BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i183–i189
DOI: 10.1093/bioinformatics/btg1024

Predicting phenotype from patterns of annotation

Oliver D. King[1], Jeffrey C. Lee[1], Aimée M. Dudley[2], Daniel M. Janse[2], George M. Church[2] and Frederick P. Roth[1,*]

[1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts, 02115, USA and
[2]Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts, 02115, USA

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Predicting the outcome of specific experiments (such as the growth of a particular mutant strain in a particular medium) has the potential to allow researchers to devote resources to experiments with higher expected numbers of 'hits'.

Results: We use decision trees to predict phenotypes associated with *Saccharomyces cerevisiae* genes on the basis of Gene Ontology (GO) functional annotations from the Saccharomyces Genome Database (SGD) and other phenotypic annotations from the Yeast Phenotype Catalog at the Munich Information Center for Protein Sequences (MIPS). We assess the methodology in three ways: (1) we use cross-validation on the phenotypic annotations listed in MIPS, and show ROC curves indicating the tradeoff between true-positive rate and false-positive rate; (2) we do a literature-search for 100 of the predicted gene-phenotype associations that are not listed in MIPS, and find evidence for 43 of them; (3) we use deletion strains to experimentally assess 61 predicted gene-phenotype associations not listed in MIPS; significantly more of these deletion strains show abnormal growth than would be expected by chance.

Contact: fritz_roth@hms.harvard.edu

Supplementary Information: Complete results are available at http://llama.med.harvard.edu/~king/pheno.html

Keywords: decision trees; phenotype; gene function

INTRODUCTION

When an organism's genome has been sequenced and its genes identified, there remains the task of determining the role of each gene in the organism, aspects of which include the gene's molecular function and the phenotypes associated with the gene's disruption. There is an interplay between a gene's functional attributes and phenotypic attributes, with each providing information about the other—see Hampsey (1997) for an overview of *Saccharomyces cerevisiae* phenotypes and their relationships to function.

Efforts to standardize the vocabulary of function and phenotype have facilitated the use of statistical methods to infer function from phenotype and vice-versa. In Clare and King (2002), decision trees were used to extract rules for inferring function from phenotype in *S.cerevisiae*. In this paper we build decision trees for inferring phenotype from function in *S.cerevisiae*. Our approach differs in many details from the approach in Clare and King (2002), but closely follows the approach used in King *et al.* (2003) for predicting functional annotations on the basis of other functional annotations. As genes may have multiple phenotypic annotations, and as there may be informative patterns among these annotations, in this paper we predict phenotype not on the basis of annotated function alone, but on the basis of both function and other phenotypic annotations.

The training data we use consists of the Gene Ontology (GO; The Gene Ontology Consortium, 2000) annotations of function from the Saccharomyces Genome Database (SGD; Cherry *et al.*, 1998), and the annotations of phenotype from the Yeast Phenotype Catalog at the Munich Information Center for Protein Sequences (MIPS; Mewes *et al.*, 2002).

We assess our methodology using three approaches: cross-validation; a literature-search on top-scoring predictions of gene-phenotype associations that are not listed in MIPS; and comparison with a high-throughput experimental determination of phenotype for a comprehensive collection of yeast deletion strains.

METHODS

Training data

We downloaded the MIPS phenotypic annotations from http://mips.gsf.de/proj/yeast/catalogues/phenotype and the SGD GO annotations from http://www.geneontology.org. The versions of the files that we used were downloaded on October 10, 2002.

*To whom correspondence should be addressed.

The 180 MIPS phenotypes are organized as a hierarchy, with more specific phenotypes descending from more general phenotypes. We associated with each phenotype j an indicator random variable X_j, with $X_j(i) = 1$ if gene i is annotated as having phenotype j and $X_j(i) = 0$ otherwise.

The GO attributes are organized as a directed acyclic graph (DAG) (like a hierarchy, but in which attributes may have multiple parents, not just multiple children) consisting of three branches: cellular component, molecular function, and biological process. The GO consists of roughly 13 000 terms, 3051 of which were associated with at least one yeast gene. We excluded annotations for the three attributes 'biological process unknown', 'molecular function unknown', and 'cellular component unknown', and we associated with each of the other 3048 attributes j an indicator random variable X_j, with $X_j(i) = 1$ if gene i is annotated as having attribute j and $X_j(i) = 0$ otherwise.

