance of our method, we used 4-fold cross-validation on 692,865 SSL-tested gene pairs (5, 11), of which 0.56% (3,868) were SSL [see Table 3, which is published as supporting information on the PNAS web site, an early version of the Tong *et al.* (5) data]. Gene pairs were randomly divided into four groups, and each group was scored by using a decision tree trained on the remaining three. Thus, every gene pair in the data set was scored without regard to its SSL status, and each tree was blind to the SSL status of gene pairs used to assess its predictive capability.

We then assessed performance on only the 692,118 pairs (99.9% of the training set) tested by SGA analysis (5, 11), because we planned to later use SGA analysis to validate predictions. To assess method performance overall, we computed the sensitivity (or true-positive rate, defined here as the fraction of SSL gene pairs correctly predicted) and false-positive rate (defined here as the fraction of non-SSL gene pairs incorrectly predicted to be SSL) at a series of score thresholds. A plot of sensitivity versus false-positive rate at various score thresholds (Fig. 1; see Table 4, which is published as supporting information on the PNAS web site, for all data points) revealed a sensitivity of 80% at a false-positive rate of 18%. This is significantly better than the false-positive rate of 80% expected from random predictions at this sensitivity ($P < 10^{-166}$). Most importantly, our performance suggests that a large-scale screen guided by our method could capture 80% of the SSL interactions by testing <20% of all gene pairs.

By using alternative score thresholds, this approach may be tuned to predict a subset of SSL interactions with higher confidence at the cost of sensitivity. For example, 20% of the interactions were detected at a false-positive rate of 0.2% ($P < 10^{-97}$). This translated to a success rate of 31% (740 SSL interactions in 2,356 predictions), far exceeding the 0.56% success rate expected of an unguided approach. Thus, when

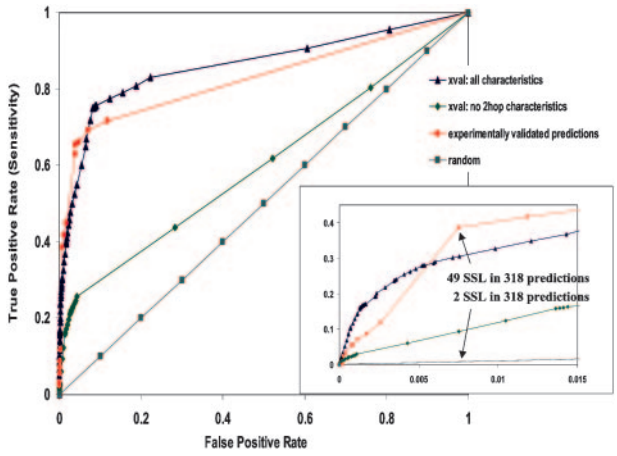


Fig. 1. SSL prediction performance in cross-validation using all gene-pair characteristics (blue triangles), in cross-validation without 2hop characteristics (green diamonds), of experimentally validated predictions using all characteristics (red circles), and performance expected by chance (gray squares).

experimental resources are limited and even a few genetic interactions would be valuable, our method can provide a list of candidate gene pairs that is highly enriched for SSL interaction.

The four trees generated in cross-validation (Fig. 4 *a–d*, which is published as supporting information on the PNAS web site) each contained between 45 and 55 nodes and were structurally similar. The top predictor of SSL pairs was consistently the characteristic 2hop SSL–SSL, in agreement with a previous finding that SSL partners of a gene tend to interact with each other in the genetic network (5). Analogously, dense local clustering in the protein physical interaction network was helpful in predicting physical interaction (21). Another top predictor, 2hop physical–SSL, may indicate compensating pathways in which two gene products, A and C, physically interact in one pathway, whereas a gene, B, belongs to a compensating pathway. When both pathways are impaired (e.g., by mutation of at least one gene from each pathway), the common biological role they can each maintain may be lost, resulting in reduced fitness. Therefore, when genes B and C were SSL and proteins A and C physically interacted, 2hop physical–SSL helped us predict that A and B were SSL (Table 1 diagram). Simultaneously excluding either all 11 2hop descriptors (Fig. 1) or only the four SSL-containing 2hop descriptors of network topology (Fig. 5, which is published as supporting information on the PNAS web site) noticeably decreased performance, further highlighting the importance of network topology information in SSL prediction.

