Modular Architecture in Biological Networks

by

Gopal Ramachandran
B.S., University of Washington, 1999

Submitted to the
Department of Mathematics
and the
Harvard/MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

Massachusetts Institute of Technology

May 2007

© Gopal Ramachandran, 2007. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper
and electronic copies of this thesis document in whole or in part in any medium now
known or hereafter created

Author ..............................................................

Department of Mathematics/Division of Health Sciences and Technology

May 23, 2007

Certified by .............................................................

Bonnie A. Berger
Professor of Applied Mathematics
Thesis Supervisor

Accepted by .............................................................

Martha L. Gray Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Director, Harvard-MIT Division of Health Sciences and Technology
Modular Architecture in Biological Networks

In the past decade, biology has been revolutionized by an explosion in the availability of data. Translating this new wealth of information into meaningful biological insights and clinical breakthroughs will require a complete overhaul both in the questions being asked, and the methodologies used to answer them. One of the largest challenges in organizing and understanding the data coming from genome sequencing, microarray experiments, and other high-throughput measurements, will be the ability to find large-scale structure in biological systems. Ideally, this would lead to a simplified representation, wherein the thousands of genes in an organism can be viewed as a much smaller number of dynamic modules working in concert to accomplish cellular functions.

Toward demonstrating the importance of higher-level, modular structure in biological systems, we have performed the following analyses:

1. Using computational techniques and pre-existing protein-protein interaction (PPI) data, we have developed general tools to find and validate modular structure. We have applied these approaches to the PPI networks of yeast, fly, worm, and human.

2. Utilizing a modular scaffold, we have generated predictions that attempt to explain existing system-wide experiments as well as predict the function of otherwise uncharacterized proteins.

3. Following the example of comparative genomics, we have aligned biological networks at the modular level to elucidate principles of how modules evolve. We show that conserved modular structure can further aid in functional annotation across the proteome.

In addition to the detection and use of modular structure for computational analyses, experimental techniques must be adapted to support top-down strategies, and the targeting of entire modules with combinations of small-molecules. With this in mind, we have designed experimental strategies to find sets of small-molecules capable of perturbing functional modules through a variety of distinct, but related, mechanisms. As a first test, we have looked for classes of small-molecules targeting growth signaling through the phosphatidylinositol-3-kinase (PI3K) pathway. This provides a platform for developing new screening techniques in the setting of biology relevant to diabetes and cancer.

In combination, these investigations provide an extensible computational approach to finding and utilizing modular structure in biological networks, and experimental approaches to bring them toward clinical endpoints.
Acknowledgements

First and foremost, I would like to thank my advisor, Bonnie Berger. Without her enduring support, encouragement to explore new directions, and enthusiasm toward myriad scientific questions, this thesis would not have been possible. I would also like to thank Daniel Kleitman and Peter Shor for insightful commentary and helpful discussions.

I am very grateful to Stuart Schreiber, for giving me the opportunity to conduct experiments in his laboratory and opening the door to many exciting new scientific avenues. Many of my research directions were strongly influenced by conversations with Stuart and my colleagues at the Broad Institute.

I would also like to thank Paul Eder, for not only introducing me to his many patients, but also taking the time to teach me many subtleties of clinical and translational medicine. Through our many conversations and interactions I have gotten a first glimpse into the arduous task of developing new therapeutics.

I am incredibly grateful to Collin Stultz. Many of the new directions now before me would not have been possible without Collin’s tireless support. In addition to providing me with continuous mentorship and encouragement, Collin also provided a model of both a clinician and a scientist that will stick with me through the years to come.

Finally, above all, I would like to thank my parents. I cannot hope to express what they mean to me, nor how grateful I am for their unconditional love and support. I am grateful for everything they have taught me, and all of the values they have instilled in me. Without them, I would never have arrived at this point.
# Table of Contents

1. Introduction ........................................................................................................... 1  
   1.1. Determination and Validation of Modular Structure ........................................... 4  
   1.2. Modularity Applied to High Throughput Screen Data ....................................... 11  
   1.3. Modular Evolution and Alignment of Biological Modules ................................. 14  
   1.4. Discovery of Modularly-Targeted, Small-Molecule Perturbagens .................... 17  
   1.5. Future Directions

2. Biological Preliminaries ..................................................................................... 22  
   2.1. Central Dogma ............................................................................................... 24  
       2.1.1. Central Dogma I: DNA ........................................................................ 25  
       2.1.2. Central Dogma II: RNA ...................................................................... 34  
       2.1.3. Central Dogma III: Protein ................................................................. 37  
       2.1.4. Central Dogma IV: Limitations and Extensions ................................... 44  
   2.2. Conclusions .................................................................................................... 47

3. Discovering and Validating Modular Structure ............................................. 49  
   3.1. Density in Randomized, Inhomogeneous Graphs ........................................... 52  
   3.2. Noise Tolerance of Edge Metrics .................................................................... 80  
   3.3. Biological Validation ..................................................................................... 96  
   3.4. Summary ....................................................................................................... 106  
   3.5. Conclusions ................................................................................................... 108

4. Modularity Applied to Analysis of High-Throughput Data ......................... 109  
   4.1. Methods ....................................................................................................... 114  
   4.2. Results .......................................................................................................... 116  
       4.2.1. Non-Specific Compound-Module Interactions ...................................... 117  
       4.2.2. Compound Specific Modular Explanations ........................................ 119  
   4.3. Conclusions ................................................................................................... 124

5. Modular Alignment and Conservation ......................................................... 126  
   5.1. Approach ...................................................................................................... 128  
       5.1.1. Approach: Pairwise Alignment of Modules .......................................... 128  
       5.1.2. Approach: Alignment of Modules Across Multiple Species .................. 130  
       5.1.3. Approach: Super-Modular Structure from a Module Tree .................... 132  
   5.2. Results ......................................................................................................... 133  
       5.2.1. Results: Pairwise Alignment of Modules ............................................. 133  
       5.2.2. Results: Alignment of Modules Across Species ...................................... 141
5.2.3.  Results: Super-Modular Structure from a Module Tree.........................143

6.  Discovering Modularly Targeted Small-Molecules........................................145
   6.1.  Screening of Commercial Compounds for U73122 & Wortmannin Modifiers in
         Budding Yeast............................................................150
   6.2.  Retesting Hits from Commercial Compound Primary Screen...................156

7.  Systematic Interpretation of Genetic Effects Using a Scaffold of Molecular
    Interactions..............................................................................160
   7.1.  Introduction........................................................................160
   7.2.  Methods..............................................................................161
       7.2.1.  Overview: Modeling Physical Pathway Explanations..................161
       7.2.2.  Formal Model and Scoring.............................................163
       7.2.3.  Search Procedure........................................................164
   7.3.  Results...............................................................................166
       7.3.1.  Construction of the Genetic and Physical Interaction Network.......166
       7.3.2.  Explaining Genetic Relationships with Physical Interaction Paths....169
   7.4.  Discussion.........................................................................173
       7.4.1.  The Physical Pathway Model Associates New Functions with Genetic
               Interactions..................................................................173
       7.4.2.  Protein Interaction Chains are the Predominant Explanation for Genetic
               Interactions....................................................................174
       7.4.3.  Genetic Interaction Predictions fill in Gaps of Knowledge in the Existing
               Genetic Network................................................................175
       7.4.4.  Significant Enrichment for Essential Proteins Downstream of Explanatory
               Paths..............................................................................176
   7.5.  Conclusion.........................................................................176

8.  Dissection of Complex Traits in S. Cerevisiae Using Small-Molecules............177
   8.1.  Introduction......................................................................177
   8.2.  Methods............................................................................178
   8.3.  Results..............................................................................179
   8.4.  Conclusions.....................................................................186

9.  Future Directions.....................................................................190
   9.1.  History of Chemotherapy....................................................190
   9.2.  Targeted Trials..................................................................199
   9.3.  Modularly Targeted Therapeutics............................................200

Bibliography..............................................................................201
1. Introduction

In the past decade, biology has been revolutionized by an explosion in the availability of data. New technologies can make experimental measurements once thought impossible. Moreover, it is now feasible to simultaneously measure the amounts of every transcript in a cell as well as the interaction partners of every protein. These approaches promise to radically change biology and medicine by allowing the cell to be viewed from a holistic point of view previously impossible in an era of reductionist science.

In 2001, the first complete draft sequence of the human genome was completed\textsuperscript{10,11}. Through this data it was possible to get a glimpse of human history and ancestry as well as to start compiling a complete “parts list” for human cells. Though this landmark accomplishment was an unparalleled step forward, it has also demonstrated the difficulty of the path ahead: utilizing this new resource would require a complete overhaul in the type of questions being asked, and such a change in questions would require rethinking scientific methodologies for these new data types. Looking forward, it became reasonable to unify many disparate aspects of biology and conceive of a dynamic model of a cell that would allow one to see all the thousands of pieces of a cell working in concert to accomplish the many varied cellular functions.

Throughout the twentieth century, most scientific endeavors relied on maximally isolating a phenomenon of interest to understand a given system one piece at a time. Specifically, in biology, the goal was often to dissect a particular cellular function by defining the role of each contributing protein, one by one. This model worked extremely well in a time wherein data was a precious commodity, and often the limiting factor.
Now, biology has been transformed from a data-poor field to a data-rich field with new genomic and high-throughput technologies capable of working synergistically with classical approaches. Such techniques are reliant on top-down and system-wide viewpoints requiring approaches from synthetic sciences such as computer science and engineering. With the change in viewpoint has come a huge paradigmatic shift in the style of questions being asked and the technologies and methods required to answer them. For instance, decades of cancer research asked the question “how is the function of p53 important to the progression of malignancy?” To rephrase this in the context of a systems biology question: “what are all of the groups of genes, that when perturbed, lead to human cancer?” The answers to these questions may shed light on large swaths of clinically-relevant biology on much shorter time scales.

While novel technologies open the possibility of making system-wide measurements, the ability to turn this data into meaningful biological conclusions presents a much larger challenge. With this in mind, it is reasonable to construct a simple roadmap of how new fields will feed into one another to transform millions of measurements into novel biological insights. Genome sequencing has become the starting point of moving toward dynamic, predictive models, by enumerating the components, both at the level of genes, and at the level of non-coding regulatory elements. The next step will require organizing these genomic elements into a more structured static picture utilizing data closer to a functional level. To choose from the many available approaches and technologies requires clarifying the underlying goal of this step: to construct an “interaction scaffold” that represents every possible interaction between groups of cellular constituents, independent of context. Limiting versions, or
instances, of this structure can be viewed as snapshots of cellular state. Representing different contextual instances on this scaffold (e.g. how cells respond to a specific stimulus such as nutrient deprivation, osmotic stress, etc.) can be accomplished using dynamic measurements such as gene expression or protein levels. With this combination of data, it is possible to see how cells work within the constraints of the scaffold to accomplish myriad cellular functions.

Perhaps the most critical roadblock in understanding the new sea of data is the ability to find high-level structure in biological systems, toward building an interaction scaffold. In many ways, this is analogous to understanding the organization of a computer by disseminating the high level components from the low level parts. By viewing the computer one transistor at a time, and trying to isolate and understand each transistor’s role, it would be virtually impossible to decipher how it works. However by starting from a top-down viewpoint all of the critical components become immediately apparent: CPU, memory, hard drive, etc. Similarly, in biological systems, there is considerable reason to believe that there is substantial high-level organization that is critical to the accomplishment of cellular functions. Uncovering this structure will allow not only the ability to understand cellular organization, it will allow the cell to be intuited as a small set of interacting functional modules rather than tens of thousands of individual genes.

With the goal of deciphering and utilizing the modular architecture of biological systems, this thesis will have four primary objectives:

1. Using computational techniques and pre-existing data from several different species, develop general tools to discover and validate
modular structure in well-characterized biological systems. This will result in the construction of a modular scaffold on which to ask subsequent questions.