Note that the files from MIPS and SGD usually contain explicit annotations only at the most specific levels that are supported by the literature. In defining X_j we also include those annotations logically implied by the structure of the MIPS hierarchy and the GO DAG, so that $X_j(i) = 1$ if gene i is explicitly annotated with term j or any descendant of term j in the MIPS hierarchy or GO DAG.

We made predictions for only the 130 most specific phenotypes, although we allowed all 180 phenotypes, as well as the 3048 GO attributes, to be used as predictors. Let $\mathbf{X}_{\sim j}$ denote the vector consisting of all those random variables X_k for which $k \neq j$ and k not an ancestor of j in the MIPS hierarchy, and let $\mathbf{X}_{\sim j}(i)$ denote the vector of values of these random variables for the gene i. For each of the 130 most specific phenotypes j we built a decision tree for predicting X_j from $\mathbf{X}_{\sim j}$, and we used this decision tree to compute

$$q(i, j) = Pr(X_j = 1 | \mathbf{X}_{\sim j} = \mathbf{X}_{\sim j}(i))$$

for each gene i. (Note that when making a prediction about whether a gene has a certain phenotype j, we do not use the ancestors of j in the MIPS hierarchy as predictors.) The score $q(i, j)$ may be interpreted as the probability that a randomly selected gene is annotated as having phenotype j, given that its other annotations, ignoring those for phenotype j and its ancestors in the MIPS hierarchy, agree with those for gene i.

Decision trees

(The presentation in this section is adapted from King *et al.*, 2003.)

See (Breiman *et al.*, 1984) or (Quinlan, 1993) for an overview of decision trees and their applications. For our purposes, the decision tree for phenotype j prescribes a sequence of tests to apply to a gene to aid in predicting whether the gene is annotated as having phenotype j. The tests are all of the form 'Is the gene annotated as having function or phenotype attribute k?' for some GO attribute k or MIPS phenotype $k \neq j$, with k not an ancestor of j in the MIPS hierarchy. Which test is applied depends on the result of previous tests—hence the tree structure.

We constructed the decision tree for phenotype j greedily, by starting with all genes g in the training set in a single root node, and then recursively splitting each node $\mathcal{N}$ by testing on the attribute k for which the *information gain* for phenotype j is maximal.

If we test on attribute k, splitting $\mathcal{N}$ into two nodes $\mathcal{N}_0$ and $\mathcal{N}_1$ where $\mathcal{N}_t = \{g \in \mathcal{N} : X_k(g) = t\}$, then the information gain is defined to be

$$H_{\mathcal{N}}(X_j) - Pr(g \in \mathcal{N}_0 | g \in \mathcal{N})\, H_{\mathcal{N}_0}(X_j) - Pr(g \in \mathcal{N}_1 | g \in \mathcal{N})\, H_{\mathcal{N}_1}(X_j).$$

Here $H_{\mathcal{N}}(X_j)$ is the *entropy* of X_j at node $\mathcal{N}$, which is defined to be $-p_{\mathcal{N}} \log(p_{\mathcal{N}}) - (1-p_{\mathcal{N}}) \log(1-p_{\mathcal{N}})$, where $p_{\mathcal{N}}$ is the probability that a gene $g \in G$ at a node $\mathcal{N}$ is annotated as having j (see e.g. Cover and Thomas, 1991). As in (Niblett and Bratko 1986), we used the estimate

$$p_{\mathcal{N}} = \frac{\#\{g \in \mathcal{N} : X_j(g) = 1\} + m\, p(j)}{\#\{g \in \mathcal{N}\} + m},$$

where $p(j)$ is the fraction of the genes in the entire training set that are annotated as having phenotype j, and m is an adjustable parameter. The term $m\, p(j)$ is used as a *pseudocount*—a small sample-size regularization term, with an interpretation as a prior probability in a Bayesian framework (see e.g. Ewans and Grant, 2001)—with m being the total number of pseudocounts; we set $m = 2$. We used $\#\{g \in \mathcal{N}_t\}/\#\{g \in \mathcal{N}\}$ as an estimate for $Pr(g \in \mathcal{N}_t | g \in \mathcal{N})$ for $t = 0$ and $t = 1$, again following (Niblett and Bratko, 1986).

When no test at a node $\mathcal{N}$ provides a positive information gain, the node is not split, but becomes a leaf. It is labelled with the estimate $p_{\mathcal{N}}$ of the probability that a gene at node $\mathcal{N}$ has phenotype j, as defined above.