Next we investigated how omission of other characteristics affected performance. We omitted information about localization, function (specifically, the same MIPS function and MIPS protein class), phenotype, or function and phenotype together. Each omission affected performance only mildly (Fig. 5), suggesting that none were critical to our performance, but each improved it slightly.

Experimental Validation of SSL Predictions. Having achieved success in cross-validation, we sought independent experimental validation for our method. We constructed one decision tree using all 692,865 pairs used in cross-validation (Fig. 6, which is published as supporting information on the PNAS web site). Next we scored a test set of 35,996 gene pairs from eight newly performed SGA screens whose query genes were chosen, as in our training set, with a preference for query genes involved in actin-based cell polarity, cell wall biosynthesis, microtubule-based chromosome segregation, or DNA synthesis and repair. The eight query genes

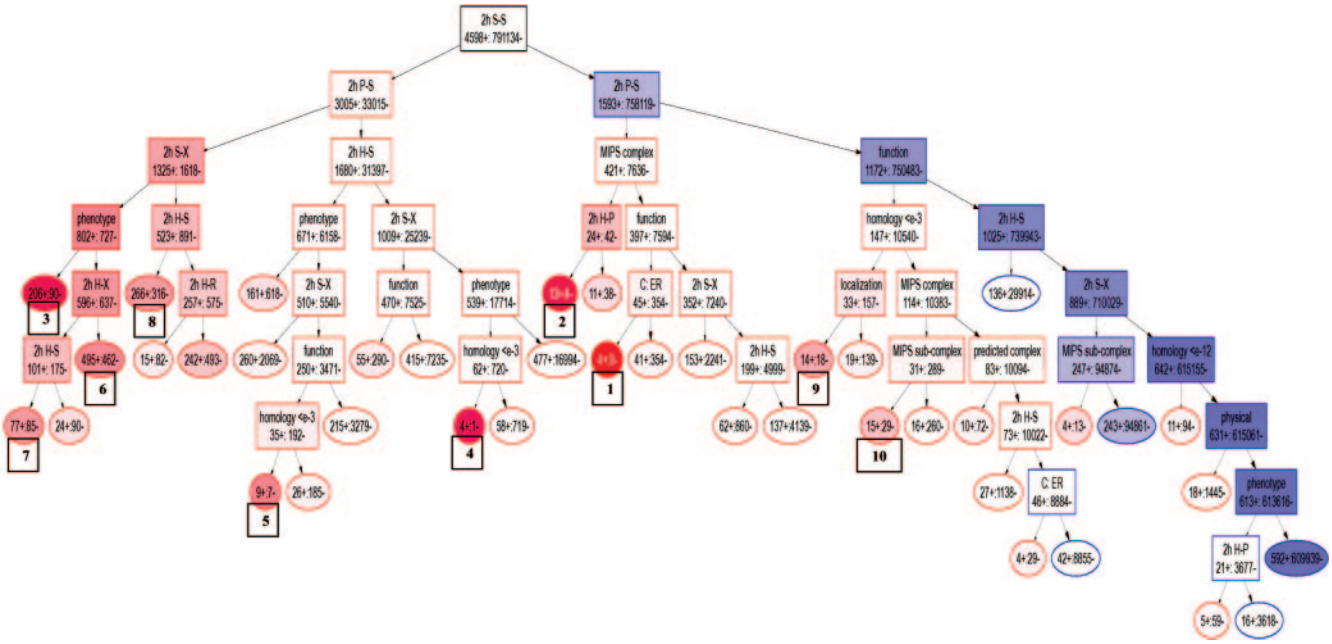


Fig. 2. Tree used to predict new gene pairs. The 10 top-scoring leaf nodes are labeled by rank. Left and right arrows point to gene pairs with and without, respectively, the characteristic that labels the node from which the arrow points. Arrowhead size is proportional to the fraction of gene pairs in the parent node that were assigned to each daughter node. Nodes with higher (lower) fractions of SSL gene pairs than the root are red (blue). Color saturation reflects the entropy with respect to SSL of gene pairs in a node relative to that of the root. Each node is labeled with the number of its gene pairs that are (+) or are not (−) SSL.

were the glycosidases *CWH41* and *ROT2*; glycosyl transferase, *ALG8*; tubulin folding factor D, *CIN1*; the DNA helicases *RRM3* and *HPR5*; ADP-ribosylation factor-like 1, *ARL1*; and *KRE1* involved in cell wall β-glucan assembly (23). Parallel biases in our training and test sets allowed us to demonstrate the capacity of our approach to predict SSL interactions in a case where the training set was representative of the test set. Comparison of our predictions to experimental results revealed a performance similar to that observed in cross-validation (Fig. 1; Table 5, which is published as supporting information on the PNAS web site, lists the data points).