2. Utilize the modular scaffold to construct an efficient reduced representation. Using this simplified, high-level representation, generate hypotheses that might explain existing system-wide experiments. This representation is meant to be equally facile at capturing Mendelian and complex traits.

3. Following the example of comparative genomics, align biological networks to elucidate principles for how modules evolve. Use conserved modular structure to further aid in functional annotation across the proteome.

4. Design experimental strategies to find sets of small-molecules capable of perturbing functional modules through a variety of distinct, but related, mechanisms.

1.1 Determination and Validation of Modular Structure

Before beginning the search for modularity in biological data, there are a number of factors to consider: what type of data to use, what properties are these data known to have that might confound the study, what approaches are amenable to determining modularity, and how can any patterns in the data be validated as bona fide modular structure? Thus, we begin with a brief introduction to the field, associated approaches, and previous work, followed by the details of our approach.
In order for modularity to be useful as a means for describing cellular functions, it must have two properties: 1) discrete tasks are accomplished by a small fraction of cellular constituents, and 2) multiple moieties must work in concert to achieve these goals. Ravasz et al. made the first demonstration of large-scale modularity in biological systems by showing that the metabolite network in 44 species of bacteria could be systematically decomposed into small subnetworks representing different functional components of intermediary metabolism. Metabolic networks provided a perfect platform in which to elegantly demonstrate the presence of modularity, due to the pristine quality of the underlying data: each interaction had been painstakingly mapped such that there was little doubt about the veracity of each link in the network. Unfortunately, the methodology of Ravasz et al. proved hard to generalize in the context of larger, noisier datasets plagued with incorrect interactions and potential systematic errors. Making this work generally applicable will require techniques capable of using structure to separate noise from genuine interactions.

Given the many types of high-throughput data now available, there are many avenues from which to look for the high-level structures in cellular systems. Foremost amongst these are protein-protein interaction (PPI), gene-expression (GE), metabolic, genetic, and chromatin-immunoprecipitation (ChIP) networks. In many ways these networks provide complementary information useful in different contexts. Specifically, modularity can be achieved through different means, and different experimental approaches can be helpful for elucidating each individually. For example, a cell’s ability to separate different functions temporally is perhaps best captured by high-throughput gene expression measurements, whereas protein-protein interactions
demonstrate modularity achieved by specificity of interaction between different molecules. In the future it will be interesting to decipher how these different types of modularity are utilized in organisms ranging from the simplest of single cells to higher mammals. A particularly important question will be how multicellular organisms use cellular boundaries as an additional level of spatial modularity to build dramatically more complicated systems. Additionally, does this additional level of modularity and organizational complexity make multicellular systems more flexible in terms of temporal and interaction modularity, allowing possible evolutionary benefits?

PPI networks provide a natural starting point toward building an interaction scaffold, since they capture all of the possible interactions as defined by molecular specificity, while they are particularly context independent. PPI networks first became available with the work of Uetz et al\textsuperscript{3} and Ito et al\textsuperscript{4,5} using Yeast Two Hybrid (Y2H) technology. Y2H allowed the sequential testing of pairs of proteins for physical interaction, irrespective of whether they shared the same spatial localization or temporal regulation. Though an extremely scaleable technique, Y2H has been demonstrated to be very noisy with both high false-positive and high false-negative rates. Subsequent technologies include Mass Spectrometry (MS)\textsuperscript{4,5} and Co-Immunoprecipitation (CoIP)\textsuperscript{12}, which provided parallel approaches and data sources to enlarge and support Y2H data. These data can be combined to represent the interactome of a given organism. Currently, interactomes have been constructed for a number of organisms including: yeast\textsuperscript{4,5,6,7}, worm\textsuperscript{13}, fly\textsuperscript{28}, and human\textsuperscript{35}. These four interactomes will be the focus of our studies.

On the path toward discovering modular structure in these networks, it will be necessary to find a systematic approach for decomposing complex biological networks.
into groups of distinct modules. Additionally, this approach must be able to toggle the size of a particular module to look at a given system and function at different levels of hierarchical specificity. For instance, it should be possible to find the entire set of proteins involved in RNA metabolism (on the order of hundreds of proteins), while simultaneously maintaining the internal structure consisting of modules involved in transcription, ribosome biogenesis, tRNA synthesis, snoRNA metabolism, etc. (each containing on the order of tens of proteins). This hierarchical modularity will allow the observation of any trait or function of interest at the appropriate corresponding component size-scale.

Given the noisiness of the data on which these analyses will be based, it is critical to demonstrate: 1) any modular structure found is significant relative to that expected at random and 2) that the computational means for finding modular structure are sufficiently resistant to noise relative to that expected in the data.

Historically, many properties of theoretical networks, such as local density can be solved for analytically. However, this is reliant on the network being homogenous, or constructed in a systematic way that aids in computing properties of interest. The construction of randomized networks, based on networks of interest, is especially critical to the evaluation of decompositions of biological networks due to their documented inhomogeneity. The first, and still considered by many to be the most important, property noted in these networks was degree distribution: before the advent of Y2H, biological networks had been assumed to have the topology of random graphs, wherein most nodes have approximately the same number of neighbors. However, upon completion of the first two large-scale Y2H datasets\textsuperscript{3,5}, it became apparent that these
networks did not exhibit the relatively homogeneous topology expected in a random graph. Upon inspection it was clear that several nodes, so-called "hubs," had many more neighbors than would be expected at random, and further analysis showed that the number of expected neighbors for a given node followed a power-law (scale-invariant)\(^46,47,49\). This is analogous to airline route maps: a very small number of airports have many more incoming and outgoing flights and serve as key routing points for airline travel. When evaluating the modular structure of a network with such inhomogeneities, it is critical to determine how much modularity is present in addition to that expected for graphs with these characteristic degree distributions.

Toward detecting modularity beyond that expected at random in a scale-invariant network, it is possible to construct a randomize graph that maintains scale-invariance, while disrupting any modularity. This allows a direct evaluation of the degree of non-random modularity observed.

If the data is represented as a graph consisting of nodes (representing the proteins) and edges (representing the interactions between them), a number of different strategies from graph theory provide a natural starting point, including: spectral methods and graph-theoretic node and edge metrics used in conjunction with clustering.

Spectral methods have been used in a variety of fields including theoretical physics, population genetics, statistics, machine learning, and mathematics. Spectral values can provide a way to rapidly characterize the gross characteristics of graphs, determine whether two graphs are similar, and how many edges must be cut to break the graph into two distinct subgraphs. Though elegant in their ability to handle homogenous graphs, spectral methods struggle to decompose inhomogenous graphs and become
highly sensitive to the method used to determine the cuts. Specifically, slight changes in how different parameters are weighted, especially the number of cuts, can preclude a non-degenerate decomposition. Perhaps the biggest failure of spectral methods is their inability to simultaneously handle modules of different sizes.

One intuitive notion of modularity is the expectation that constituents of a module interact much more with one another within the module than they do with elements outside of the module. This intuition can be represented by iterating through pairs of nodes in a graph, and asking how many neighbors they have in common. It is expected that nodes within a module will share many more interacting neighbors than those in separate modules. Mathematically, this notion is characterized by edge metrics: scores that weight the similarity of interaction between two edges. Perhaps the simplest way to decompose a graph is to start with weights for each edge followed by removing edges with successively higher weights (where edges with higher weight indicate more neighbors in common). At first, large loosely connected modules will be present, eventually being split into tighter submodules focused around progressively narrower functions.

1.1.1 Determination and Validation of Modular Structure: Approach

As a starting point, we will modularly decompose interaction networks from yeast, fly, worm, and human. For each network, we will perform multiple decompositions utilizing different metrics in conjunction with hierarchical clustering, to assess their individual performance. The decomposition will result in a “modular tree,” representing the interactome of each species. The majority of validation or our approach
will not require the tree structure; the distribution of scores across the edges will suffice. The tree structure, and the modules that comprise it, will be key for investigations in subsequent sections.

In order to validate our decompositions we will construct three sets of interaction networks for each original network: 1) randomized networks that maintain the degree distribution, but have different connectivity than the original network, 2) networks that, starting with the original network, have add edges randomly removed, and 3) networks that, starting with the original network, have had edges randomly added. We discuss each in turn below.

Comparison to randomized networks will allow a rigorous, quantitative evaluation of a given decomposition method’s ability to detect modularity irrespective of other topological features, especially inhomogeneous degree distribution. One question that still remains to be answered is, what is the appropriate type of randomization? Randomizing to maintain degree distribution is one possibility, while randomizing to maintain joint degree distribution across edges is another. For our purposes, we will randomize the original network while maintaining degree distribution.

Similarly, addition and removal of nodes, will determine how robust a given decomposition method is to different amounts of noise: both false positives and false negatives. Using a family of networks that have had edges added or removed, different edge metrics can be compared head-to-head, as well as provide an absolute assessment of their noise tolerance. Thus, we will score the noisy graphs with all of the edge metrics and look for correlations between the scores of edges common to the original graphs and the new noisy graphs.
Though the preceding tests provide a test of the statistical validity of these methods, they in no way support the biological significance. To assess the performance of our decomposition methods in discovering biologically meaningful modules, we will compare them to known annotations. For predictions made on the order of the entire proteome, the only tools available for benchmarking are Gene Ontology (GO) annotations\textsuperscript{40}. GO annotations are derived primarily from low-throughput experiments making them useful as a collective “gold standard.” For a given protein, they detail the biological process it is involved in, its molecular function, and which cellular compartment it resides in. In the budding yeast Saccharomyces Cerevisiae, GO annotations are available for almost 75\% of proteins\textsuperscript{72}. Thus, we will look at the modules found through our approach to determine if proteins within a module are significantly enriched for a function or process according to GO annotation.

By assessing our approach from purely topological and purely biological viewpoints separately, we are able to show that these modules are a robust feature in the data that is representative of meaningful biological structure. Next, we apply these modular structures to the analysis of large biological datasets.

1.2 Modularity Applied to High Throughput Screen Data

If one of the primary goals of recasting biological phenomena from a modular viewpoint is the ability to understand complicated data, a perfect test is whether modular structure can generate hypotheses capable of explaining high-throughput screen data. For example, screening technologies, as employed by biotech and pharmaceutical companies are capable of generating thousands to millions of data points per day using robotics and
miniaturized screening approaches. With these technologies comes the promise of being able to more quickly discover both new drug targets and novel therapeutics. To make these advances will require being able to organize and understand an incredible amount of data, providing a perfect test bed on which to benchmark these approaches.

A key difference that separates our approach, outside of the particulars, is the intent: analyses of most large datasets strive to find a “hit.” Where a “hit” is defined as a single gene or small-molecule that explains the phenomenon of interest. Our approach is not to find the needle in the haystack, as many more complicated phenomena are not explained by a single entity. Instead, we attempt to organize biological systems into processes of different scopes and scales, and determine which impinge most of our disease or process under investigation. Thus, we describe the dataset of interest with a discussion of Mendelian and complex traits, followed by the details of our approach.

In 2003, Rosetta Inpharmatics performed a groundbreaking screen that combined a classical genetic approach with new genomic technologies to better understand the action of common therapeutics. Rosetta started with a collection of ~3,500 yeast strains, each heterozygous deleted for a given gene, creating a unique genetic Achilles heel for each strain. By subjecting each strain to 78 compounds (many common clinical agents such as: aspirin, lovastatin, 5-FU, etc.), combined with a highly efficient pooling strategy, they were able to validate known targets for several compounds as well as suggest targets for molecules with unknown targets. This provided a means to use a classical genetic approach on a much larger scale, but still was limited to looking for strong effects limited to a single gene. Unfortunately, the majority of compounds achieve their effects through
actions not solely through the action of a single gene, and might be better understood in terms of synergistic interactions in the context of a module. This underscores an important turning point at the intersection of medicine and genetics/genomics.