A tree grown in this manner will usually overfit the training data, and consequently perform poorly on the held-out test data. A standard way of combating this is to prune away some of the branches after the tree is grown. We used the Bayesian Information Criterion

$$\mathrm{BIC} = -2 \ln Pr(\mathrm{data} \mid \mathrm{model}) + (\ln M) K,$$

which is asymptotically equivalent to the Minimum Description Length (MDL) (Schwartz 1978), for model selection during pruning (see e.g. Friedman and Goldszmidt, 1996). Here K is the number of free parameters in the model (which in our case coincides with the number of leaves in the decision tree), and M is the number of

samples in the data-set (which in our case is the number of genes in the training set). The first term measures the goodness-of-fit of the model to the data, and the second term penalizes model-complexity. We pruned the tree in a bottom-up fashion, starting at the leaves and working toward the root, pruning away any branch whose removal caused the tree's BIC score to decrease. In computing the BIC score we treated the genes as independent, so that the likelihood $Pr(\text{data} \mid \text{model})$ factored as the product of the likelihood for each gene. (This may not be strictly true, due to homology between genes for example.)

The score $q(i, j)$ was then just $p_{\mathcal{N}}$, where $\mathcal{N}$ is the leaf that gene i ends up at in the decision tree for phenotype j.

One notable difference between our approach and the approach used in Clare and King (2002) to predict gene function is that they built a single decision tree to predict what combination of functions a gene has, whereas we build a separate decision tree for predicting each phenotype.

RESULTS

Cross-validation

Let G denote the set of 6898 yeast genes listed in the SGD and T the set of 130 most-specific MIPS yeast phenotypes. We assessed our decision trees using 10-fold cross-validation. We randomly partitioned the set G of yeast genes into 10 sets of equal size (± 1). For each of the 10 sets of genes, we built 130 decision trees using the remaining nine sets (combined) as training data. Then for each gene i in the held-out set, we used these decision trees to compute $q(i, j)$ for each phenotype j in T.

The scores $q(i, j)$ for each of the 10 folds of the cross-validation were pooled together, and for each threshold $t \in [0, 1]$ we computed the true-positive rate

$$TP_t = \frac{\#\{(i, j) \in G \times T : q(i, j) \geq t \;\&\; X_j(i) = 1\}}{\#\{(i, j) \in G \times T : X_j(i) = 1\}}$$

and the false-positive rate

$$FP_t = \frac{\#\{(i, j) \in G \times T : q(i, j) \geq t \;\&\; X_j(i) = 0\}}{\#\{(i, j) \in G \times T : X_j(i) = 0\}}.$$

Figure 1 shows Receiver Operating Characteristic (ROC) curves, plotting TP_t versus FP_t. To demonstrate the value of using MIPS phenotypes combined with GO attributes as predictors, we have included ROC curves for decision trees in which only GO attributes were used as predictors and in which only MIPS phenotypes were used as predictors. We have also included the ROC curve for a model in which all phenotypes and attributes are treated as mutually independent, so that $q(i, j)$ is just the fraction of the genes in the training set that are annotated as having phenotype j. (This gives higher scores for predictions of more common phenotypes, independent of gene.)

Of the $6898 \times 130 = 896\,510$ examples (i, j) in the set $G \times T$, 3029 were positive (i.e. had $X_j(i) = 1$) and the remaining 893 481 were negative. Thus, for example, in the top graph in Figure 1, when using MIPS phenotypes combined with GO attributes as predictors, at the point on the ROC curve where the true-positive rate is 0.1, 303 of the 3029 positive examples are correctly classified as such, and 307 of the 893 481 negative examples are misclassified.

There is the possibility of some circularity in this cross-validation, since certain GO annotations are informed by phenotype annotations. When predicting whether gene i has phenotype j, we do not look at whether gene i is annotated as having phenotype j or any of the ancestors of phenotype j in the MIPS hierarchy, since annotations for these ancestral phenotypes might be derived from an annotation for phenotype j, which we are trying to predict. Similarly, there might be GO annotations for gene i that are derived from an annotation for phenotype j, and using these as predictors for phenotype j could give artificially inflated performance.

