Prospectus: Using Molecular Descriptors and Functional Genomics Data to Predict the Effect of Small Molecules on *S. cerevisiae*

Tara Gianoulis
Computational Biology and Bioinformatics

Specific Aims

One of the major unresolved questions in small molecule design is how a protein "sees" a molecule (Zanders, Bailey et al. 2002). In other words, how is it that two "chemically similar" compounds can have different targets? The simplest explanation would be that we have not yet discovered what chemically similar actually means in the context of a biological system. By characterizing small molecules both in terms of their phenotypic profile and in terms of their molecular descriptors, the goal of my project is to examine the relationship between chemical similarity and phenotypic effect. I will determine whether the structural features of a compound can be used to predict its effect on living organisms by combining molecular similarity analysis, functional genomics data, and phenotype mapping in *S. cerevisiae*. Specifically, I will:

(1) Assemble available datasets of functional genomics data for *S. cerevisiae*

a. Collect large and small scale phenotypic data.

b. Integrate functional genomics data.

c. Calculate molecular descriptors for small molecules.

(2) Clustering, Feature Selection, and Prediction

a. Determine phenotypic signature and molecular fingerprint via principal component analysis and hierarchical clustering.

b. Implement genetic algorithm to extract minimal set of molecular descriptors.

c. Construct decision tree on minimal sets found in b.

(3) Perform chemogenomic profiling on several uncharacterized compounds.

Introduction and Significance

Small molecule compounds have been used both as potent therapeutic agents and as chemical inhibitors to probe protein function. These molecules typ ically are identified using either screens of natural products or combinatorial chemical libraries (Stockwell 2000). Despite the development of a number of new technologies to address issues involved in high throughput screening, target validation remains a large bottleneck. Predicting the biological effect of a small molecule using structural characteristics would winnow the amount of experimental effort required and thus be a very valuable tool.

I will classify small molecules using molecular descriptors and combine this with phenotypic information about known small molecules to develop a prediction method which I will apply to uncharacterized small molecules. To establish this method, I will use the following tools: molecular descriptors, genetic approaches, functional genomics information, and feature selection algorithms.

Molecular Descriptors

Molecular similarity analysis (MSA) quantifies the similarities between compounds. Classification criteria can be simple descriptors, such as the types of atoms and their connections. Previous work categorized 5120 known drugs, and half of these were classified by just 32 “molecular frameworks” based on simple descriptors (Bemis and Murcko 1996). More complex characteristics, such as, shape, electrotopological states, and other electronic interactions can provide additional descriptors of small molecules (Mannhold, Kubinyi et al. 2000). By looking at features of both the substructures and the overall molecule, a quantitative picture of the small molecule can be developed.

Genetic Approaches

There are four basic types of genetic approaches: forward genetic, reverse genetic, forward chemical genetic, and reverse chemical genetic. Classically, forward genetics was used to identify the gene causing a particular phenotype; whereas, in reverse genetics, the gene is known; but the phenotype is not. Both types of methods can be implemented using small molecules. Traditionally these tools relied on genetic mutants, in chemical genetics, small molecules can mimic genetic mutations by specifically inhibiting a particular gene product (reverse) or serving as the phenotype of interest (forward) (Stockwell 2000). The advantage of using small molecules is that this approach generates "conditional" and reversible mutants.

Regardless of whether the pure genetic or chemical genetic approach is pursued, target identification, that is linking a phenotype to a particular genotype, remains a significant hurdle. Failure to achieve mutant saturation, accumulation of suppressor or enhancer mutations, and a variety of other problems can obscure the "true" cause of a particular phenotype (Winzeler, Shoemaker et al. 1999). A major breakthrough in target identification came as a result of the construction of the yeast knockout library (Giaever, Chu et al. 2002). As mentioned above, the use of mutagenic agents to generate mutants introduces multiple mutations making disentangling the cause and effect of individual genetic changes both difficult and error-prone. By constructing deletion strains for every gene in the *S. cerevisiae* genome, a number of these issues have been ameliorated.