Before completion of sequencing of the human genome, the effect of genetics on medicine was primarily executed through diagnosis of extremely rare monogenic, Mendelian diseases. Though these diagnoses are critical to the care of these patients, they are limited in their scope. Looking forward to the era of post-genomic biology, a fusion of genetics and computational biology can impact common diseases, such as hypertension, diabetes, myocardial infarction, and prostate cancer. Though these diseases have been shown to have a substantial genetic component, their inheritance is far more complicated than Mendelian disorders, and has been difficult to fully decipher. As genetics and genomics unlock the critical sets of genes in these diseases, the challenge will be to construct a coherent system-wide picture of their interconnections. Moreover, translating that picture into new therapeutics will require similar system-level understanding of the effects of target compounds, and the ability to match such a picture to a disease with a similar modular profile. The Rosetta dataset provides an ideal platform from which to attempt to understand the multifaceted effects of commonly used medications and tool compounds.

Although the function of these molecules in yeast is not of direct clinical relevance, it may be a step toward better understanding the myriad secondary effects of commonly used small molecules as many proteins are conserved between yeast and humans.
1.2.1 Modularity Applied to High Throughput Screen Data: Approach

By overlaying this data onto a modular scaffold, it may become apparent how the effects of a given molecule are spread across different functional components. To evaluate the usefulness of this approach, we will project treatment with each compound onto the modular tree discussed in the previous section, to determine whether the partitioning represented by the tree captures the structure of each phenotype.

Our analyses of the Rosetta data begin with the modular tree mentioned in the previous section. The tree is meant to capture all of the cellular organization covered by the subset of the proteome that was used to construct it. For each trait and the data from the Rosetta study, we will score each module to determine whether the dataset can be meaningfully partitioned modularly. The benchmark of this application is the ability of a small number of modules to explain a phenotype that appears complex at the gene level. We proxy this ability by measuring how a module scores relative to randomized versions of the dataset. To further stratify the strengths and weaknesses of the approach on traits of different complexity, we utilize the partition constructed by Rosetta: based on the profiles of each compound, the grouped the compounds into three disjoint groups based on how many genes showed a significant effect in the presence of a specific compound.

1.3 Modular Evolution and Alignment of Biological Modules

The field of genomics has utilized the comparison of genomes from different organisms to determine many features of biological information storage and processing such as microRNA discovery, mechanisms of gene regulation, how genomes control genetic variation, and how environmental stimuli mold genomic evolution. Though
genomics can observe many of the effects of evolutionary pressures on genes through changes in gene and protein sequences, it cannot provide functional hypotheses as to how a given gene and the functions to which it contributes affect the fitness of the organism.

To begin tackling these problems will require observing how functional components have evolved over thousands to millions of years. As the genome provided a picture of how the constituents changed, the structure of the proteome reflects how large groups of proteins have changed to adapt an organism's diverse capabilities.

In conjunction with work on molecular evolution, this approach will provide a complimentary way to view how biological systems change, where they can afford to be malleable, and where they must be rigid to maintain robustness in essential functions.

1.3.1 Modular Evolution and Alignment of Biological Modules:

Approach

Starting with the modular decomposition, we will attempt to construct alignments between interactomes in two parts: 1) score alignments between pairs of modules in different networks, and 2) construct global alignments of multiple networks based on the pairwise module scores. The latter is a preliminary investigation, but provides a completely novel means of comparing large-scale structure in biological systems.

The pairwise score will be constructed from the pairwise sequence alignment of each protein in one module to each protein in the other module. This will provide us with the equivalent of a sequence similarity metric for comparing two modules, though it is not normalized to the size of the modules being considered, or the network from which they are drawn (two considerations that are factored into sequence alignments tools such
as BLAST). Though size of the modules could be normalized directly by using their sizes as parameters in the pairwise module score, it is more difficult to directly adjust for the networks from which modules are being drawn. Instead, it is easier to construct a set of randomly constructed modules of similar size for comparison. By building these random modules from the same starting networks, they will correctly normalize for the background distribution of sequences of the starting network, while automatically correcting for the sizes of the modules. Pairwise modular scores can be combined to find meta-modules that represent the modular equivalent of orthologues. These are especially interesting as they define the evolutionary currency of function: these are functional building blocks that have remained relatively unchanged despite speciation and millions of years of subsequent divergent evolution. This will allow us to investigate several fascinating questions: is there a set of universal building blocks, and if so, does evolution occur primarily by changing the connections between building blocks through minor changes in the constituents? Also, if there are some building blocks that are widely conserved, are there features that make them more robust or reusable?

Utilizing the pairwise module scores to construct global alignments, we will seek to construct a meta-decomposition that shows where different networks exhibit similarity in their modular structure. The metadecomposition will proceed first by searching for metamodules (modular templates that represent modules that are conserved across multiple networks), and second by combining metamodules into metanetworks. Metanetworks will attempt to resolve similarities between networks at the super-modular level: how modules connect to one another.
Just as modular structure will be used to make annotative predictions of function, conserved modular structure might provide a more accurate way to make similar predictions. To make predictions, we will start with the yeast network, wherein many functions are known. We will then predict the GO annotations of aligned modules in other interactomes. For validation, we will use the GO annotations that are available over our predictions. In the future, it might be possible to combine sequence, modular structure, and conserved modular structure into a robust framework for understanding and annotating gene function in a high-throughput, scalable way.

1.4 Discovery of Modularly-Targeted, Small-Molecule Perturbagens

Small-molecule discovery has gone through a number of distinct phases: initially molecules were discovered according to their ability to modify phenotypes or symptoms directly without knowledge of the underlying mechanism. As knowledge of the underlying systems has increased, critical effectors have been identified, and new biochemical approaches have been introduced to target these molecules directly. Now as the set of critical effectors related to a given trait or disease increases, it becomes increasingly important to have methodologies that can find groups of molecules capable of affecting these traits through distinct mechanisms. Unfortunately, simply parallelizing earlier targeted biochemical strategies is extraordinarily inefficient. Instead, synthetic genetic methods, hinged on intrinsic modular dependencies could be effectively deployed to find these molecules. Two weaknesses have prevented these genetics approaches from being useful in this setting: 1) the inability to build disease relevant systems to screen
(inability to construct appropriate genetic models in higher-throughput outside of model organisms), and 2) the lack of related genes due to lack of modular neighbor knowledge.

Toward a first attempt at a modular-targeted genetic screen, we will describe our results in a two-part screen aimed at phosphatidyl-inositol-3-kinase (PI3K). In the first stage, we screened a collection of 2,500 known bioactive compounds, searching for a suppressor of wortmannin, a known antagonist of PI3K.

From the three molecules that showed the ability to rescue cell growth in the presence of toxic levels of wortmannin, we chose one for further study, U73122, a known phospholipase-C inhibitor. U73122, showed an extremely interesting and useful property: it was able to inhibit cell growth in isolation, but this effect was suppressed by wortmannin. To search for similar molecules, both wortmannin and U73122 were screened across a library of ~75,000 compounds. We show the results of follow-up/retesting of ~750 molecules that demonstrated enhancer or suppressor phenotypes.

1.5 Future Directions: Modularly-Targeted Combination Therapeutics

Bringing computational biology and genomics to bear on clinical problems will require novel strategies to recast our understanding of diseases from a holistic viewpoint. It is clear that the next step in medical research is observing the many facets of a given disease, instead of focusing only on the most prominent. This will require the capability to measure all of the ways that dysregulated cells differ from normal cells, the means to organize this information into an efficient and clear picture of the perturbation, experimental techniques to find small-molecules focused on the module or function of interest, and methods to find combinations of targeted small-molecules that can correct
the dysfunction in these cells. We present one potential path forward, suggesting potential approaches for the three mentioned objectives.

Perhaps the most direct way to determine how normal and diseased cells differ, is to look at the differences in the levels of each gene or protein in the cell. Despite all of the recent technological advances, accurately measuring protein levels in high-throughput is still nearly impossible. Aside from the measurement of protein-protein interactions, as discussed earlier, most high-throughput measurements are made at the level of nucleotides. These data come in primarily two flavors: DNA and RNA.

In addition to gene sequencing, diseases of somatic genetics such as cancer are making the quantification of DNA levels of genes extremely important — e.g. copy number variation and polymorphisms — in addition to somatic mutations captured by resequencing of known oncogenes and tumor suppressors. In contrast to most genetic diseases, the dynamics of cancer take place at the level of DNA. By combining measurements of gene dosage (copy-number) and mutated regulatory genes (resequencing), it is possible to molecularly dissect the genetic defects and mechanisms underlying malignancy.

With the advent of microarray technologies, the last decade has shown that almost every disease has a unique signature at the level of the transcriptome. Thus, it is reasonable to view gene expression as a general place to start in the hunt for the site of cellular dysfunction. As the technology continues to improve, the amount and quality of information has increased dramatically. Gene expression microarrays can now yield a list of almost every gene disrupted by a disease process, providing an ideal data source from which to understand large-scale cellular changes.
If this information is to be used productively, it will be important to organize it into the larger functional differences gleaned from the tens of thousands of measurements. Using modularity, as discussed above, it might be possible to project gene expression data onto a modular map of the cell, looking for areas where normal cells diverge from diseased cells. This would provide a big-picture view of disease that might indicate which cellular functions are affected. Ideally, the output would be a set of modules that would provide a large set of therapeutics targets.

The completion of the human genome was expected to lead to immediate medical benefit. One bottleneck in this process is the ability to find and validate new targets in high-throughput. The paradigm of drug development in the pre-genomic era relied on having a handful of putative targets for a given disease, and using biochemical techniques that are far from scalable due to their initial overhead. The future of drug development will rely on the ability to evaluate many drug targets in parallel, moving the emphasis on techniques from biochemistry to genetics.

To evaluate the druggability of every constituent of a module will require a rich set of genetic reagents that allow parallel, scalable, high-content measurements. New strategies including RNAi and image-based screening will allow the extremely challenging and time consuming genetics of the past to be rapidly deployed at a genome-wide scale. Using these approaches, it will be possible to rapidly determine the ability of a given target to alter the disease state of interest. By combining these methods with small-molecule screening, new lead molecules can be found based on their ability to copy the phenotypes discovered in the genetics experiments. Genome-level screening
approaches and drug target validation should increase the number of potential therapeutics by orders of magnitude.

The final step is the construction of combinations of small-molecules that can be tailored to a given patient\textsuperscript{88,89,90}. As the study of common, complex diseases has already demonstrated, the genetic complexity is overwhelming. One therapy per disease will not be sufficient to improve treatment of any of the most common diseases. It will be key to move from one molecule per disease toward one molecule per genotype and eventually many molecules per genotype. When the number of drug targets per disease ranges in the hundreds, there will be sufficiently many options to precisely tailor a medication to the disease of a given patient. Learning to construct these combinations will rely on many of above-mentioned components to measure how each drug perturbation affects disease state. Working iteratively, the relationship between genotype and phenotype will become clear enough to rationally construct the optimal therapeutic strategy. When genomic measurements reshape the way the drugs are developed and administered, the era of genomic medicine will truly have arrived.
2. Biological Preliminaries

Before we begin, it is necessary to cover some preliminaries in biology to understand the many ways that biological systems enact information handling and use. Approximately ever decade a new form of control is discovered, before which it was assumed that all modes of regulation were known. Even after the most recent discovery of the role of non-coding RNAs, it is reasonable to think many as of yet undiscovered information processing mechanisms remain to be discovered in the genomes of higher organisms.

Living organisms can be distinguished from one another in many ways: prokaryotes vs. eukaryotes, unicellular vs. multicellular, etc. Though these distinctions are critical when choosing the optimal system to investigate a phenomenon of interest, the similarities that cross these boundaries are equally interesting: all life on earth is carbon-based, uses adenosine triphosphate for energy, and carries its genetic information in virtually the same form.