SGD curators use ten evidence codes when assigning GO annotations (see http://www.geneontology.org/GO.evidence.html). Phenotype-derived SGD GO annotations should usually have evidence code IMP (inferred from mutant phenotype), TAS (traceable author statement), NAS (non-traceable author statement), IC (inferred by curator), or NR (not recorded), as opposed to other codes such as IPI (inferred from physical interaction). To reduce the influence of phenotype-derived GO annotations, we re-ran the cross-validation, this time making predictions only for the 4835 genes with no annotations of type IMP, TAS, NAS or NR. (This left 1318 positive examples and 627 232 negative examples.) The resulting ROC curves (the lower two graphs in Figure 1) show only a slight reduction in prediction performance. (Note also that in the top graph, circularity is not an issue when we use only MIPS phenotypes as predictors.)

Literature-based assessment

While the cross-validation performed above demonstrates that gene-phenotype associations may often be predicted accurately on the basis of other annotations, this would be of little use if all of the phenotypes associated with each gene were already known. But this is almost certainly not the case, and we operate under the premise that those genes i and phenotypes j for which $q(i, j)$ is large are good candidates for being genuinely associated, even if the association is not listed in MIPS. While the present method is not the ideal way to deal with missing phenotypic data, as discussed in King *et al.* (2003) it has the virtue of being computationally tractable, and is formally identical to a treatment of missing data that has

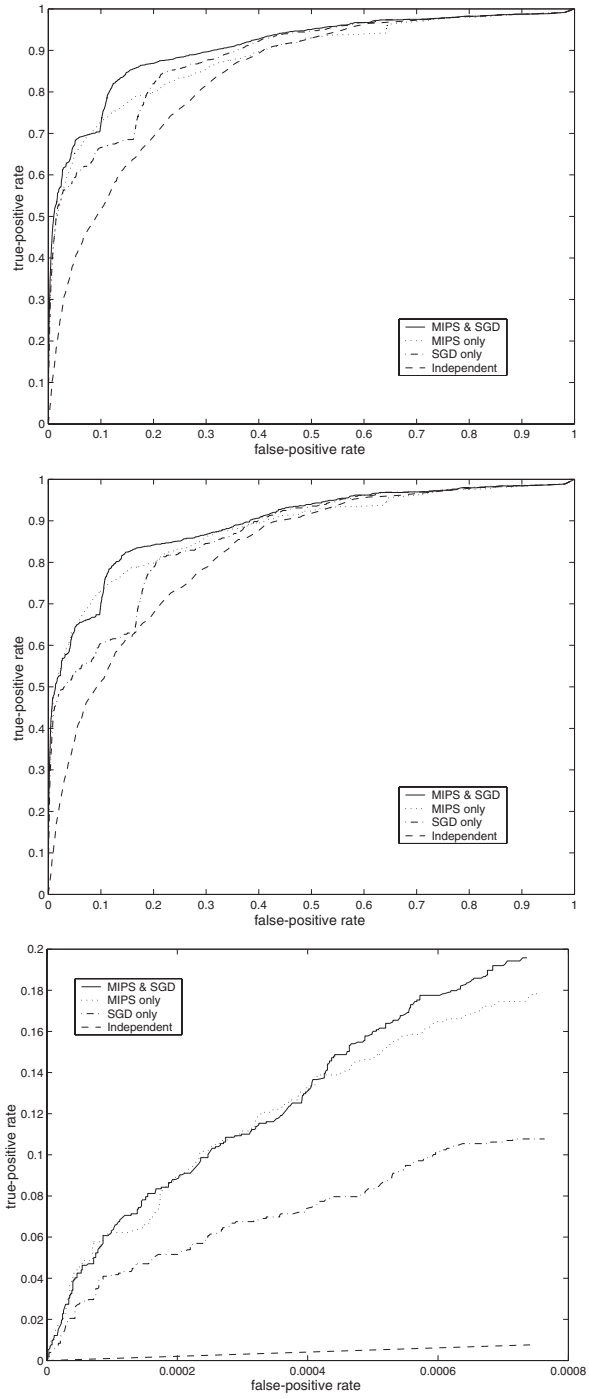

Fig. 1. ROC curves for cross-validation of phenotype predictions (1) using both MIPS annotations and SGD GO annotations as predictors; (2) using just MIPS annotations as predictors; (3) using just SGD GO annotations as predictors; and (4) modeling all MIPS phenotypes and GO attributes as independent. In the top figure we use all genes for testing, and in the middle figure we use only those genes with no SGD GO annotations of type IMP, TAS, NAS, IC or NR, to avoid circularity. The bottom figure is a detail of the middle figure at low false-positive rates, with the axes rescaled.

been used profitably in collaborative filtering applications (see e.g. Breese *et al.*, 1998).