In addition, since the deleted gene was replaced with a cassette flanked by two unique twenty nucleotide sequences, one upstream and one downstream, each strain in essence is barcoded. This unique feature allows the strains to be grown together in competitive growth assays. Barcode DNA can be PCR amplified, hybridized, and labeled to a microarray chip containing the complementary barcodes (Giaever, Chu et al. 2002). This allows for a massively parallel phenotypic assay, where the effect of a particular condition on every strain can be seen in terms of differential growth rate. In a second variation, rather than using the barcodes, automated pinning machines spot the entire collection on plates under a variety of different conditions, and differential growth rates can be measured through either automated or manual colony scoring (Winzeler, Shoemaker et al. 1999). These two approaches have been used in a variety of different studies from identifying genes involved in toxicity (Birrell, Giaever et al. 2001), to examining genes affected by UV sensitivity (Birrell, Giaever et al. 2001), to genomewide identification of human disease genes (Steinmetz, Scharfe et al. 2002) . The datasets that will be included in my project are limited to those involving the use of small molecules. This particular application of the yeast knockout library is commonly referred to as chemogenomic profiling.

Chemogenomic Profiling

In chemogenomic profiling, all of the yeast deletion strains can be pooled together with one set being treated with the drug and the second in wild type conditions, the barcode DNA can then be PCR amplified, labeled, and hybridized to a microarray that contains the complements of the barcodes. Those strains that drop out most quickly when grown in media containing compound compared to those grown in rich media are deemed most sensitive to the drug. Alternatively, the strains can be plated on compound containing medium, and the relative growth rate between deletions grown on wild type and on compound containing plate can be compared.

Mechanism of Action and Drug Target Identification

How can the information obtained from chemogenomic profiling experiments be levied to ascertain the drug's mechanism of action or its target? One approach is to determine drug induced haploinsufficiency. Haploinsufficiency is a phenomenon where having only one copy of the gene, as is the case for heterozygous deletions, results in an abnormal phenotype when compared to wild type. Additionally, it was discovered more than twenty years ago that increasing gene copy number that encodes for drug resistance not unsurprisingly leads to an increase in drug resistance (Rine, Hansen et al. 1983). The main assumption is that the increase in gene copy number leads to a corresponding increase in synthesis of the gene product, and this overexpression results in the increase in resistance (Launhardt, Hinnen et al. 1998). If increasing the gene copy number of the drug target leads to drug resistance, would decreasing gene copy number result in drug sensitivity? In a study using 233 heterozygous (one wild type copy) deletion strains, Giaever, et al. tested this hypothesis. They found that those deletion strains which grew more slowly in sublethal concentration of drug with respect to those grown under wild type conditions were in several cases the known target of the drug (Giaever, Shoemaker et al. 1999). With the completion of the yeast knockout library, the experiment was extended to include the entire genomewide collection. Nine different small molecules were profiled in this manner, and by using the differential growth rate of the strains, several compounds with previously unknown targets were characterized (Giaever, Flaherty et al. 2004).

Concurrently, a study using a subset of 3500 of the deletion strains profiled 78 compounds (Lum, Armour et al. 2004). These 78 compounds were divided into three groups according to the number of strains that were sensitive to the given compound. Group I compounds had no changes in growth rates among any of the strains; group II triggered growth defects in just a few strains, and group III compounds caused growth defects in a large number of deletion strains. The authors concluded that the heterogeneity of responses could be used to determine the mechanism of action for a particular compound (Lum, Armour et al. 2004). They postulated that since group I compounds triggered no growth defects, these drugs may not have a protein target. For example, it has been established that cisplatin targets DNA rather than a particular protein. The chemogenomic data showed that there were no sensitive strains identified for cisplatin (Lum, Armour et al. 2004). Group III compounds could represent those compounds that affect general processes, for example transporters. The group II compounds are the most likely compounds to have one or just a few protein targets. By examining the number of strains that exhibit fitness defects, the subset of drugs that have particular protein targets can be separated from those using a different mechanism of action.

One caveat is that at the time of the study the entire genomewide collection was not available, so it is possible that the subset used may have particular biases, but overall it is expected that the trend observed should be preserved. In total, 100 small molecules have been profiled either using the bar tag method or a similarly conducted plate assay. These studies provide a wealth of data that I plan to use in training and building my predictive model.

Although chemogenomic profiling has resulted in a tremendous leap in the ability to link phenotype to genotype, identifying the specific target of a particular small molecule using this data remains problematic for three reasons. A careful accounting of these discrepancies will allow me to reanalyze the data in a potentially less noisy manner. The first reason, as described above, is that not all drugs act on a particular protein. This limitation can be easily sidestepped by filtering drugs where there are no strains that exhibit drug sensitivity(Lum, Armour et al. 2004). Two additional complications are nonspecific drug effects and genetic interactions which can obscure the “true” target of the small molecule. Although a number of different studies have been performed, each addresses only one of these issues, and there has yet to be a systematic analysis incorporating all three. In Table 1, I have detailed the possible outcomes of a chemogenomic profiling experiment, as well as some interpretations.