Before we discuss their similarities, it is useful to cover some of the mentioned ways in which organisms can be distinguished, as this is key to understanding how results in a simple fungus can be extended to higher mammals. For example, most of our current understanding of human cancer, relies on knowledge of cell division cycle checkpoints, which were discovered and characterized in the budding yeast Saccharomyces Cerevisiae\textsuperscript{106}, a simple unicellular eukaryote.
Cells can be broadly divided into two groups: eukaryotes and prokaryotes. The distinguishing feature being the presence (eukaryotes) or absence (prokaryotes) of a nucleus – a small envelope that compartmentalizes a cell’s DNA. Prokaryotes only exist as single celled organisms, are the most diverse organisms on the planet, and constitute the majority of the biomass. The importance and prevalence of bacteria are often missed, since they are often overshadowed by their more complicated and interesting neighbors. What they lack in complexity, they make up for in diversity and ubiquity. Though their relevance to humans, and hence human disease, is extremely limited, they are by far the most studied of model organisms, with the most facile genetics. All of the work we will discuss later is based on work that originated in bacteria.

Eurkaryotes run the gamut of organismal complexity from single-celled fungi to higher mammals (containing on the order of $10^{13}$-$10^{14}$ cells). Most interesting, is that in addition to sharing the presence of a nucleus, eukaryotes share a very large percentage of their genetic information. As this genetic information is documented through sequencing, it becomes increasingly clear that how humans differ from fungi is not determined at the level of constituency of their genetics, but how those constituents are used. This suggests that nature is heavily reliant on the reuse of “functional parts,” and the organization of those parts is critical to understanding the many functions comprising these organisms. One outgrowth of these similarities is the ability to find a eukaryotic system to model a phenomenon of interest.

Many of the greatest breakthroughs in biology have come from attacking a given question in an organism that allows the question to be maximally isolated. Examples abound. Cell cycle checkpoints were elucidated due to the simple genetics of yeast.
Programmed cell death, which plays a central role in development and cellular maintenance, was understood by watching the fates of individual cells in the much less complicated brains of worms. The rules by which cells pass on their genes were discovered in peas.

We will make extensive use of the similarities amongst eukaryotes by designing and testing many approaches in S. Cervisiae (budding yeast), and then extending this work first to the higher model organisms Caenorhabditis Elegans (round worm) and Drosophila Melanogaster (fruit fly), and eventually to Homo Sapiens (human).

2.1 Central Dogma

A natural point to start a discussion of how information is handled in biological systems, is the Central Dogma, which applies to all living organisms:

\[ \text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein} \]

In short the Central Dogma says that cellular information begins at the level of deoxyribonucleic acid (DNA), is transduced in the form of ribonucleic acid (RNA), into its final working form of protein. The central dogma can be viewed as the biological equivalent of a three-tier client-server architecture. This applies not only to how cells interact and exchange information, but also to how they are organized spatially.

DNA represents the data layer, wherein the organism’s vast instructions are stored and accessed within the nucleus. Changes at the level of transcription and RNA levels, represent the equivalent of functional process logic: this is the channel through which approximately static information at the level of DNA in connected with dynamic information about local environment and cellular state communicated through proteins.
Thus the interface level is mediated through proteins, which communicate what is happening outside of the nucleus back to the information processing center.

Although this analogy is limited in its usefulness, it does make a striking point: all biological systems demonstrate properties of well-engineered systems. This turns out to be true at a number of levels, and will become evident later in a variety of contexts, such as error and attack tolerance in scale-invariant networks and modular organization of biological functions.

We will explore each successive level of the central dogma to determine where is most appropriate point to begin the search for higher-level modular structure. And finally, there will be a brief discussion of additions to the central dogma that provide additional insights leading to new experimental and therapeutic strategies.

2.1.1 Central Dogma I: Deoxyribonucleic Acid (DNA)

As mentioned above, there are many caveats, but the Central Dogma provides the core framework for how biological systems turn information into function. We begin with the cellular information repository: DNA. It will require a rather lengthy digression to establish some important properties of how different organisms organize their DNA, which will be important for motivating later problems. Specifically, the organization of the genome as we understand it does not provide the information for determining the larger building blocks of function. Thus, we will have to look beyond DNA to find modular structure in the eukaryotic systems of interest.

Cells maintain their genetic information in the form of deoxyribonucleic acid (DNA), which is analogous to storage of information in computers, though biological
systems represent this information in base-4 rather than binary. DNA is a long, polymeric molecule built out of a repeating unit consisting of a combination of phosphate groups, deoxyribose (a five-carbon sugar), and a nitrogenous base.

DNA provides the blueprint for the entire organism, and is the center of information processing in a cell. The complete assortment of a given organism’s DNA is referred to as its genome. Historically, and in the context of the Central Dogma, the genome was meant to refer to the complete set of genes (pieces of DNA that provide the instructions on how to construct proteins). As we will see later, genes account for only a very small fraction of the DNA in higher organisms, and are only one of many important functional elements in the genome. The genomes of prokaryotes consist of one long circular DNA molecule, though they can often contain more. Eukaryotes, on the other hand, contain several linear “chromosomes,” wherein each chromosome has a prespecified set of genes on it. Near the middle of each chromosome is a centromere, an area of the chromosome that contains no genes, and provides a point of attachment for physical manipulation of the DNA.

The simplest functional unit of DNA is the gene: a short piece of DNA (relative to the size of the genome) that provides the information for a single protein. Initial investigations demonstrated that genes involved in similar functions were located close to one another along the DNA, in so-called “operons.” The most famous of these is the lac-operon, which was characterized by Jacob and Monod. The lac-operon controls a bacterium’s ability to switch carbon sources from glucose, its preferred carbon source, to lactose. This happens only when glucose is absent and lactose is present. For decades after these groundbreaking discoveries, people
continued to find similar operonic structures throughout prokaryotes\textsuperscript{17}. This suggests that the prokaryotic genome is modular organized, with an easy mapping to function. Unfortunately, the modular genome of prokaryotes is not maintained across kingdoms.

In eukaryotes, there is still no evidence of genes being grouped within the chromosome by function; more correctly, there is no demonstration that genes are more functionally related to their neighbors than would be expected at random. These observations become more important, as they indicate that knowledge of the genome in isolation is insufficient to decipher the functional organization of an organism.

Understanding these findings, or attempting to rationalize them, relies on distinguishing how eukaryotes and prokaryotes pass on their genes during sexual and asexual reproduction. Before discussing the evolutionary benefits and drawbacks of sexual/asexual reproduction, it is necessary to make a few remarks regarding the fitness of an organism as a function of genetic diversity.

Laboratory studies of genetics are often constrained to populations that are virtually, genetically identical (inbred)\textsuperscript{2,107}. The only variation within laboratory populations, be it bacteria, yeast, or mice, is usually the gene or genes under investigation. Natural populations cannot exist under such circumstances. Incontrovertible evidence has demonstrated that populations must maintain genetic diversity to survive over evolutionary timescales of thousands to millions of years. The simplest toy example of this is the evolutionary bottleneck: a harsh stimulus to a population, wherein only select individuals survive dependent on their possession of a specific trait, which is most often rooted in their genetics. If the entire population has the
same genetics, they will eventually encounter a stimulus that they cannot survive. Their odds increase substantially if the population is genetically diverse.

Evidence of these bottlenecks (strong negative selections) is difficult to measure because there is no way to capture the genetics of the subpopulations that don’t survive. However, the benefit of genetic diversity can be appreciated in a similar case: a new genetic variant gives a slight increase in fitness to those that possess it. One of the clearest examples of this is lactose-tolerance. Approximately 10,000 years ago humans in Northern Europe, and possibly elsewhere, acquired the ability to tolerate lactose and hence, drink cow’s milk. The nutritional benefit for otherwise undernourished people was enormous, and lactose-tolerance has since spread such that lactose-intolerance is the less frequent phenotype. The genetic diversity that gave rise to lactose-tolerance was certainly beneficial to the individuals that it first occurred in, and the populations that it has since spread through.

A slight oversimplification can be concluded: an increase in genetic diversity is beneficial to the survival of a given organism at the population level. Although we will not explore population dynamics and caveats of this statement, we will be interested in how genetic diversity is manifest in the organisms of interest, and the interplay between genetic diversity and genomic organization.

Relative to asexual reproduction, sexual reproduction has been shown to maintain a much higher level of genetic diversity in populations. The clearest demonstration of this, though not direct, is the finding that species that reproduce asexually have much shorter times to extinction. This is not meant to indicate that species that reproduce primarily asexually cannot survive, but that they are at a disadvantage and must
compensate in other ways. More importantly, species that reproduce asexually tend to organize their genomes in a fundamentally different way from those that reproduce sexually.

Organisms, such as many forms of bacteria, that rely primarily on asexual reproduction, tend to organize pieces of their genomes into functional units: often times, a set of genes sufficient for a specific function will exist on a plasmid (small piece of circular DNA) in isolation. This allows that piece to be copied by itself and passed to neighboring bacteria. Bacteria possess three mechanisms for this type of genetic transfer: conjugation, transformation, and transduction\(^8\). It is important to note that none of these mechanisms constitute sexual reproduction, though conjugation is often mistakenly deemed sexual reproduction. Additionally, all of these mechanisms allow DNA from one bacterium to be transferred to another bacterium, though the medium of transfer is different (pilus, environment, and virus, respectively). For later comparison to sexual reproduction, the fact that the odds of successful transfer are extraordinarily low, certainly less than 1:1000, determines the critical difference.

In contrast, sexual reproduction produces offspring that are mixtures of their parents, and contain exactly equal amounts of parental genetic contribution. From this, if we were to watch how quickly a new gene introduced into a population becomes widespread, the disparity is clear. In a sexually reproducing population, half of the offspring of the first cell carrying the new gene would possess it. For an asexually reproducing population to produce similar levels of transmission requires the first cell to make thousands of copies and attempt transmission thousands of times, which is absolutely infeasible especially under natural conditions of nutrient restriction – the
energetic cost of replicating DNA requires allocating most of an organism's metabolic reserves. Building off of these observations, it is possible to draw the conclusion that organisms that reproduce in distinct ways require different genomic organization to maximize fidelity in the face of different errors.

Just as it has been noted that the genomes of bacteria contain many small functional modules, there are characteristic features of eukaryotic genomes that provide a contrasting picture of eukaryotic genomic organization. As we will see, the pressure on eukaryotes is to maintain critical functions in the midst of potential errors introduced in the process of generating genetically diverse offspring.

One common technique in genetics is the knocking out of a single gene. This allows the function of the gene to be explored by seeing how the organism functions, or doesn't, in the gene's absence. After the S. Cerevisiae (budding yeast) genome was sequenced, it became possible to systematically knock out each gene in parallel. A consortium, headed by groups at Stanford University, completed this task in 1999. This was the first opportunity to determine on a large-scale, which genes were absolutely essential for a cell to function.

Years of knockout data in aggregate, the combined wisdom of several hundred research groups world-wide, projected the fraction of essential genes in the genome to be greater than one half. In reality, ~1100/6000 yeast genes were required for function. Though the disparity between the estimate and the reality is explained by a simple ascertainment bias, the reality is still quite striking: under normal laboratory growth conditions, yeast cells do not require approximately 80% of their genes. More
interesting is how essential genes are distributed throughout the genome, which turns out to be far from random.

Before a eukaryote can pass its DNA on to its progeny, it must choose which copy of each chromosome it wants to pass on. Although the copies of the chromosomes that it maintains for its own function represent one from each parent (homologous chromosomes), the chromosomes it passes to its offspring will be a mixture of the two. This is accomplished by recombination, wherein each pair of homologous chromosomes sustains several breaks in identical positions and are then rejoined to produce a new mixed pair. There are two important points to keep in mind: 1) recombination has a very high rate of doing damage close to the region where the breaks are made, and 2) studies have demonstrated that recombination happens at different rates at different points in the genome. Not surprisingly, it was discovered that essential genes tend to cluster in areas where recombination rates are lowest, especially near the centromere, where due to physical constraints, the recombination rate approaches zero.