To test the approach, in the process of doing the cross-validation we also compiled a list of the gene-phenotype pairs $(i, j) \in G \times T$ for which $X_j(i) = 0$ but $q(i, j) > 0.5$. There were 542 such pairs, 160 of which were for 'other' phenotypes (such as 'other DNA replication mutant' and 'other cytoskeleton mutants') and were removed. Of the remaining 382 pairs (i, j), we looked at the 100 with the highest scores $q(i, j)$ and manually assessed the plausibility of gene i being associated with phenotype j by looking up gene i in the Yeast Proteome Database (Costanzo *et al.*, 2001; http://www.incyte.com/sequence/proteome/databases/TPD.shtml), following up in MEDLINE when warranted.

We used a four-level rating scheme. The highest-scoring 50 pairs (i, j) and their ratings are given in Table 1—the full list of 100 ratings, along with other supplementary data, is available at http://llama.med.harvard.edu/~king/pheno.html; below we summarize the number of gene-phenotype pairs that received each rating:

Rating	Number
(1) Null mutant has phenotype	17
(2) (Non-null) mutant has phenotype	26
(3) No decisive evidence	50
(4) Contradictory evidence	7

These results indicate that our success rate for these 100 predictions was at least 43% and at most 93%.

Experimental assessment

We also experimentally assessed 61 of the predicted gene-phenotype associations not listed in MIPS. These were not the 61 predictions with the highest q scores, but were those predictions with $q(i, j) > 0.5$ and with phenotype j among eleven phenotypes (listed in Table 2) with available high-throughput assay results. These eleven phenotypes were tested in 4710 yeast deletion mutants (Winzeler *et al.*, 1999; Giaever *et al.*, 2002) (Research Genetics), based on growth on solid agar media in various conditions. Further details on the phenotype screens, including the quantitation and normalization of mutant growth, will be given in a separate publication (Dudley, Janse, and Church; manuscript in preparation), but are available upon request. We briefly describe the methodology below.

All phenotypes were examined in a homozygous diploid background (BY4743). With the exception of YPG, which contained 3% glycerol as the sole carbon source, all media were prepared as YPD (Rose *et al.*, 1990) containing 2% glucose with the concentration of drugs or chemicals as

Table 1. List of the 50 highest-scoring gene-phenotype pairs (among those not already listed in MIPS) and their ratings based on Yeast Proteome Database and MEDLINE searches

Gene name or ORF	MIPS phenotype	Rating
CDC10	Actin cytoskeleton mutants	2
PPH22	Actin cytoskeleton mutants	2
HKR1	Actin cytoskeleton mutants	3
SEC3	Actin cytoskeleton mutants	3
TWF1	Actin cytoskeleton mutants	4
MHP1	Actin cytoskeleton mutants	3
BUD4	Actin cytoskeleton mutants	3
CKA2	Actin cytoskeleton mutants	3
BEM4	Actin cytoskeleton mutants	3
MSH1	Alkylating agents sensitivity	3
RAX2	Bud localization	1
RAX1	Bud localization	1
PKC1	Calcofluor white sensitivity	2
LRE1	Calcofluor white sensitivity	3
PSA1	Calcofluor white sensitivity	3
OST4	Calcofluor white sensitivity	3
MSS4	Calcofluor white sensitivity	4
HKR1	Calcofluor white sensitivity	3
YGR166W	Calcofluor white sensitivity	3
YGR229C	Calcofluor white sensitivity	1
SLT2	Calcofluor white sensitivity	1
SET1	Calcofluor white sensitivity	3
LAS21	Calcofluor white sensitivity	2
FPS1	Calcofluor white sensitivity	3
RPL10	Calcofluor white sensitivity	3
BNI4	Calcofluor white sensitivity	3
YNL322C	Calcofluor white sensitivity	3
PFY1	Calcofluor white sensitivity	3
HHT1	Heat-sensitivity (ts)	3
MSS4	Heat-sensitivity (ts)	2
CDH1	Heat-sensitivity (ts)	3
YGR099W	Heat-sensitivity (ts)	3
CAP2	Heat-sensitivity (ts)	3
DOM34	Heat-sensitivity (ts)	3
SSN3	Heat-sensitivity (ts)	2
ECM15	hygromycin B sensitivity	2
ECM2	hygromycin B sensitivity	2
ECM33	hygromycin B sensitivity	2
ECM31	hygromycin B sensitivity	2
ECM11	hygromycin B sensitivity	2
ECM29	Hygromycin B sensitivity	4
ECM14	Hygromycin B sensitivity	2
YMR308C	Nuclear mutants	2
CBS1	Respiratory deficiency	1
CBS2	Respiratory deficiency	1
COX20	Respiratory deficiency	1
DNM1	Respiratory deficiency	1
MDJ2	Respiratory deficiency	4
TFB3	UV light sensitivity	2
TFB2	UV light sensitivity	2