		Cmpd1	Cmpd 2
		Inhibits	inhibited by
	Strains	Prot A	Prot A
Prot A			
Ess in	A/A	Lethal	No Defect
YPD	?a/A	Lethal	Some Defect
Prot A	A/A	No Defect	No Defect
NONess	?a/A	No Defect	Some Defect
in YPD	?a/?a	No Defect	Lethal
A is	A/A B/B	No Defect	No Defect
S Lethal	A/A ?b/B	Some Defect	No Defect
w/B	A/A ?b/?b	Lethal	No Defect

Table 1 Expected Outcomes and Interpretation of Chemogenomic Profiling Experiment

For simplicity, imagine a drug and a protein can have only one of two possible interactions. The drug can either specifically inhibit a protein, or the drug itself can be toxic, and the protein sequesters or neutralizes the drug preventing the toxicity. One precaution then is that the most sensitive strain to a toxic drug does not necessarily represent the drug target. How can one distinguish between when the most sensitive strain is the actual drug target and when it is not? One solution is to use the synthetic lethal interactions (Parsons, Brost et al. 2004). In the case of synthe tic lethality, the strain missing gene A or gene B remains viable, but the deletion of both gene A and gene B is lethal. If compound 1 inhibits protein A, and gene A (encodes protein A) and gene B have a synthetic lethal interaction, then the knockout of B in the presence of compound 1 should result in lethality; therefore, understanding the mechanism behind synthetic lethality is critically important in accurately interpreting the results of a chemogenomic profiling experiment.

Synthetic Lethality and Target Identification

Two approaches, synthetic genetic array screen (SGA) (Tong, Evangelista et al. 2001) and diploid-based synthetic lethality analysis on microarrays (dSLAM) (Pan, Yuan et al. 2004), have been developed to perform high throughput screens for synthetic lethal interactions. Using the SGA screen, it was found that there are on average 34 genetic interactions per gene, making the putative genetic interaction network eight times as dense as the protein-protein interaction network (Tong, Lesage et al. 2004) . Synthetic lethal interactions can indicate that two genes are components of the same complex, that they function in the same pathway, or even that they function in separate pathways, but share the same end goal (Sharom, Bellows et al. 2004).

The second method dSLAM has proven to be an extremely versatile tool. Mapping of synthetic lethal interactions was performed using both the classical genetic approach of crossing two different deletion strains and selecting for double deletion transformants, as well as, using a chemical genetic approach using small molecules to generate a "chemical knockout" (Pan, Yuan et al. 2004). The basic assumption is that if a knockout of gene A in the presence of compound that inhibits protein B results in the yeast dying, then A has a synthetic lethal interaction with B. The higher the resolution of genetic mapping, the greater confidence one can have in the results of the chemogenomic profiling identifying the drug target.

Thirdly, a drug can only act on its target if it can reach its targets. Those strains defective in transporters , efflux pumps, and lipid biosynthesis will be sensitive to a large number of compounds not because the compound physically binds any of them, but rather because these defects increase drug accessibility (Bauer, Wolfger et al. 1999; DeRisi, van den Hazel et al. 2000; Mukhopadhyay, Kohli et al. 2002) . A number of genes have been implicated in multidrug resistance and general nonspecific response to small molecules (DeRisi, van den Hazel et al. 2000; Parsons, Brost et al. 2004). Lastly, there are some problems with the deletion collection including missing strains, incorrectly annotated strains, and development of secondary mutations (Grunenfelder and Winzeler 2002). In addition, after sequencing the barcodes, it was found that 31% of the tags differed by at least one base pair from what was believed to have been introduced (Eason, Pourmand et al. 2004). It was shown, however, that in most cases, these differences did not significantly alter hybridization. Despite these limitations, the yeast knockout library is still orders of magnitude cleaner and more robust than the random mutagenesis methods.

Notwithstanding major advances in both biology and chemistry, target identification remains a tedious process. By examining the relationship between molecular descriptors, processed phenotypic data, and functional genomics, I plan to develop machine learning algorithms to predict subsets of potential targets and to provide new insight in the linkage of genotype and phenotype.