This suggests that the genomes of eukaryotes are organized to, among other things, reduce the risk of damage to essential genes in the process of generating genetic diversity through recombination. There are many other questions that naturally arise: Are genes for which diversity is beneficial localized to high-recombination areas of the genome? Do different organisms have different sets of essential genes? Not surprisingly, both of these questions have been breached, and for the most part answered, as we discuss below.

Just as some genes must be copied and maintained faithfully, some genes benefit an organism through their diversity. Undoubtedly the best examples come from
pathogens, and the immune systems meant to combat them. Pathogens and immune responses are locked into an unending battle: pathogens develop new attacks by changing their molecular repertoire, and immune systems seek to identify and destroy new threats. Though both systems have ways to engineer genetic variation, whether to avoid or detect, both are partially reliant on their inherited genetic diversity. Thus, after sequencing of the human genome, wherein it became possible to measure genetic variability across the genome, it was not surprising to find that the locus (position in the genome) with the greatest variability was the Major Histocompatibility Locus (MHC)\(^5\). The MHC locus is critical both for the identification of pathogens by the immune system, and the distinguishing of self from non-self, so as not to cause an autoimmune response. Sequencing has further shown that loci such as the MHC locus change more rapidly than areas that are thought to be neutral from a fitness point of view, indicating that genes whose diversity is key, reside in areas of the genome with high recombination rates.

Unlike yeast, with its extraordinarily facile genetics, higher eukaryotes provide enormous technical challenges in the knocking out of genes. In the last two years, the refinement and expansion of RNA interference (RNAi) techniques have led to the possibility of exploring essential gene functions beyond the simplest of model organisms. Though this data is not available yet, there is another direction from which to approach the question of universally essential genes.

As organisms evolve over millions of years, speciate from common ancestors, and diverge in their genomes, genes are gained and lost. One would expect that genes that are essential to the ancestral organism would be very difficult to lose, and would require gross changes at the organismal level to allow their function to become inessential.
leading to the possibility of their loss. Conversely, by looking at genes that are present across different species we can deduce how long these genes have been around. Since these timescales can only be resolved as far back as the last common ancestor, this provides a conservative lower bound for the lifetime of a gene. Again, by comparing essential and non-essential genes in yeast, we find that essential genes are far more broadly conserved. Given that further diverged species have much older common ancestors, it is exceedingly likely that a very large fraction of essential genes have been successfully maintained for millions of years. More importantly, most of the genes essential to yeast are present in virtually every eukaryote. Although these genes could work in completely distinct functions, it turns out they are clustered into distinct primordial functions. We will return to this later.

In total, this lengthy digression has suggested that the organization of elements at the level of DNA is built to maximize evolutionary fitness - simultaneously maximizing fidelity and diversity - within the constraints of how a given organism reproduces. Unfortunately, since eukaryotes do not provide functional modularity at the level of the genome, such as operonic structure in bacteria, their genomes provide little information about the functional and biochemical interactions between their constituents. This “parts list,” though not helpful for determining functional structure, will provide a starting point by enumerating the building blocks. And though bacteria may seem a logical starting point, due to the functional information encoded in their genome, operons do not provide a picture of higher-level (super-operonic) interactions, in addition to being inextensible to eukaryotes. Additionally, as the eventual goal is to understand the organization of human
cells to provide insight into mechanisms of disease, bacteria are not an appropriate system. Thus, we next consider the second step in cellular information flow: RNA.

2.1.2 Central Dogma II: Ribonucleic Acid (RNA)

As DNA provides a cellular blueprint, RNA provides a transducer for bringing those instructions toward a functional form. RNA has virtually the same repeating polymeric structure as DNA. In fact, the polymeric units of DNA and RNA differ at only a single atom. RNA is made in a process wherein it is copied directly from DNA, called transcription. Though this discussion will focus primarily on only one type, RNA molecules serve many functions.

The three earliest discovered, and most important, types of RNA are ribosomal RNA (rRNA), transfer RNA (tRNA), and messenger RNA (mRNA). rRNA, synthesized in a special part of the nucleus, the nucleolus, provides one piece of the machinery required to transcribe DNA into RNA. tRNA, as we will see shortly, is required in the final step of the Central Dogma for conversion of mRNA into protein. mRNA, which will be our focus, is the photocopy of DNA that is turned into protein, the functional machine of the cell. For many years, these molecules were thought to be the extent of cellular RNA. However, two major discoveries toppled this belief, the second of which demonstrates one of the substantial shortcomings of the Central Dogma.

While deciphering the mechanics of transcription, especially as part of rRNA synthesis, it was shown that there are a number of small non-coding RNAs in the nucleus that are required for various tasks, including: rRNA construction, gene regulation, and splicing. Though these findings demonstrated a new set of molecules were necessary for
a fundamental cellular process, they did not challenge or enlarge our view of how cellular activities are conducted or regulated.

In 1999, Fire and Mello$^{42}$ made the astonishing discovery that small artificial RNA molecules (small-interfering RNAs or siRNAs) could change the expression of genes there were complimentary to. More surprising, upon subsequent investigation, was the finding that most eukaryotic cells had the machinery to bring about such an effect. This suggested that this method of gene silencing was another mechanism by which cells could regulate gene expression. This led to a simultaneous hunt for both the endogenous RNA molecules used in this process, and the mechanism by which genes were silenced using siRNAs.

Shortly after Fire and Mello’s breakthrough, it was shown that cells produce a set of small non-coding RNAs called micro RNAs or miRNAs that were virtually identical to siRNAs. Several of these genes had been discovered in previous genomic analyses, but were mistaken for processed pseudogenes – genes that are no longer used by an organism, but remain in the genome as evolutionary fossils. Subsequent analyses have shown that there are many miRNAs in the genomes of higher eukaryotes, and the hunt for new classes continues$^{109}$.

As miRNAs were being discovered, the machinery behind RNA interference (RNAi) was also being worked out. It was shown that a number of uncharacterized protein complexes worked together to silence genes post-transcriptionally$^{109}$. Since these genes are not known to be involved in other processes, searching for them in all sequenced genomes provided a quick measure of how prevalent RNAi is. It turns out that all higher eukaryotes, and a number of simple multicellular eukaryotes, possess the
machinery for RNAi. In most of these systems, RNAi has since been shown directly to be a functioning system critical to the regulation of many functions in these organisms. The finding that RNAi acts at the post-transcriptional level flew directly in the face of the Central Dogma and will be critical to expanding understanding of the many facets of gene regulation.

mRNA molecules are made as needed, and provide the first glimpse into how cells respond dynamically. If RNA is regarded as a photocopy of the instructions encoded in DNA, these photocopies are made for each gene as the gene is needed. In bacteria, one RNA molecule may contain photocopies of multiple genes (polycistronic mRNA) that are contiguous in the genome, part of the same operon, and functionally related. Polycistronic mRNA's provide an extremely efficient means for bacteria to produce entire functional units at once. In contrast, eukaryotes make one RNA molecule per gene (monocistronic mRNA). This finding is not surprising, given that neighboring genes in the genome are not expected to be functionally related, and hence should not have coupled transcription.

Unlike cellular DNA, which consists of a small number of molecules that will remain unchanged throughout the life of an organism, hundreds of thousands of RNA molecules are being made and degraded every second in a living cell. This tremendous turnover can be used to capture inordinate amounts of information: at a small-scale it provides a proxy for how much of a given gene is required and whether that level is increasing or decreasing, and at a large-scale, it provides a picture of which genes are required in tandem. Looking toward a means to understand how genes work in concert to enact cellular functions, mRNA levels provide a first step: by looking across many
conditions and environments, and looking at which genes tend to have similar patterns of expression, it is possible to start grouping genes based on similarity. Given the ability to see cellular dynamics through parallel RNA measurements, RNA would seem an ideal point to look for cellular organization and modularity. However, any conclusions based off of RNA measurements are limited to the conditions in which they were measured. This context dependence is extremely useful for a condition of interest, but more general conclusions require a context independent measurement (e.g. genome sequencing is context independent).

2.1.3 Central Dogma III: Protein

The final step in the Central Dogma is translation, wherein the instructions photocopied onto RNA are converted to protein. Proteins are the workhorses of cell, executing all of the instructions encoded in DNA: synthesizing ATP from available nutrients, detecting and responding to extracellular signals and stimuli, and duplicating the DNA and organelles required to produce offspring. Quite in contrast to the regular homogenous structure of nucleotides (DNA and RNA), proteins have an incredibly rich set of properties and capabilities that will provide the majority of subsequent material.

Currently, two problems dominate the world of protein biology and chemistry: 1) the correct prediction of a protein’s shape based only on the DNA sequence that encodes it, and 2) elucidating how and which proteins interact with one another to build complicated functions. Unlike DNA and most forms of RNA that will be discussed here, the three dimensional shape of a protein is absolutely critical to its function. Loss of a protein’s ability to maintain the correct three-dimensional conformation is at the heart of
a variety of diseases. Though the structure, and ultimately the function, of a single protein is fascinating in its own right, most biological processes require tens to hundreds of distinct proteins. As a result, the last decade has seen the birth of a new field aimed at determining how an organism’s collection of proteins, its proteome, uses these molecular machines to accomplish cellular tasks. These problems more so than those previously mentioned, will yield a substantial picture of biological function.

Before beginning a discussion of the protein folding problem in its many forms, it is necessary to introduce some nomenclature. Proteins are composed of one or several peptides, which are in turn linear polymers of amino acids. Amino acids, like the mononucleotides that constitute DNA and RNA, have inherent polarity that gives rise to polarity in the larger peptides; their ends are defined by the terminal group at each end: the amino terminus and the carboxy terminus. Each amino acid consists of a backbone (common to all amino acids) and a side chain (specific to each amino acid).

Compared to the four bases that can be used to build nucleotides, peptides can be built out of 20 standard amino acids and a handful of other uncommon and unnatural amino acids. Recently, it has become possible to incorporate unnatural amino acids into mammalian cells, toward engineering completely novel biological functions as well as opening up uncharted capabilities in synthetic chemistry\textsuperscript{110}.

Protein structure can be derived at several levels. The most basic, aptly named primary structure, is simply the listing of a peptide’s amino acids, typically written starting at the amino terminus. The primary structure, also called the protein sequence provides the starting point for investigations into a protein’s shape. Within each amino acid’s backbone are a hydrogen bond donor and a hydrogen bond acceptor. Such
hydrogen bonding groups on each amino acid leads to two types of regular hydrogen bonding structure: alpha helices and beta sheets. And additional interactions that help a single peptide maintain its three-dimensional shape are grouped under the umbrella of tertiary structure: hydrophobic interactions, electrostatic interactions, disulfide bridges, etc. The final level of protein structure is quaternary structure, which involves the shape of multiple peptides spatially woven together to build a protein. The ability to connect these disparate levels of description is the heart of the protein-folding problem.

The protein-folding problem is to be able to predict a protein’s complete three-dimensional shape from its primary structure. Though simple in its statement, solution to this problem has eluded scientists since its postulation in the wake of the pioneering work by Christian Anfinsen and colleagues in the 1950’s. Though there are many clues to solving different levels of structure, the combinatorial complexity of possible solutions prevents accurate reconstruction from piecemeal solutions. Despite these challenges, the extraordinary potential benefits in solving the folding problem make it as interesting, if not more so, than it was fifty years ago.

Due to the difficulty of solving the folding problem directly, it has been recast into slightly simpler subproblems, and many of the potential benefits of a solution have been realized through other channels. The most important benefits would come in three primary areas: protein design, functional annotation of genomic sequences, and rational drug design.