A rating of 1 means that evidence was found that null mutants have the phenotype; 2 means that evidence was found that (non-null) mutants have the phenotype; 3 means that no decisive evidence was found; and 4 means that contradictory evidence (i.e. evidence that mutants do not have the phenotype) was found. The list is sorted alphabetically by phenotype

indicated in Table 2, following (Hampsey, 1997). UV sensitivity was measured by growth on YPD following UV irradiation with 100 Joules/m^2. All growth measurements were repeated twice using fresh strains from a frozen stock. Mutant growth on each medium was quantitated by image analysis of agar plate images using GenePix 4.0 (Axon Instruments). The growth of each mutant in each condition was normalized to growth in the control condition (YPD) by computational comparison of the image analysis results files. The normalized scores took five possible values $(-2, -1, 0, 1, 2)$, with negative scores indicating decreased growth in the condition tested (sensitivity) and positive scores indicating increased growth in the condition tested (resistance).

In the following analysis, we use a fairly relaxed standard of evidence and consider a predicted gene-phenotype association to be validated if either of the two replicates of a screen for sensitivity has a negative score, or if either of the two replicates of a screen for resistance has a positive score. Using this criterion, 12 of the 61 predictions we tested were validated.

To confidently conclude that a gene is associated with a phenotype, one would want to use more stringent criteria and additional controls as in Bianchi *et al.* (2001), but we account for the relaxed standard when assessing the statistical significance of our results:

Let t_k denote the number of predictions tested for the kth phenotype ($k = 1, \ldots, 11$), let v_k denote the number of predictions validated from among these t_k predictions, and let u_k denote the total number of the 4710 deletion mutants that displayed the kth phenotype using our relaxed standard of evidence. (The values of t_k, v_k and u_k are listed in Table 2.) Suppose we were to exchange the t_k genes we predicted to be associated with the kth phenotype with t_k genes selected at random (from among the 4710 genes tested for phenotype k) for each k, to get a new set of 61 predictions. Then by the linearity of expectations, the expected number of these random predictions that would be validated is $\sum_{k=1}^{11} t_k u_k/4710 = 2.6$.

The probability that exactly i of the t_k random predictions for the k-th phenotype are validated is $C(u_k, i)\, C(4710 - u_k, t_k - i)/C(4710, t_k)$, where $C(r, s) = r!/(s!(r-s)!)$ is the binomial coefficient. (The number of validated predictions for the k-th phenotype follows a hypergeometric distribution, as we are picking t_k genes without replacement.) We computed a p-value—the probability that 12 or more of these random predictions are validated—via a Monte Carlo simulation, by summing samples from the eleven appropriate hypergeometric distributions ten million times; the p-value is 9×10^{-6}.

It should be noted that this p-value is conservative, since we are assessing predictions only for gene-phenotype associations not listed in MIPS, but our totals u_k include those associations that are listed in MIPS. If we adjust for

Table 2. List of the eleven phenotypes assayed experimentally

MIPS phenotype	Assay	t	v	u
Benomyl sensitivity	15 μg/ml benomyl	9	0	125
Caffeine sensitivity	2 mg/ml caffeine	5	2	303
Cycloheximide resistance	0.18 μg/ml cycloheximide	1	0	91
Cycloheximide sensitivity	0.18 μg/ml cycloheximide	1	0	260
Divalent cations and heavy metals sensitivity	0.7 M $CaCl_2$	11	2	224
Hygromycin B sensitivity	50 μg/ml hygromycin B	6	0	361
Osmotic sensitivity (Osm)	1.2 M sorbitol	4	0	40
Rapamycin resistance	0.1 μg/ml rapamycin	1	0	113
Rapamycin sensitivity	0.1 μg/ml rapamycin	2	0	166
Respiratory deficiency	YPG (3% glycerol)	12	6	275
UV light sensitivity	100 Joules/m^2 UV	9	2	95

The column headed t gives the number of predictions we tested for the phenotype, the column headed v gives the number of these predictions that were validated, and the column headed u gives the total number of the 4710 deletion mutants screened that displayed the phenotype

this, the expected number of random predictions validated drops to 2.4, and our p-value drops to 3×10^{-6}.