In addition, we correctly predicted interactions more frequently for some of the eight screens than for others (Fig. 7, which is published as supporting information on the PNAS web site), suggesting that performance of our method may vary from gene to gene. Most importantly, though, both validation approaches demonstrated that a subset of SSL interactions could be predicted with high confidence, suggesting that SSL predictions can dramatically reduce the number of gene pairs that must be tested, while maintaining high sensitivity.

Characteristics Most Useful in Predicting SSL Interactions. To identify combinations of gene-pair characteristics predictive of SSL interaction, we trained a decision tree using previously tested pairs and examined characteristics associated with leaves that were enriched for SSL pairs. To assemble our training set, we began with gene pairs systematically tested for SSL interaction by SGA and SGA-associated analyses [using the published version of the SSL data (5, 11)]. To take advantage of non-systematically derived data, we then supplemented our training set with an additional 367 SSL pairs reported in the MIPS database (15). MIPS, however, reports only gene pairs positive for SSL interaction and does not provide negative training examples important to our model. Therefore, to maintain the 0.58% (4,207/728,066) frequency of SSL pairs found by the systematic screens, we also included 67,299 randomly selected gene pairs, treating them as non-SSL (relatively few were expected to be SSL, because the frequency of SSL is low). In total, our training set comprised 795,732 gene pairs, including 4,598 identified SSL interactions involving 1,296 (≈20%) genes. To mitigate the bias in SGA query gene selection, we prohibited our decision tree from using subcellular localization characteristics (e.g., colocalization the nucleus) directly related to the SGA bias, leaving 111 characteristics.

The resulting tree (Fig. 2) comprised 79 nodes and used characteristics from 13 major categories, 10 represented in previous trees and 3 new ones appearing in low-scoring areas of the tree (Table 1). Thus, this new tree used similar characteristics to those used by previous trees.

Each combination of characteristics leading to our top-scoring leaves describes subtypes of known SSL gene pairs and offers insight into mechanisms underlying genetic robustness. Here, we focus on two examples.

Gene pairs in the highest-scoring leaf possess the characteristics 2hop physical–SSL, the same function, and colocalization in the endoplasmic reticulum. For example, the SSL pair, *ALG5* and *WBP1* (Fig. 3*a*), maps to this leaf. Alg5 is a UDP-glucose: dolichyl-phosphate glycosyltransferase, and Wbp1 is a member of the oligosaccharyl transferase glycoprotein complex. The pair's 2hop physical–SSL relationship stems from a physical interaction between Alg5 and Swp1, a subunit of the oligosaccharyl transferase glycoprotein complex, like Wbp1. Both genes belong to the MIPS functional category and protein modification and localize in the endoplasmic reticulum (15).

The most predictive characteristic of this highest-scoring leaf, 2hop physical–SSL, suggests a model in which gene pairs in this leaf respectively belong to two compensating pathways (Fig. 8*a*, which is published as supporting information on the PNAS web site). Two other predictive characteristics associated with this leaf, assignment to the same functional category and the same subcellular location, are consistent with this idea. The tree also indicates that gene pairs mapping to this highest-scoring leaf are not annotated with the 2hop SSL–SSL characteristic, further suggesting involvement of two compensating pathways. In ad-

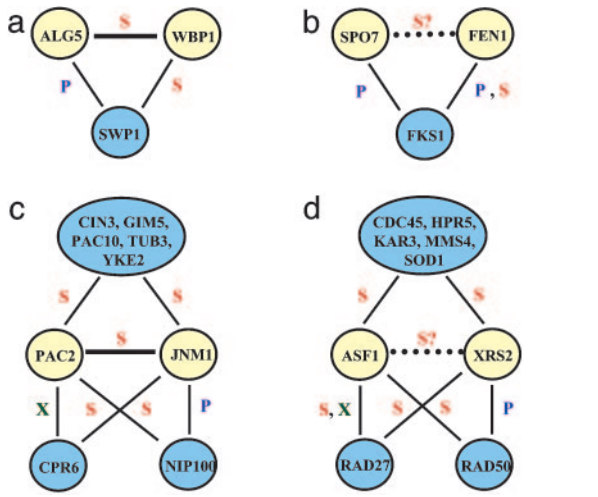


Fig. 3. Gene-pair relationships. (*a* and *b*) Known (*a*) and predicted (*b*) SSL gene pairs from the highest-scoring leaf of the decision tree. (*c* and *d*) Known (*c*) and predicted (*d*) SSL gene pairs from the third-highest-scoring leaf. P, physical interaction; S, synthetic sick or lethal interaction; X, correlated mRNA expression.