The ability to design proteins from scratch opens many possibilities for hijacking pieces of biological machinery and co-opting it for other purposes. A few keys facts make this even more appealing: 1) proteins with similar sequences have similar three
dimensional shapes, 2) only a small fraction of protein sequences will maintain stable folds, and 3) the universe of naturally occurring protein folds is much smaller than expected given the universe of possible sequences. It is worth mentioning that many folds provide enzymatic scaffolds, on which many diverse functions can be built with very small changes in sequence. Also, it is should be noted that the majority of enzymes (proteins capable of catalyzing chemical reactions) fall into a very small set of structural classes, indicating that most enzymes are built on molecular scaffolds.

The above findings underscore the logic behind using solutions to the folding problem for the elucidation of a wider variety of biological questions. The common mapping connecting the genetic information encoded at the level of DNA and primary structure, to the more complicated functional picture is as follows:

Sequence → Structure → Function

Thus, if one could solve the folding problem, they would be able to accurately construct the first mapping. Though, as mentioned above, a given structure can support many molecular functions, knowledge of a protein’s structure still dramatically restricts its universe of possible functions.

Taken together, the restricted universe of naturally occurring protein structures and the structural similarity of neighbors in sequence space suggest that attempts to design proteins should start by searching areas of sequence space already shown to be rich in function through naturally occurring proteins. This is not exclude the possibility of folds and structures undiscovered by nature, but a more expeditious route to protein design. Perhaps the simplest way to mine these areas of sequence/structure space, would be to start with an enzymatic scaffold, and look for neighbors in sequence space that
might produce similar stable folds with different enzymatic functions. This gives rise to the Inverse Folding Problem: given an arbitrary sequence, does it fold into a given structure of interest? When combined with a functional screening strategy, this could provide an extremely efficient means to combine rational and combinatorial protein design. Given the inhomogeneous distribution of enzymatic folds, solving the Inverse Problem for such a reduced set would provide the vast majority of benefit of solving the more general problem. This is supported by observing the distribution of naturally-occurring enzymatic folds: the top one percent of enzymatic folds account for more than ninety percent of enzymes.

Another problem that would benefit from solution of the folding problem, is functional annotation of genomic sequences. As genomes continue to be sequenced, it becomes increasingly important to find computational means of determining each newly sequenced gene’s function. In the pre-genomic era, sequences were usually only determined for genes or proteins that were being functionally characterizing, so that the annotation and sequencing rates were matched. Nowadays, the rate of sequencing has grossly outpaced experimental functional annotation. One way to predict function would be to predict the structure, which would provide a very good indication of function. Fortunately, most of the functional information provided from structural similarity can also be gleaned from sequence similarity, alleviating the need for accurate structure prediction for these purposes.

The final area that was espoused to derive benefit from solution to the folding problem was rational drug design, wherein small-molecules could be designed to provide a perfect lock-and-key fit to the protein target of interest and modify its behavior. One
alternative to computational prediction of protein structure is experimental determination through X-ray crystallography. Crystallography, being both difficult and extremely time-consuming, was viewed as too inefficient to apply to all drug targets of interest. As experimentally solved structures have since saturated the space of targets, the need for computational determination has decreased. One application however still remains: allele-specific targeting.

In many diseases, most prominently cancer, a specific gene will be dysregulated in subtly different ways in different patients. As therapeutics evolve to not only target a specific protein, but specific alleles (slight variants in sequence) of that protein, the ability to use a crystallography-based method diminishes. Daunting as the prospect of solving the equivalent folding problem may be, it is unnecessary to predict the entire structure as the allelic variant differs from the sequence of the solved structure in only a handful of amino acids. The combinatorial complexity of this problem is well within the bounds of current computational methods, and has allowed pharmaceutical companies to effectively design next-generation inhibitors with relative ease. To date, the best example is the set of Aurora Kinase inhibitors now making their way through clinical trials.

The other major problem of interest involving proteins is how they are organized in the proteome to accomplish cellular functions. Progress in understanding the organization of the proteome would shed light on a number of areas already mentioned: determining functional constraints from an evolutionary point of view, functionally annotating genomic sequences, and discovering new therapeutic targets. Networks of interacting proteins provide a perfect venue in which to look at cellular organization: new technologies can systematically measure protein-protein interactions across an entire
proteome, the proteome is the cellular layer at which functions are achieved, and the measurement technologies are context independent.

As searching for meaningful structure in the proteome is the primary focus of this work, and will be introduced in greater detail later, only its relation to other related problems will be covered here. Though solution to the folding problem and elucidation of the proteome are often viewed as entirely different problems, it is worth noting that they have much to offer one another: solution of the folding problem would provide a completely novel avenue to interrogate protein-protein interactions, and demonstration that proteins interact, says volumes about their structure building upon knowledge of their protein sequence.

Despite dramatic technological innovations in high-throughput assays for detecting protein-protein interactions, only the simplest of conclusions can be drawn for each interaction in the form of a yes or no answer. Additionally, these binary answers come with substantial noise in the form of false-positives and false-negatives. This leads to a completely static picture that does not factor in changes in protein structure due to the presence of allosteric effectors or additional proteins. If it were possible to determine the tertiary and quaternary structure of the putative interactors, it is reasonable to believe it would be possible to computationally search a much larger set of interactions under a variety of conditions. Specifically, it has been shown in a number of contexts that groups or complexes of proteins reorganize dynamically to adjust their function. This type of information is difficult to capture without either measurements of every cellular configuration, or the possible structural and interactive states proteins can occupy. With
knowledge of protein structure dynamics, it would be possible to construct a picture of changing intermolecular interactions, drawing closer to the reality within a living cell.

As will be discussed later, interacting proteins enforce substantial evolutionary constraints on one another. This suggests, that one could look across species for pairs of homologous proteins that are known to interact. For example, suppose we have two proteins - A and B - that are known to interact. If we were to compare A to all of its homologues that do not interact with homologues of B, versus comparison to all of those that do interact with a homologue of B, it would be possible to construct a loose partition of the sequence of A into pieces key to the interaction with B, and those uninvolved. Notably, this approach could potentially find residues that are far in space from the interaction surfaces of A, that are still critical to the structural basis of interaction.

2.1.4 Central Dogma IV: Limitations and Extensions

Though the Central Dogma provides a picture of how cells turn information into function, it is limited in several regards. The clearest illustration of these limitations comes in the form of exceptions to, or missing links in the Central Dogma. Amongst the many that have recently come to light, three show signs of having tremendous functional relevance: 1) DNA packaging at the level of chromatin, 2) small-molecules, and 3) miRNAs. As miRNA’s were covered earlier, we focus on the first two: chromatin biology, and the chemical biology of small-molecules.

For decades, evolutionary theories have assumed that all of the information passed from parent to child is transmitted through the base sequence of DNA. For the most part, this only allows the information of the ancestry of the parents, plus any
mutations accrued during the process of copying the parents' genetic material. Just as humans attempt to pass wisdom to their children through education, cells pass information to their descendants through a number of so-called epigenetic mechanisms. To date, the most studied is how DNA is packaged, and how this dynamic process can encode substantial information about the environment the parents inhabit. In retrospect, it makes intuitive sense that organisms would want to pass on slight genetic optimizations that helped them adapt to their environment. The combination of mutation and recombination, though they provide benefits at a population level, do so over large timescales and do not provide benefits to individuals relative to their peers. Proving the evolutionary significance of this argument is beyond the scope of this work.

In brief, the DNA of eukaryotes is packaged by wrapping it around globular proteins called histones (the so-called balls on a string model); the combination of DNA and histones is referred to as chromatin. Due to physical constraints, when DNA is wrapped tightly around a single histone octamer (the functional unit of a histone is a hetero-octamer), called a nucleosome, it is inaccessible for transcription, so-called heterochromatin. By contrast, when DNA is stretched between two adjacent nucleosomes, it is accessible for transcription, so-called euchromatin. Though not completely understood, the movement of histones along the DNA is governed by a series of enzymes that add methyl and acetyl groups to fixed positions on the histones, and sequence specificity to the DNA. Though much of the histone code remains to be deciphered, as it appears to be different in different organisms. Substantial evidence has demonstrated the ability of pass epigenetic information through these mechanisms. The area that promises both to benefit most from understanding these mechanisms, as well as
lead to the most substantial medical benefit, is the stem cell biology, which may very well be controlled almost exclusively at the level of chromatin.

Since the advent of molecular biology in the 1970’s and 1980’s, the relative importance of small-molecules in biological systems has been obscured. Moreover, the Central Dogma, in an attempt to be pristinely general, ignores the fact that many cellular systems are influenced both internally by metabolite levels, and externally by signaling molecules secret by neighboring cells. The rich biology of these overlooked molecules cannot be ignored, especially as they provide a source of incredibly innovative chemistry with unending clinical potential. The veracity of this claim has been repeatedly demonstrated through the dominance of natural product chemistry in the development of new therapeutics; the majority of current treatments in immunosuppression and oncology are derived from natural sources\(^8\). It is important to understand the diverse roles these molecules play in their native milieu, and how this distinguishes standard metabolites from stringently derived chemical probes.

Just as many genes are conserved broadly across eukaryotes, so are the main metabolites involved in the synthesis and breakdown required to sustained life: conversion between and storage of myriad chemical fuels, and synthesis of protein and nucleic acids, and amphoteric molecules. In fact, the majority of these molecules are not only conserved across eukaryotes, the vast majority are conserved across prokaryotes and hence all living organisms. However, a small set of molecules have been found that are not used for the above purposes, and are more important for competition between microbes for common resources\(^8\).
Unlike laboratory experimental conditions, microbes in the wild are constantly under stress due to lack of nutrients and other harsh environmental stimuli. To survive in these conditions requires not only optimizing their internal functions, but also outcompeting their neighbors for available resources. To this end, many organisms have developed a series of small-molecules meant to target critical proteins in neighboring organisms. However, since synthesis of such molecules is extraordinarily complicated, and cannot be adapted to different neighbors dynamically, the target of interest must be widely conserved across species. As mentioned above, such widely conserved protein targets are likely to be important in primordial pathways across organisms, and are highly enriched for therapeutic targets in human disease.

Limited research into the small-molecule biology of a handful of organisms has led to the discovery of a number of small-molecule inhibitors that have revolutionized the natural history of a number of diseases: FK506 and rapamycin in immunosuppression, and taxol, doxorubicin, and 17-AAG (derived from the naturally-occurring geldanamycin) in the treatment of cancer. As the number of species of microbes on the planet is estimated to be in the billions, this provides an unending resource for therapeutic development that far exceeds that generated by human ingenuity.

2.2 Conclusions

By exploring each step in the Central Dogma, as well as some exceptions/extensions, this section was meant to motivate the choices and assumptions made later. Most important amongst those is the choice of protein-protein interaction
(PPI) networks as the system in which to investigate modularity in biological systems. Ultimately, each level of information processing has an important contribution and affects cellular organization and function at some level. In the future, as these different layers of cellular organization become clearer, it will be possible to integrate many different data types into functional models.
3. Discovering and Validating Modular Structure

The primary goal of discovering modularity, was to use large-scale structure to better understand high-throughput datasets. Unlike datasets of old, new technologies allow parallel measurements across thousands of genes. For instance, sequencing a single gene could take days to weeks, allowing plenty to time understand each gene being sequenced. Currently, it is possible to not only sequence hundreds of genes in an hour, it is possible to measure the expression level of every gene in real time.

With these novel capabilities comes a new set of challenges: simplifying massive datasets, filtering out noise, and transforming the filtered product into biological insights. As a step toward overcoming these challenges, we will construct an approach to: 1) organize biological systems according to their inherent structure, 2) use this information to contextualize current publicly-available datasets (such as Lum et al\textsuperscript{43}), and 3) assess whether the additional information has contributed to our understanding of those datasets. Specifically, we will attempt to partition the set of proteins in an organism into a small set of modules. Each module will contain a set of proteins that interact with one another more than with proteins outside of the module. Using these modules, we will assess whether they accurately capture the partitioning of cellular function and can explain data sets of interest.