RESEARCH PLAN

Specific Aim 1: Compilation of Available Datasets and Molecular Descriptor Calculation

Assembling Available Datasets

Since the sequencing of the S cerevisiae genome in 1996 (Goffeau, Barrell et al. 1996), a large number of genomewide studies have been conducted. The first step of my project will be to assemble these datasets. Large scale data exists for subcellular localization (Kumar, Agarwal et al. 2002; Huh, Falvo et al. 2003), essentiality (Winzeler, Shoemaker et al. 1999), conservation with close yeast (Kurtzman and Robnett 2003), protein-protein interactions (Ito, Tashiro et al. 2000; Schwikowski, Uetz et al. 2000; Ito, Chiba et al. 2001), genetic interactions (Tong, Lesage et al. 2004), and a host of problem because they are (1) capable of efficiently exploring high dimensional search spaces and ??(2) able to incorporate interactions not just individual features ADDIN

Num>20</RecNum><record><rec-number>20</rec-number><ref-type name="Journal Article">17</ref-type><contributors><authors><author>Hughes, T. R.</author><author>Marton, M. J.</author><author>Jones, A. R.</author><author>Roberts, C. J.</author><author>Stoughton, R.</author><author>Armour, C. D.</author><author>Bennett, H. A.</author><author>Coffey, E.</author><author>Dai, H.</author><author>He, Y. D.</author><author>Kidd, M. J.</author><author>King, A. M.</author><author>Meyer, M. R.</author><author>Slade, D.</author><author>Lum, P. Y.</author><author>Stepaniants, S. B.</author><author>Shoemaker, D. D.</author><author>Gachotte, D.</author><author>Chakraburtty, K.</author><author>Simon, J.</author><author>Bard, M.</author><author>Friend, S. H.</author></authors></contributors><auth-address>Rosetta Inpharmatics, Inc.,

Preprocessing Phenotypic Data: An Extension of Marginal Essentiality

Since I am conscious of the limitations of the phenotypic data, I can incorporate the three classes of problems outlined earlier into a data preprocessing step. By cataloguing those genes involved in nonspecific drug effects, genetic interactions, and known problems with the yeast knockout library, I will construct filters and flag those strains which both exhibit drug sensitivity and are found in the catalog (Bauer, Wolfger et al. 1999; DeRisi, van den Hazel et al. 2000; Mukhopadhyay, Kohli et al. 2002). This preprocessing will allow for a more detailed picture of the small molecule's mechanism of action. In addition, I will perform the analysis on both the filtered and unfiltered set. By examining the differences between those two analyses, I will develop a measurement of general drug responsiveness. This measurement is an expansion of marginal essentiality, a calculation developed by Dr. Yu in our lab (Yu, Greenbaum et al. 2004). Marginal essentiality (M) for each gene i is calculated using the following formula

$$M_i = \frac{\sum_{j \in J_i} F_{i,j} / F_{max,j}}{J_i}$$

where $F_{i,j}$ is the value of the gene i in dataset j, and the datasets J were derived from four large scale phenotypic studies. $F_{max,j}$ represents the largest value in dataset j. The genes were binned according to their marginal essentiality to examine the relationship between various topological parameters and the essentiality or marginal essentiality of the given gene (Yu, Greenbaum et al. 2004). I plan on recalculating marginal essentiality on the chemogenomic profiling data. By redefining marginal essentiality as a measure of the specificity of a gene's response to small molecules or target specificity index, I will assign a probability of a putative target being a specific target as opposed to a general response. Developing such a catalog is important in the analysis of the individual compound, but it will also provide a global view of how drugs affect *S. cerevisiae*.

Potential Complications and Alternative Methods

As this is an iterative process and currently only 100 compounds have been profiled, I expect that the probability score for an individual gene will initially vacillate substantially, but as more compounds are profiled I expect the scores will level off. In addition, there are biases in the initial dataset. For example, antifungal compounds are overrepresented, and one common target of antifungal compounds is the genes of the ergosterol pathway. Just using the marginal essentiality measurement may lead to the conclusion that ergosterol genes have a low target specificity which may or may not be the case. I plan on partitioning the dataset in a number of different ways including by therapeutic goal, as well as, by molecular descriptors. Those genes with low target specificity across multiple data partitions can be assigned a higher confidence value than those found only using one data partition. The different marginal essentiality scores can then be searched for systematic biases.

Molecular Descriptor Calculations

There are a number of different types of molecular descriptors. At the outset, I have decided to use a partial subset of the chemical descriptors chosen to characterize small molecules by the curators of Harvard's new chemical database ChemBank (http://chembank.broad.harvard.edu). The full set of features include 130 structural feature counts, 77 E-state indices, 42 physical and charge properties, and 96 topological and connectivity features (Strausberg and Schreiber 2003). The structural feature counts calculations are trivial to program; however, the physical and charge properties, as well as, the topological and connectivity features, present more of a challenge, and sophisticated software packages have been developed to handle these calculations. The E-state indices can easily be calculated using E-state (Kier and Hall 1999). Initially, I have chosen to focus on the 130 structural feature counts. Other features can be incorporated if necessary.