It should also be noted that here we are just testing null mutants. Since a prediction by our approach is a prediction that *some* allele of gene i has phenotype j, our estimate of success using assays of null mutants is conservative in the sense that some predictions not confirmed by assay of a null allele might be confirmed with some other allele.

CONCLUSIONS

We have demonstrated using cross-validation that our models can be useful for predicting gene-phenotype associations already listed in MIPS. We have also demonstrated that they are useful for predicting gene-phenotype associations not listed in MIPS: over 40% of the top 100 predictions for associations not listed in MIPS were supported by a literature-search, and high-throughput experimental phenotype assays using deletion strains were successful significantly more often than would be expected by chance.

One might ask 'What is the use of predicting yeast phenotypes, given that deletion strains are readily available and that the phenotype assays described here are relatively straightforward and are being performed in high-throughput for all genes?' We expect that phenotype prediction will be particularly useful for organisms in which mutant strains are less available and for phenotypes that are more difficult to assay (e.g. the inattentive mother phenotype in mouse Brown *et al.*, 1996).

ACKNOWLEDGMENTS

The authors would like to thank G.Berriz for programming assistance, F.Winston, Z.Moqtaderi, and S.Buratowski for helpful discussions, and the anonymous referees for their suggestions. This research was sponsored in part by Aventis Pharmaceuticals, and by an institutional grant from the HHMI Biomedical Research Support Program for Medical Schools. O.D.K. was supported by an NRSA Fellowship from NHGRI. J.C.L. was supported by an NIH training grant. G.M.C., A.M.D. and D.M.J. were supported by grants from the DOE.

REFERENCES

Bianchi,M.M., Ngo,S., Vandenbol,M., Sartori,G., Morlupi,A., Ricci,C. *et al.* (2001) Large-scale phenotypic analysis reveals identical contributions to cell functions of known and unknown yeast genes. *Yeast*, **18**, 1397–1412.

Breese,J., Heckerman,D. and Kadie,C. (1998) Empirical analysis of predictive algorithms for collaborative filtering. *Microsoft Research Technical Report MSR-TR-98-12*.

Breiman,L., Friedman,J.H., Olsen,R.A. and Stone,C.J. (1984) *Classification and Regression Trees*. Chapman and Hall, New York.

Brown,J.R., Ye,H., Bronson,R.T., Dikkes,P. and Greenberg,M.E. (1996) A defect in nurturing in mice lacking the immediate early gene fosB. *Cell*, **86**, 297–309.

Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Clare,A. and King,R.D. (2002) Machine learning of functional class from phenotype data. *Bioinformatics*, **18**, 160–166.

Costanzo,M.C., Crawford,M.E., Hirschman,J.E., Kranz,J.E., Olsen,P., Robertson,L.S. *et al.* (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, New York.

Ewans,W.J. and Grant,G.R. (2001) *Statistical Methods in Bioinformatics: An Introduction*. Springer, New York.

Friedman,N. and Goldszmidt,M. (1996) Learning Bayesian Networks with Local Structure. *Proceedings of 12th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, pp. 252–262.

The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Véronneau,S. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

Hampsey,M. (1997) A Review of Phenotypes in *Saccharomyces cerevisiae*. *Yeast*, **13**, 1099–1133.

King,O.D., Foulger,R.E., Dwight,S.S., White,J.V. and Roth,F.P. (2003) Predicting gene function from patterns of annotation. *Genome Res.*, in press.

Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Niblett,T. and Bratko,I. (1986) *Learning Decision Rules in Noisy Domains*, Developments in Expert Systems, Bramer,M. (ed.), Cambridge University Press, pp. 25–34.

Quinlan,J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Rose,M.D., Winston,F. and Hieter,P. (1990) *Methods in Yeast Genetics: A Laboratory Course Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schwartz,G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B. *et al.* (1999) Functional characterization of the *Saccharomyces cerevisiae* genome by precise deletion and parallel analysis. *Science*, **285**, 901–906.