It should be noted that the optimization of any one part of the process is secondary, and necessary only inasmuch as it aids in assessing the feasibility of the whole process. Thus, though we will attempt to justify choices along the way, the results from the whole process are the primary goal.
To find modular structure, we begin by specifying the approach and the methods to validate our findings. Again, it is worth noting that there are an endless number of approaches we could have chosen, some of which may provide better results, but we chose this approach for its simplicity. Moreover, we aim to choose a method that will isolate the phenomenon of interest, while not being unduly sensitive to the specifics of the approach.

All of the subsequent investigations will be focused on the concept of modularity, which will be measured in graphs based off of biological data. Our intent is to find modules in a systematic way based off of the topology of the graph. The modularity can be loosely defined as: given an edge of interest \( E \), and spanning nodes \( A \) and \( B \), what is the intersection of the sets of neighbors of \( A \) (\( N(A) \)) and \( B \) (\( N(B) \))? There are caveats to this definition that we discuss below. Optimally, after we construct modules by partitioning the set of nodes, two nodes in the same module would have many neighbors in common, and far fewer neighbors in other modules.

In order to find edges spanned by nodes with such properties, we attach a weight to each edge using an edge metric that is a function of \( N(A) \) and \( N(B) \). Once the edge metric has been applied to each edge, the graph can be decomposed into disjoint subgraphs by removing edges with successively increasing weights (where heavier edges correspond to areas of higher local density), leading to progressively smaller, denser subgraphs. Thus, the tree (dendrogram) produced by hierarchical decomposition (or hierarchical clustering in reverse) will have a root node corresponding to the original graph, and each child node is a proper subgraph of its parent. We will refer to this structure as the "module tree" for each organism.
Ultimately, all subsequent analyses hinge on the success or failure of the edge metric. Thus, our notion of modularity will have to be further refined: the modularity should measure how many neighbors nodes A and B share, relative to the degree of A and the degree of B. Instead of choosing the appropriate corrections for the degrees of A and B, we propose four modularity metrics to be compared:

- **Jaccard**: \( \frac{I(N(A),N(B))}{U(N(A),N(B))} \)
- **Meet/Min**: \( \frac{I(N(A),N(B))}{\min(|N(A)|,|N(B)|)} \)
- **Geometric**: \( \frac{I(N(A),N(B))^2}{|N(A)||N(B)|} \)
- **Hypergeometric**: \( \sum C(K,i)C(N-K,n-i)/C(N,n) \quad (k <= i <= n) \)

Where \( I(N(A),N(B)) \) is the intersection of \( N(A) \) and \( N(B) \), \( U(N(A),N(B)) \) is the union of \( N(A) \) and \( N(B) \), and \( C(a,b) \) is the binomial coefficient “a choose b.” From a graph theoretic perspective, each of these metrics is querying how many triangles a given edge is part of, and how significant that is giving the degree of the spanning nodes. We investigate this further below.

It should be noted that if the graph from a given organism contains multiple disjoint subgraphs, there will be multiple trees for that organism. However, in all cases under consideration, > 90% of the graph is in one connected component, and hence one tree. Additionally, the remaining disjoint subgraphs do not play a prominent role in our results, and can be ignored from this point forward.

After the module tree has been constructed in each case, we will validate it in two ways based on topology alone: 1) demonstrate that it contains modularity, and hence modules, that are not expected at random in a graph with an identical degree distribution,
and 2) show that the means used to derive it are robust enough to overcome the error in experimental techniques used to generate the graph in the first place.

As a final measure of the robustness of this technique, we use hand-curated annotations to assess the biological relationship to this topological procedure.

3.1 Density in Randomized, Inhomogeneous Graphs

Before the advent of technologies such as Yeast Two Hybrid (Y2H), it was assumed that the topology of biological networks was identical to that of a random graph. However, early Y2H results demonstrated that this was unlikely to be true. In fact, they were shown to be scale-invariant and have a power-law degree distribution. Recent evidence has called into question the veracity of a power-law degree distributions in biological networks by showing that equivalent sparse sampling from other topologies could give rise to similar results\(^{86}\). Though we will not directly use the scale-invariance, we will assume that the biological networks under consideration are extremely inhomogeneous at a variety of levels: some nodes have many more neighbors than others, some subgraphs are much denser than others, etc. This assumption necessitates the use of stringent tests for significance of any density findings, as they might be insignificant given the background degree distribution.

As a suitable comparison on which to base significance, for each graph, we have generated 1000 graphs with identical degree distributions that have been randomized as follows: given an edge X spanned by nodes A and B, and a second edge Y spanned by C and D, X and Y are removed and two new edges are added. The two new edges are created:
This procedure assumes that neither of the new edges was present before this step. On inspection, it is clear that this procedure does not change the degree of any of the nodes. To assure that each graph has been sufficiently randomized we used the following empiric procedure:

1. Graphs were randomized 10,000 times, and distributions of edge weights were compiled for each set of graphs (one set per starting PPI network from each species) under each edge metric.

2. Graphs were subsequently randomized 100,000, 1,000,000, and 10,000,000 times, with similar distributions collected at each step. All graphs were independently randomized starting with the original PPI network to avoid bottlenecking at each step and false indications of homogeneity at each step in randomization.

3. Distributions of edges weights under each metric were compared across randomization groups and to the original PPI network using a two-tailed Kolmogorov-Smirnov (KS) test. Stabilization of the average value of the KS test indicates convergence of the randomization process.

Since all of the metrics being applied in these analyses are querying the number of triangles completed by a given edge, it is important to know whether we expect any difference in triangle content between PPI networks and randomized networks. The following histogram shows the number of triangles found in the randomized graphs (the number of triangles in the corresponding original PPI network is listed in each caption).
Figure 3.1. Normalized histograms showing the number of triangles present in randomized networks derived from Worm (CE), Fly (DM), Human (HS), and Yeast (SC). The number of triangle in the corresponding PPI network is listed in each caption.

In the case of the Worm PPI network, the number of triangles is much smaller than that seen in randomized networks. Though this suggests that on average, there is less clustering than expected at random, there may still be specific places in the network that have greater modularity than expected. For this reason, we include the Worm network in further investigations. It should be noted that the Worm network is by far the smallest and sparsest network, which may provide an explanation for its strikingly different behavior relative to the other three networks. For Fly, Human, and Yeast the number of triangles is far larger, suggesting that there is significant clustering beyond what is expected at random.
One caveat to these results lies in the notion of joint-degree distribution across the edges. Moreover, randomizing according to our procedure may very well change certain other features of the topology, which indirectly affect the clustering. One example is changing the distribution of edges spanned by nodes with pair of degrees $X,Y$. If the PPI networks have a different joint-degree distribution across their edges, our randomization procedure might change this. To determine this affect, we constructed a second set of randomized networks that maintain the joint degree distribution. From an intuitive point of view, this procedure is identical to the above procedure, except that nodes $A & D$ must have identical degree, as must nodes $B & C$. This type of randomization provides a much more stringent test, as this second family of randomized networks is much closer to the original PPI network. As with the first family of randomized networks, we present a series of histograms counting the number of triangles in the second family of randomized networks:
Figure 3.2. Normalized histograms showing the number of triangles present in randomized networks derived from Worm (CE), Fly (DM), Human (HS), and Yeast (SC). Networks were randomized to preserved joint-degree distribution across the edges. The number of triangle in the corresponding PPI network is listed in each caption.

As before, the networks of Fly, Human, and Yeast show a significantly larger number of triangles in the PPI network than the randomized network. Using this type of randomization, the Worm PPI network now shows a slightly larger number of triangles, in contrast to the previous randomization scheme. The more important conclusion is that the other three networks show substantial clustering beyond that expected at random, even under the stricter normalization scheme of preserving joint degree distribution across the edges. We do not use this family of graphs for any further investigations. We make this choice due to the constraint imposed: in many cases, certain edge classes are
too small to be meaningfully randomized. Though it proved useful in this context, it
leads to inhomogeneous randomization across the graph.

We return to the family of randomized graphs preserving degree distribution, but
not preserving joint degree distribution. All randomized graphs were scored using the
four edge metrics. For a given graph and a given metric, an edge weight distribution was
collected. This distribution was compared to the edge weight distribution in the original
network using a two-tailed Kolmogorov Smirnov (KS) test. The following tables lists the
D-score and P-values from each comparison of interactome to randomized networks.

<table>
<thead>
<tr>
<th>Metric</th>
<th>C.Elegans</th>
<th>D.Melanogaster</th>
<th>H.Sapiens</th>
<th>S.Cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometric</td>
<td>0.20</td>
<td>0.14</td>
<td>0.25</td>
<td>0.20</td>
</tr>
<tr>
<td>Hypergeometric</td>
<td>0.22</td>
<td>0.16</td>
<td>0.26</td>
<td>0.21</td>
</tr>
<tr>
<td>Jaccard Index</td>
<td>0.16</td>
<td>0.10</td>
<td>0.25</td>
<td>0.18</td>
</tr>
<tr>
<td>Meet/Min</td>
<td>0.22</td>
<td>0.15</td>
<td>0.27</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Figure: D-scores from Kolmogorov-Smirnov Test Comparing Edge Distribution in PPI
Networks to Randomized Networks with identical Degree Distributions.

<table>
<thead>
<tr>
<th>Metric</th>
<th>C.Elegans</th>
<th>D.Melanogaster</th>
<th>H.Sapiens</th>
<th>S.Cerevisiae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geometric</td>
<td>0.016</td>
<td>0.007</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Hypergeometric</td>
<td>0.015</td>
<td>0.006</td>
<td>0.004</td>
<td>0.005</td>
</tr>
<tr>
<td>Jaccard Index</td>
<td>0.018</td>
<td>0.008</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td>Meet/Min</td>
<td>0.015</td>
<td>0.006</td>
<td>0.004</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Figure 3.3: P-values from Kolmogorov-Smirnov Test Comparing Edge Distribution in
PPI Networks to Randomized Networks with Identical Degree Distributions.

These tables suggest that the edge metrics are able to distinguish statistically significant
differences in network structure between PPI networks and randomized networks with
identical degree distributions. One weakness in the use of a KS test, is its lack of
resolution at the edges of the distribution. Specifically, the majority of edges have very
small intersections between neighbors of their spanning nodes. As a result, the KS test is
unable to make fine distinctions between the metrics. Although this makes the KS test
less helpful for choosing a metric, it provides an extremely stringent test for
distinguishing the original PPI networks and randomized networks. Thus we can conclude that all four metrics are able to distinguish the true PPI network from randomized networks, but we are unable to evaluate the metrics head-to-head.

In order to help choose an edge metric the following histograms further illustrate the differences in network density between PPI and randomized networks, as well as highlighting the strengths and weaknesses of each metric. Parameters for each network is included in the Appendix; this includes number of nodes and edges, and some of the experiments that made major contributions to each network.