Specific Aim 2: Clustering, Feature Selection, and Prediction

Once the data is assembled, I will first perform an exploratory analysis of the data by clustering the small molecules on the basis of their phenotypic profile and on the basis of their structural descriptors. Since the Tanimato coefficient is an accepted standard of chemical similarity (James, Weininger et al. 2005), this distance metric will be used to cluster the molecular descriptors. In order to calculate the Tanimato coefficient, I will first convert graphical representations of a compound into a text string encoding known as SMILES (James, Weininger et al. 2005). Applying sub-structural fingerprinting, the SMILES are chopped into 3-4 atom pieces, and each fragment is converted into a bit string. I will take combinations of these fragments and perform bit string similarity comparisons; then using pairwise Tanimato similarity indices (Butina 1999) relate the compounds to one another. Similarity as defined by Tanimato indices is defined as BC/(B1+B2-BC) where BC is the number of bits two fingerprints have in common and B1 and B2 are the total number of bits in compound 1 and compound 2 respectively (Todeschini and Consonni 2000). I will then use this distance metric to cluster the compounds. This clustering will serve only as a crude measurement of molecular similarity. Once a fingerprint has been generated, I can, if necessary, incorporate other features (as described above).

Finally, I will map the functional genomics data on to the target clusters. This type of clustering followed by mapping approach was recently used to identify multidrug resistance genes (Parsons, Brost et al. 2004), to predict synthetic lethal interactions (Wong, Zhang et al. 2004), and finally to map hot spots of toxicity modulation (Begley, Rosenbach et al. 2004). The basic question I am investigating is whether there is a one-to-one relationship or some other mapping between molecular descriptors of small molecules and their biological effect on yeast.

As these examples show, functional relationships can be revealed using this methodology; however, a major flaw in any clustering method is that they are not robust. That is small changes to the initial dataset could result in vastly different clusters. This is of particular concern because the chemical space is many orders of magnitude larger than that sampled here. Therefore, as more compounds are added the results would differ dramatically. The main advantage is that this type of analysis is relatively simple.

Further analysis is required to determine if the relationship between structural descriptors and phenotypic data is strong enough to allow for prediction. By incorporating clustering, feature selection, and prediction, I will determine if it is possible to predict the effect of a new compound given its molecular descriptors.

Collapsing the Output: Clustering the Gene Attribute Table

Chemogenomic profiling ultimately gives a vector of 6000 growth rates, one for each knockout strain. Given a new drug, it would be computationally infeasible to predict which one of the genes encodes the protein that represents the exact drug target. Instead, I plan to first cluster the genes on the basis of their functional genomics data. This will collapse the output vector. The prediction will not be able to classify whether or not gene *i* is a target of Drug 1, but instead the algorithm would generate rules that the small molecule has a probability *x* of hitting gene class *Y* where gene class *Y* has attributes *i*, *j*, *k*, and *l*. Explicitly, the statements I expect would be in the form: "Targets of compounds with three or more rings and at least one heteroatom have a probability *x* of being in pathway *Y*."

Feature Selection

Once the output vector has been shrunk, the problem becomes more manageable. Since the larger the number of attributes used to describe a given instance, the larger the amount of training data, reducing the number of inputs can reduce the computational complexity even further. Selecting the optimal subset of molecular descriptors can be restated in terms of a feature subset selection problem. In other words, given the list of molecular descriptors, select the minimal number of those features necessary to characterize the small molecule. In addition to minimizing the amount of computation that is necessary, reducing the number of descriptors is also important from a chemical point of view. Dividing those descriptors that are absolutely required for activity from those that are marginally or completely irrelevant allows greater flexibility in compound optimization.

Filter and Wrapper Approach

There are two basic approaches to feature subset selection: filter and wrapper. In a filter approach, all of the features are first fed into the feature selection algorithm; then those subsets become the input to the inductive algorithm, and at no point in this process is their feedback between the feature selection algorithm and the inductive learning algorithm. In contrast, in a wrapper approach the feature subset is evaluated in terms of its performance in the learning algorithm. In other words, the inductive algorithm both generates and evaluates the performance of the subset. I plan to implement the wrapper approach where feature selection algorithm is a genetic algorithm and a decision tree is used as the learning algorithm (Figure 1).