dition, protein pairs corresponding to the gene pairs in this leaf are not found in the same complex according to MIPS, suggesting that the compensating pathways do not physically interact via these pairs of genes (although they may do so upstream or downstream).

The third-highest-scoring leaf provides another example in which decision trees suggested informative combinations of characteristics. Gene pairs in this leaf possess the characteristics 2hop SSL–SSL, 2hop physical–SSL, 2hop SSL–coexpression, and the same phenotype. For example, the SSL pair, *PAC2* and *JNM1* (Fig. 3*c*), maps to this leaf. Pac2 is the tubulin-folding cofactor E, and Jnm1 is a coiled-coil domain protein required for proper nuclear migration during mitosis. The pair has five 2hop SSL–SSL relationships involving *YKE2*, *CIN8*, *TUB3*, *PAC10*, and *GIM5*, respectively. Their 2hop physical–SSL relationship is attributed to a physical interaction between Jnm1 and the microtubule-binding protein Nip100, and a SSL interaction between *PAC2* and *NIP100*. The 2hop SSL–coexpression interaction stems from a SSL interaction between *JNM1* and the chaperone-encoding gene *CPR6* and to correlated mRNA expression of *PAC2* and *CPR6*. In addition, *PAC2* and *JNM1* both belong to the MIPS phenotype category, "tubulin cytoskeletal mutants."

One model suggested by the 2hop SSL–SSL characteristic involves three or more compensating pathways for which loss of any two is lethal (Fig. 8*b*). The 2hop physical–SSL and 2hop SSL–coexpression characteristics suggest relationships between the compensating pathways. Consistent with this interpretation, genes paired in this leaf have similar single-mutant phenotypes. This combination of characteristics is also consistent with an alternative model in which proteins encoded by a gene pair are each members of a protein complex for which the loss of either member alone is tolerated, but loss of both is lethal.

These insights are particularly interesting because compensating pathways are difficult to identify and, as a result, have not been well studied. By contrast, duplicate genes, also thought to underlie genetic robustness, are systematically identified by sequence homology and have been actively investigated (1, 24–26). Homologous genes, however, comprise only an estimated 2% of SSL gene pairs (5), suggesting that compensating pathways or other explanations must underlie the majority of SSL interactions. The combinations of characteristics used by decision trees to predict can also identify genetic interactions that arise due to compensatory pathways.

New SSL Predictions. Finally, the decision tree we used above to describe predictive characteristics was also used to generate predictions among all yeast gene pairs potentially testable by SGA (i.e., pairs for which at least one gene was on the SGA array). Table 6, which is published as supporting information on the PNAS web site, lists the 5,000 top-scoring predictions. For example, one of the highest-scoring pairs (mapping to the highest-scoring leaf) is *FEN1* and *SPO7* (Fig. 3*b*). *FEN1* is a long-chain fatty acid elongase. Mutants in *FEN1* exhibit defects in budding and sporulation, likely due to altered membrane phospholipid content (27). *SPO7* is dispensable for mitosis but is required for premeiotic DNA synthesis, a normal mutation rate, recombination, meiosis I and II, glycogen degradation, and sporulation. SSL interaction between *FEN1* and *SPO7* may result from defects in meiosis completion and sporulation. Another high-scoring pair (mapping to the third-highest-scoring leaf) was *ASF1* and *XRS2* (Fig. 3*d*). Asf1 is an antisilencing protein causing derepression of silent loci when overexpressed, and Xrs2 is involved in DNA repair. Validation of these predictions awaits further study.