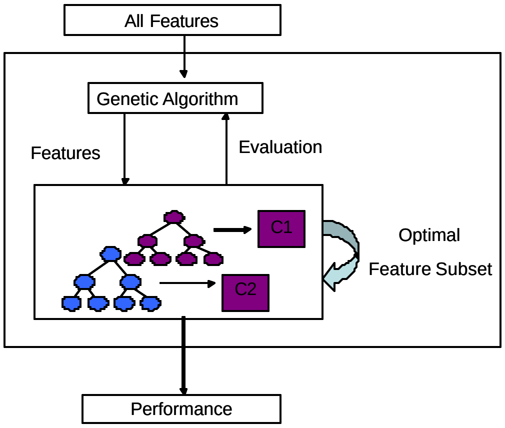

Figure 1 Genetic Algorithm Wrapper Approach
Initial feature selection is performed by the genetic algorithm. The decision tree performance serves as the genetic algorithm's fitness function.

Wrapper Genetic Algorithm

Although there are a number of different feature selection algorithms, genetic algorithms are particularly well suited to this problem because they are (1) capable of efficiently exploring high dimensional search spaces and (2) able to incorporate interactions not just individual features (Yang and Honavar 1998). The basic concept is that new hypothesis or feature subsets are "evolved" by crossing "individuals" hypothesis, and selecting the offspring with the highest fitness. Evolution proceeds as it does naturally through random variation and selection bias (Narayanan, Keedwell et al. 2002). My goal is to pick the minimal subset of molecular descriptors necessary to classify a small molecule. The subsets of descriptors are the individuals, and each Individual is represented as a binary bit string where 1 is the presence of a molecular feature and 0 is the absence. Given two individuals 01110 (features 2, 3, and 4) and 11001 (features 1, 2 and 5), genetic diversity will be introduced by applying the genetic operators: mutation and crossover (Narayanan, Keedwell et al. 2002). For example, mutation can be introduced by flipping position 3 in individual 01110 which results in a new individual 01010. I can cross the two individuals at position 4 to generate new offspring 01100 and 11011. Next, I evaluate the fitness of my new offspring and add the most fit to my population. This process will be repeated until either the search fails or a fitness threshold is reached. Finally, this optimal subset will be used to construct the decision tree.

Crossover (4)		
Parents	10110	01001
Offspring	10100	01010
Mutation (3)		
Wild Type	10110	
Mutant	10010	

Box 1 Genetic Operators
Examples of genetic operators used to generate new offspring.

Building the Decision Trees

The general outline in construction of a decision tree is each node in the tree represents a test for an attribute, and each possible outcome for the test represents a new branch. Finally, the leaves of the tree classify the instances. Since I am implementing a wrapper approach building the decision tree is inextricably linked to the subset selection scheme. I will use a toy problem to explain the details of what I propose. In this toy example, there are three molecular descriptors, the attribute name and all of their possible values are listed in Table 2. I also have data for compounds A, B, and C (Table 3). Compounds A, B, and C represent the training data for building a decision tree to predict whether or not a new compound D (test instance) has targets in the ergosterol synthesis pathway. The first step is feature selection. I randomly decide to use all three features creating individual 111. A decision tree is constructed on the basis of the training data. Next, compound D is classified on the basis of this tree. It is easy to see that this tree completely misclassifies the new instance. The fitness function in this case is simply accuracy(x) where accuracy(x) is the percent of correctly classified instances.

Attribute	***Values***
Ring	Present (P) Absent (A)
Size	Large Small
Pattern2	Present (P) Absent (A)

Table 2 Molecular descriptors and their possible values for toy problem

Features				**Outcome**
Compound	Ring	Size	Pattern2	Ergosterol genes as target
A	P	Large	P	Yes
B	A	Small	A	No
C	A	Large	P	Yes
D	P	Small	P	No

Table 3 Hypothetical data for toy problem Compounds A, B, and C represent training data. Compound D is shaded and was used for testing.

The fitness of this subset is 0, and a new subset is generated. The process is repeated until a fitness threshold is reached. In this toy example, it is trivial to observe that the values of size perfectly partition the compounds into those where ergosterol is a target and those where it is not.

Since I am selecting a subset amongst 130 features, the filter approach would be much faster, but the accuracy would be much worse. In contrast, the wrapper approach will improve accuracy although it may be extremely slow.

The initial training will be done on 90% of the chemogenomic profiling data. The remaining ten percent can be used to test the algorithm. Estimation of the classification accuracy can be determined using ten-fold cross validation (Duda, Hart et al. 2001).

Decision trees are particularly useful in this type of problem because the classification of the instances is transparent. Since each classification generates a specific set of rules, these rules can be treated as hypotheses, and their validity tested experimentally.

Specific Aim 3: Chemogenomic Profiling NCI Challenge Collection

NCI Compound Collection

In order to test the validity of the in silico predictions, I plan to perform some additional fitness profiling experiments using a set of uncharacterized compounds that affect cell proliferation in yeast which is available from the National Cancer Institute. This collection is comprised of 50 compounds that render yeast strains sensitive to radiation and/or cell proliferation (http://dtp.nci.nih.gov/yacds/index.html).

Methods

The yeast homozygous deletion collection will be treated with each of the 50 compounds, and the resulting effect on each yeast strain measured using barcode profiling of growth rates across 20 generations. Specifically, all the strains will be pooled together, and one set will be grown in the presence of the drug and another set in rich media. Genomic DNA will be extracted through zymolase treatment and ethanol/isopropanol precipitation. Dye-labeled primers allow for one step amplification and labeling reaction. The control or strains grown in wild type will be labeled with Cy5, and the experimental or drug treated will be Cy3 labeled. Because of known biases with these fluorescent dyes, I will also perform a dye swapping experiment. Accuracy of PCR primers will be checked by running products and no DNA controls on a 2% agarose gel. To reduce the levels of nonspecific hybridization, PCR products will be mixed with a set of blocking oligonucleotides before hybridizing samples to array (Pan, Yuan et al. 2004). Additionally, DTT supplemented hybridization solution will be used to reduce the oxidative degradation of Cy5. Hybridization is performed overnight, followed by washing, and drying of arrays. The arrays will then be scanned using GenePix 4000B scanner. Analysis of arrays will be performed using a combination of R and perl code.

A particular consideration in using the deletion collection is the development of strain specific error models. Since 15% of all viable homozygous deletion strains exhibit a slow growth phenotype (Giaever, Chu et al. 2002), I will hybridize together two sets of wild type strains: one set labeled with Cy3, and one labeled with Cy5, in order to derive the stochasticity of each individual deletion strain’s growth. These control growth arrays will be used to develop strain specific error models. Strain defects are quantitated by taking the ratio of Cy5/Cy3 and correcting for average growth rate determined from the control experiments. Strains that exhibit significant growth defects will be retested individually for the effect of the compound on cell growth. These findings will be compared with the predicted results. I hope to find that strains defective in a particular pathway are sensitive to the drug and that this pathway harbors the predicted targets.

Potential Complications and Alternative Methods

Since many of the compounds exhibited their effect in the presence of radiation, it is possible that cells grown in rich media may not exhibit an effect for cells grown under rich medium. Thus, I may need to subject the cells to varying levels of irradiation to observe an effect.

Outlook

In conclusion, this is an exciting time to be involved in chemical genetics. A number of laboratories have announced plans to embark on high throughput chemogenomic profiling which should lead to a dramatic increase in the number of compounds. The larger the dataset and specifically the better the coverage of chemical space, the more generalizable the predictions from my algorithm will be. Characterizing a small molecule both in terms of its structural descriptors and biological activity will yield new insight on specific gene function. On a global level, by building a target specificity index, one can develop a more detailed understanding of general drug response. Since Steinmetz, et al. found a number of human disease gene homologs in *S. cerevisiae* it is not inconceivable that ultimately the underlying principles of structural features and biological activity discovered in *S. cerevisiae* may be transferable to other organisms, such as humans.

References

Bauer, B. E., H. Wolfger, et al. (1999). "Inventory and function of yeast ABC proteins: about sex, stress, pleiotropic drug and heavy metal resistance." Biochim Biophys Acta **1461**(2): 217-36.

Begley, T. J., A. S. Rosenbach, et al. (2004). "Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping." Mol Cell **16**(1): 117-25.

Bemis, G. W. and M. A. Murcko (1996). "The properties of known drugs. 1. Molecular frameworks." J Med Chem **39**(15): 2887-93.

Birrell, G. W., G. Giaever, et al. (2001). "A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity." Proc Natl Acad Sci U S A **98**(22): 12608-13.

Butina, D. (1999). " Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets." J. Chem. Inf. Comput. Sci **39**(4): 747-750.

DeRisi, J., B. van den Hazel, et al. (2000). "Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants." FEBS Lett **470**(2): 156-60.

Duda, R. O., P. E. Hart, et al. (2001). Pattern classification. New York; Chichester [England], Wiley.