Because SSL-containing 2hop characteristics were important to our success in cross-validation and experimental validation, we were curious about the performance of our predictions involving genes absent in the SSL training network. In other words, how well could we predict SSL interactions involving genes with no previously known SSL partners? Leaf 9 (Fig. 2) was the highest-scoring leaf that could have generated predictions involving genes without SSL interactions in the training set, because its gene pairs were not required to possess any SSL-containing 2hop characteristics. Specifically, 65% (547/844) of its predictions involved two genes with no SSL interactions in the training set. Next, we checked the SSL status of these 547 pairs in the Yeast Proteome Database (28), which was not consulted in training our model. Surprisingly, 31 (Table 7, which is published as supporting information on the PNAS web site) were annotated as SSL. Unfortunately, we were unable to compute our precise success rate, because the majority of these pairs had not been tested for SSL interaction, and we had no way of determining how many had been tested (pairs tested but found negative for interaction are not reported in available databases). Therefore, our success rate lies between 5.7% (31/547, assuming that all 547 pairs were assessed for SSL interaction) and 100%, with 57% being a reasonable estimate (assuming that 10% of pairs have been assessed for interaction; this is a conservative estimate, considering that the most systematic study to date has tested only ≈3.5% of all gene pairs).

Conclusion

We have demonstrated that it is possible to successfully predict genetic interactions by integrating genomic and proteomic information. Specifically, we predicted SSL gene pairs in *S. cerevisiae* with a success rate such that 80% of the interactions may be discovered by testing <20% of the pairs. In addition, when experimental resources permit only small-scale studies, our method can provide a set of candidate pairs that is highly enriched for SSL interactions.

So what do we know about genetic interactions now? SSL interactions buffer an organism from random mutation. Surprisingly, relatively few (<3%) SSL-interacting genes share sequence homology (5), which likely arises from gene duplication (29). Although, as expected, many share similar Gene Ontology functional categories (5), many may be functionally unrelated (29). Here, we found that the strongest predictors of SSL interaction were the 2hop characteristics (measuring local topology around a gene pair), the same mutant phenotype, physical interaction, and the same function, suggesting that gene pairs with these traits

increase an organism's tolerance for random mutation. Moreover, combinations of characteristics describing SSL interactions help us further classify genetic interactions. We discussed two combinations of characteristics suggestive of gene pairs belonging to compensatory pathways, which are often difficult to identify. We also showed how gene pair characteristics, especially those describing network topology, can help us visualize biological relationships that may not be apparent from genetic interaction screens. Thus, our findings and our methodology offer further insight into genetic robustness and biological networks.

Our prediction approach offers several other applications. Predicted SSL interactions can guide experimental identification of genetic interactions and may be used to infer gene function by predicting that a gene has similar function to its interacting partner(s). This framework may also be useful in predicting SSL interactions involving more than two genes; in predicting other genetic interactions such as epistasis, suppression, or interactions with phenotypes more difficult to score than cell growth; or in predicting other gene-pair characteristics.

Our success in *S. cerevisiae* suggests that genetic interactions may be predicted in higher organisms. Most immediately, in fly and worm, high-throughput phenotype (30–32), protein interaction (33, 34), and mRNA expression (35, 36) data will complement existing yeast data, which can be useful via sequence homology in higher organisms. In addition, high-throughput genetic interaction studies in worm, using RNA interference to simultaneously suppress two genes or to suppress one gene in the background of a germline mutation in a second gene, will provide a training set of SSL interactions from which to build predictive models to guide genetic interaction discovery in higher organisms.

Reliable prediction in higher organisms has strong potential to impact medicine. Because tumor suppressor genes are frequently inactivated in cancer, using a drug to inhibit their SSL partners may selectively impair cancer cells while normal tissue persists (9, 37). Reliable prediction could also dramatically improve the statistical power of multigenic disease mapping. Testing for association between a disease and all possible combinations of genes requires an unreasonable number of tests, severely diminishing the statistical power of association studies (6). SSL prediction allows a candidate gene pair approach, in which only combinations of genes most likely to genetically interact are tested, thereby making multigenic disease mapping feasible using reduced patient populations.

In yeast and in higher organisms, predicting genetic interactions offers biological insight and the potential for medical impact, even though exhaustive assessment of genetic interactions is incomplete and will remain so for years to come.

We thank F. Gibbons and G. Berriz for programming advice and G. Bader for MCODE's predicted protein complexes. S.L.W., L.V.Z., O.D.K., and D.S.G. were supported by fellowships from the Ryan Foundation, the Fu Foundation, the National Human Genome Research Institute, and the National Science Foundation, respectively. This work was also supported by National Institutes of Health/National Human Genome Research Institute, a Howard Hughes Medical Institute institutional grant to Harvard Medical School, the Milton Fund of Harvard University, the Canadian Institute of Health Research (to B.A. and C.B.), Genome Canada (to B.A., C.B., and H.B.), Genome Ontario (to B.A. and C.B.), and Genome Quebec (to H.B.).

1. Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. & Li, W. H. (2003) *Nature* **421,** 63–66.
2. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Smith, J. M. (1997) *Nature* **388,** 167–171.
3. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. (2003) *Nature* **421,** 848–852.
4. Hartman, J. L., Garvik, B. & Hartwell, L. (2001) *Science* **291,** 1001–1004.
5. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G. F., Brost, R. L., Chang, M., *et al*. (2004) *Science* **303,** 808–813.
6. Hoh, J. & Ott, J. (2003) *Nat. Rev. Genet.* **4,** 701–709.
7. Hartwell, L. (2004) *Science* **303,** 774–775.
8. Badano, J. L. & Katsanis, N. (2002) *Nat. Rev. Genet.* **3,** 779–789.
9. Kamb, A. (2003) *J. Theor. Biol.* **223,** 205–213.
10. Ooi, S. L., Shoemaker, D. D. & Boeke, J. D. (2003) *Nat. Genet.* **35,** 277–286.
11. Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., *et al*. (2001) *Science* **294,** 2364–2368.
12. Edwards, J. S. & Palsson, B. O. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 5528–5533.
13. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
14. Schwarz, G. (1978) *Ann. Stat.* **6,** 461–464.
15. Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. & Weil, B. (2002) *Nucleic Acids Res.* **30,** 31–34.
16. Huynen, M., Snel, B., Lathe, W., III, & Bork, P. (2000) *Genome Res.* **10,** 1204–1210.
17. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2896–2901.
18. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. (2002) *Nature* **417,** 399–403.
19. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 4285–4288.
20. Huynen, M. A. & Bork, P. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5849–5856.
21. Goldberg, D. S. & Roth, F. P. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 4372–4376.
22. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984) *Classification and Regression Trees* (Wadsworth International Group, Belmont, CA).
23. Consortium, G. O. (2001) *Genome Res.* **11,** 1425–1433.
24. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., *et al*. (2000) *Science* **287,** 2204–2215.
25. Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P. & Li, W. H. (2002) *Mol. Biol. Evol.* **19,** 256–262.
26. Lynch, M. & Conery, J. S. (2000) *Science* **290,** 1151–1155.
27. Abe, M., Nishida, I., Minemura, M., Qadota, H., Seyama, Y., Watanabe, T. & Ohya, Y. (2001) *J. Biol. Chem.* **276,** 26923–26930.
28. Costanzo, M. C., Crawford, M. E., Hirschman, J. E., Karanz, J. E., Olsen, P., Robertson, L. S., Skrzypek, M. S., Braun, B. R., Hopkins, K. L., Kondu, P., *et al*. (2001) *Nucleic Acids Res.* **29,** 75–79.
29. Wagner, A. (2000) *Nat. Genet.* **24,** 355–361.
30. Kiger, A., Baum, B., Jones, S., Jones, M., Coulson, A., Echeverri, C. & Perrimon, N. (2003) *J. Biol.* **2,** 27.
31. Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., *et al.* (2003) *Nature* **421,** 231–237.
32. Carpenter, A. E. & Sabatini, D. M. (2004) *Nat. Rev. Genet.* **5,** 11–22.
33. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al*. (2003) *Science* **302,** 1727–1736.
34. Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., *et al*. (2004) *Science* **303,** 540–543.
35. Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N. & Davidson, G. S. (2001) *Science* **293,** 2087–2092.
36. Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W. & White, K. P. (2002) *Science* **297,** 2270–2275.
37. Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W. & Friend, S. H. (1997) *Science* **278,** 1064–1068.
38. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al*. (2002) *Science* **298,** 799–804.
39. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 4569–4574.
40. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al*. (2000) *Nature* **403,** 623–627.
41. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., *et al*. (2002) *Nature* **415,** 180–183.
42. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al*. (2002) *Nature* **415,** 141–147.
43. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., *et al*. (1998) *Mol. Cell* **2,** 65–73.
44. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., *et al*. (2000) *Cell* **102,** 109–126.
45. Bader, G. D. & Hogue, C. W. (2003) *BMC Bioinformatics* **4,** 2.