Eason, R. G., N. Pourmand, et al. (2004). "Characterization of synthetic DNA bar codes in Saccharomyces cerevisiae gene-deletion strains." Proc Natl Acad Sci U S A **101**(30): 11046-51.

Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.

Giaever, G., P. Flaherty, et al. (2004). "Chemogenomic profiling: identifying the functional interactions of small molecules in yeast." Proc Natl Acad Sci U S A **101**(3): 793-8.

Giaever, G., D. D. Shoemaker, et al. (1999). "Genomic profiling of drug sensitivities via induced haploinsufficiency." Nat Genet **21**(3): 278-83.

Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.

Grunenfelder, B. and E. A. Winzeler (2002). "Treasures and traps in genome-wide data sets: case examples from yeast." Nat Rev Genet **3**(9): 653-61.

Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-26.

Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.

Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A **98**(8): 4569-74.

Ito, T., K. Tashiro, et al. (2000). "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins." Proc Natl Acad Sci U S A **97**(3): 1143-7.

James, C., D. Weininger, et al. (2005). "Daylight Theory Manual." from http://www.daylight.com/dayhtml/doc/theory/theory.toc.html#Table%20of%20Contents.

Kier, L. and Hall (1999). Molecular Structure Description. New York, Academic Press.

Kumar, A., S. Agarwal, et al. (2002). "Subcellular localization of the yeast proteome." Genes Dev **16**(6): 707-19.

Kurtzman, C. P. and C. J. Robnett (2003). "Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses." FEMS Yeast Res **3**(4): 417-32.

Launhardt, H., A. Hinnen, et al. (1998). "Drug-induced phenotypes provide a tool for the functional analysis of yeast genes." Yeast **14**(10): 935-42.

Lum, P. Y., C. D. Armour, et al. (2004). "Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes." Cell **116**(1): 121-37.

Mannhold, H., H. Kubinyi, et al. (2000). Handbook of Molecular Descriptors. Germany, Wiley.

Mukhopadhyay, K., A. Kohli, et al. (2002). "Drug susceptibilities of yeast cells are affected by membrane lipid composition." Antimicrob Agents Chemother **46**(12): 3695-705.

Narayanan, A., E. C. Keedwell, et al. (2002). "Artificial intelligence techniques for bioinformatics." Appl Bioinformatics **1**(4): 191-222.

Pan, X., D. S. Yuan, et al. (2004). "A robust toolkit for functional profiling of the yeast genome." Mol Cell **16**(3): 487-96.

Parsons, A. B., R. L. Brost, et al. (2004). "Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways." Nat Biotechnol **22**(1): 62-9.

Rine, J., W. Hansen, et al. (1983). "Targeted selection of recombinant clones through gene dosage effects." Proc Natl Acad Sci U S A **80**(22): 6750-4.

Schwikowski, B., P. Uetz, et al. (2000). "A network of protein-protein interactions in yeast." Nat Biotechnol **18**(12): 1257-61.

Sharom, J. R., D. S. Bellows, et al. (2004). "From large networks to small molecules." Curr Opin Chem Biol **8**(1): 81-90.

Steinmetz, L. M., C. Scharfe, et al. (2002). "Systematic screen for human disease genes in yeast." Nat Genet **31**(4): 400-4.

Stockwell, B. R. (2000). "Chemical genetics: ligand-based discovery of gene function." Nat Rev Genet **1**(2): 116-25.

Strausberg, R. L. and S. L. Schreiber (2003). "From knowing to controlling: a path from genomics to drugs using small molecule probes." Science **300**(5617): 294-5.

Todeschini, R. and V. Consonni (2000). Handbook of Molecular Descriptors. New York, Wiley-VCH.

Tong, A. H., M. Evangelista, et al. (2001). "Systematic genetic analysis with ordered arrays of yeast deletion mutants." Science **294**(5550): 2364-8.

Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.

Winzeler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." Science **285**(5429): 901-6.

Wong, S. L., L. V. Zhang, et al. (2004). "Combining biological networks to predict genetic interactions." Proc Natl Acad Sci U S A **101**(44): 15682-7.

Xenarios, I., L. Salwinski, et al. (2002). "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions." Nucl. Acids Res. **30**(1): 303-305.

Yang, Y. and V. Honavar (1998). "Feature Subset Selection using a Genetic Algorithm." IEEE **3**(2): 44-49.

Yu, H., D. Greenbaum, et al. (2004). "Genomic analysis of essentiality within protein networks." Trends Genet **20**(6): 227-31.

Zanders, E. D., D. S. Bailey, et al. (2002). "Probes for chemical genomics by design." Drug Discov Today **7**(13): 711-8.