Figure 3.4. Histogram of Geometric edge weights, from Worm (C.Elegans) PPI network and randomized networks.
Histogram Comparing Hypergeometric Edge Weights in the Interactome of C. Elegans To Randomized Networks

Figure 3.5. Histogram of Hypergeometric edge weights, from Worm (C. Elegans) PPI network and randomized networks.
Histogram Comparing Jaccard Edge Weights in the Interactome of C.Elegans to Randomized Networks

Figure 3.7. Histogram of Jaccard edge weights, from Worm (C.Elegans) PPI network and randomized networks.
Figure 3.7. Histogram of Meet/Min edge weights, from Worm (C.Elegans) PPI network and randomized networks.
Histogram Comparing Geometric Edge Weights in the Interactome of D.Melanogaster to Randomized Networks

Figure 3.8. Histogram of Geometric edge weights, from Fly (D.Melanogaster) PPI network and randomized networks.
Figure 3.9. Histogram of Hypergeometric edge weights, from Fly (D.Melanogaster) PPI network and randomized networks.
Figure 3.10. Histogram of Jaccard edge weights, from Fly (D.Melanogaster) PPI network and randomized networks.
Figure 3.11. Histogram of Meet/Min edge weights, from Fly (D.Melanogaster) PPI network and randomized networks.
Figure 3.12. Histogram of Geometric edge weights, from Human (H.Sapiens) PPI network and randomized networks.
Figure 3.13. Histogram of Hypergeometric edge weights, from Human (H.Sapiens) PPI network and randomized networks.
Figure 3.14. Histogram of Jaccard edge weights, from Human (H.Sapiens) PPI network and randomized networks.
Figure 3.15. Histogram of Meet/Min edge weights, from Human (H.Sapiens) PPI network and randomized networks.
Figure 3.16. Histogram of Geometric edge weights, from Yeast (S.Cerevisiae) PPI network and randomized networks.
Figure 3.17. Histogram of Hypergeometric edge weights, from Yeast (S.Cerevisiae) PPI network and randomized networks.
Figure 3.18. Histogram of Jaccard edge weights, from Yeast (S. Cerevisiae) PPI network and randomized networks.
Upon visual inspection, there are two striking features: 1) marked differences between species, and 2) substantially different strengths between the metrics. We handle each in turn.

Between the species, there is a substantial difference in modularity as measured by these four metrics. Specifically, in human (H.Sapiens) and yeast (S.Cerevisiae), there is a noticeably larger difference between the original PPI network and randomized, than in worm (C.Elegans) or fly (D.Melanogaster). Although this difference could be explained by true differences in proteomic organization, a far more likely explanation comes from how the networks were constructed: virtually the entire PPI network (>90%) for both worm and fly come from a single high-throughput Y2H experiment. In contrast,
the yeast and human PPI networks have much larger contributions coming from small high-quality datasets; one large Y2H contributes to the human PPI network, and four large high-throughput studies (two high-throughput Y2H studies and two high-throughput Mass Spectrometry studies) contribute to the yeast PPI network. Further details on the construction of these networks are included in the Appendix.

The Jaccard Index appears not to perform as well as the Hypergeometric, Geometric, and Meet/Min. Perhaps more interesting is the comparison between the Hypergeometric/Geometric and Meet/Min: the former tend to be stronger in separating out high density areas of the original interactome graph, whereas the latter is stronger at finding medium to low density areas. As our primary interest will be high modularity subgraphs, the Geometric and Hypergeometric appear better choices than the Meet/Min. From these analyses, the Geometric and Hypergeometric appear the best choices. However, the discrepancy forces us to consider the properties of both the networks and the metrics further.

To further dissect the differences between the original network and randomized network, we present the following colormaps that represent the different edge classes, defined by the degree of each spanning node of a given edge. The colormaps are constructed by determining how often an edge spanned by nodes with degree A and B, occurs both in the true PPI network and randomized networks. For each edge class (defined by a degree pair, representing the degrees of the spanning nodes), a distribution is constructed from the randomized networks. A Z-score is calculated for the original PPI network based off of this distribution.
Figure 3.20. Colormaps demonstrating the enrichment of edge classes in the interactomes of worm (CE), fly (DM), human (HS), and yeast (SC). The (x,y) coordinate in each plot correspond to the degree of each node spanning the edge class represented by the pixel at that point. The colormaps are limited to edges spanned by nodes with degree less than one hundred to highlight the regions of the colormap with greatest coverage.

Scores from the colormap were derived by comparing how often an edge with spanning degrees (A,B) occurs in the PPI networks under consideration, compared to the corresponding randomized networks. There are no prominent trends across the colormaps. For the most part, regions of the colormap have mixed coloring indicating both over-representation (green) and under-representation in the PPI network relative to randomized networks. One trend that is noteworthy is the slight enrichment for edges spanned by nodes of similar degree in the yeast (SC) network. As this trend is not observed in the other networks, its significance is unclear.
To evaluate the biases of each metric, we present four similar colormaps that indicate how each edge class scored relative to randomized networks, according to each metric. For example, for each edge class the average score of that edge class is computed for each network in the family of randomized networks. This distribution is used to normalize the score of the corresponding edge class in the original PPI network. It is important to note that these colormaps are used for exploratory purposes and should not be used to make strong conclusions.

Figure 3.21. Colormap depicting the distribution of edges scores according to a Geometric edge weight in the interactomes of worm (CE), fly (DM), human (HS), and yeast (SC). The (x,y) coordinate in each plot correspond to the degree of each node spanning the edge class represented by the pixel at that point.
Immediately noticeable in the colormaps is the difference in the distribution, which is appreciated by both the different background coloring of each colormap (these are edge classes that do not occur in the intersection of the edge classes between the original PPI network and the randomized networks), and the different ranges of accompanying colorbar scales. This indicates that the human (HS) and yeast (SC) networks are significantly higher scoring at virtually every edge class than worm (CE) or fly (DM).

Figure 3.22. Colormap depicting the distribution of edges scores according to a Hypergeometric edge weight in the interactomes of worm (CE), fly (DM), human (HS), and yeast (SC). The (x,y) coordinate in each plot correspond to the degree of each node spanning the edge class represented by the pixel at that point.

As with the geometric, the hypergeometric shows an enrichment for higher scores for all edges classes. Again, is especially prominent in the case of human (HS) and yeast (SC).
Figure 3.23. Colormap depicting the distribution of edges scores according to a Jaccard Index edge weight in the interactomes of worm (CE), fly (DM), human (HS), and yeast (SC). The (x,y) coordinate in each plot correspond to the degree of each node spanning the edge class represented by the pixel at that point.

The Jaccard Index again shows a substantially different background between networks, while preserving the same ordering of modularity across species.
Figure 3.24. Colormap depicting the distribution of edges scores according to a Meet/Min edge weight in the interactomes of worm (CE), fly (DM), human (HS), and yeast (SC). The (x,y) coordinate in each plot correspond to the degree of each node spanning the edge class represented by the pixel at that point.

Finally, the Meet/Min shows a similar background across species, while preserving order. Comparing the Meet/Min colormap to its earlier histogram suggests that it has enriched scores for low modularity, mid-to-high degree edges (the areas that show enrichment in the earlier histogram and the colormap). Unfortunately, this cannot be confirmed directly from these data.

In total these analyses do not directly affect the choice of metric, but they do suggest the next step in these types of comparisons: a more appropriate control method. Specifically, the histograms demonstrated that certain metrics, were able to find non-
random structure in less clustered areas of the graph (e.g. the Meet/Min), while others found non-random structure in more clustered areas of the graph (e.g. the Geometric and Hypergeometric). As was shown earlier, the edge classes in the PPI networks do not differ dramatically from those found in randomized networks with preserved degree distribution. Thus, to determine the scoring behavior of these metrics in more detail will require constructing families of networks that are randomized while maintaining not only the degree distribution, but also the joint degree distribution across edges. Posing these two questions together will separate questions about the topology at the level of degree distributions from modularity questions in the presence of, possibly, extremely non-random topologies. Unfortunately, there are technical limitations to this suggested approach, due to the inability to homogeneously randomize across a graph while maintaining joint-degree distribution. This is due to the lack of sufficiently many edges within a given edge class to allow changes in connectivity during the randomization process. The development of more general approaches in this direction may help to characterize the subtleties of these metrics.

To further assess the ability of these metrics to capture modularity in biological networks, we measure how robust they are to noise in the input data.

### 3.2 Noise Tolerance of Edge Metrics

All of the high-throughput technologies used to generate the data represented in the PPI graphs under investigation here, have substantial false-positive and false-negative rates. To validate any modular structure found in these graphs, it is necessary to demonstrate that mild to moderate changes in the underlying graph would not
dramatically alter the extracted modular structures. To model these aberrant data, we construct a family of graphs based on the original PPI graphs, with edges either added or removed. One shortcoming of this approach is that it treats the assessment of robustness of the approach to false-positives and false-negatives separately. To fully model the noise in these data would require both adding and subtracting edges simultaneously. However, for practical reasons we perform these analyses separately. Specifically, by exploring the effects of addition and removal of edges separately, it is possible to observe the behavior of our approach at higher noise thresholds. Thus, we define two procedures, one to look at the distribution of edge weights according to each metric, and one to compare the modules produced by the PPI network and noisy networks. Our approach to look at the distribution of edge weights is listed below:

1. Starting with the original PPI graph from each species, we produce families of 1,000 graphs each. The families are defined by different amounts of edges being either added or removed as follows: the first family has edges added equaling 10% of the number of edges in the original PPI graph. The second family has 20%, and so on, increasing in increments of 10% up to an increase of 80%. Similarly, edges are removed in increments of 10%, going from 10% to 80% of the original PPI graph. Each graph generated starts with the original PPI graph, to prevent bottlenecking of the resulting graphs.

2. Each graph is scored according to each edge metric.

3. For edges held in common between the original PPI graph and each graph with edges added or removed, a Pearson Correlation Coefficient and a Spearman Rank Correlation Coefficient are computed.
4. Pearson and Spearman coefficients are averaged across each family of graphs. The use of both Pearson and Spearman tests was important to determine the robustness of the metrics, but also how information encoded at this stage would affect the decomposition procedure used later in our approach.

The Spearman Rank Correlation between the original PPI graph and graphs with added noise provides an estimate of whether the relative density information encoded by each metric is changing with noise. This is critical in comparing different areas of the graph, as is performed later using hierarchical clustering. Under these circumstances, the Spearman is expected to provide a less stringent metric for noise tolerance, than the Pearson, of each metric, as the case of a high Spearman score and a low Pearson score is not expected and was not observed. All Spearman values were observed to be greater than 0.87, demonstrating the relative density measures are preserved by all four metrics. Thus, we assess the noise tolerance primarily on the Pearson scores.

The Pearson Correlation between the original PPI graph and graphs with added noise provides an estimate of how noisy each metric was relative to the changing background distribution of edge weights; the edge weight distribution is expected to change in graphs with added or removed edges as the degree distribution is changing.

The following figures demonstrate the robustness of each metric in the face of increasing noise in the underlying network.
Figure 3.25. Tolerance of Edge Metrics to Noise in the Worm (C. Elegans) PPI Network. Average Pearson Correlation Coefficient as a function of edges added or removed to construct noisy networks.

Figure 3.26. Tolerance of Edge Metrics to Noise in the Fly (D. Melanogaster) PPI Network. Average Pearson Correlation Coefficient as a function of edges added or removed to construct noisy networks.
Noise Tolerance of Edge Metrics in the Interactome of H. Sapiens

![Graph showing noise tolerance of edge metrics in the human network.](image)

Figure 3.27. Tolerance of Edge Metrics to Noise in the Human (H. Sapiens) PPI Network. Average Pearson Correlation Coefficient as a function of edges added or removed to construct noisy networks.

Noise Tolerance of Edge Metrics in the Interactome of S. Cerevisiae

![Graph showing noise tolerance of edge metrics in the yeast network.](image)

Figure 3.28. Tolerance of Edge Metrics to Noise in the Yeast (S. Cerevisiae) PPI Network. Average Pearson Correlation Coefficient as a function of edges added or removed to construct noisy networks.
The Geometric and Jacard Index perform uniformly poorly (relatively), in this test. Interestingly, the Meet/Min performs extremely well, roughly equaling or outperforming the Hypergeometric in all cases.

Since the eventual use of these metrics is the decomposition of PPI networks into modules, a more relevant and direct test of the performance of each metric, is their ability to produce the same modules with noise, as in the previous plots. Thus, we took the networks with added or removed edges, performed hierarchical decomposition on them, and determined what percentage of modules from the decomposition of the original network were present in the decomposition of the noisy networks. Specifically, for the comparison of a given noisy network to the original PPI network, all modules were conserved pairwise (for the decomposition due to each metric). The number of modules that were exactly preserved was measured, and averaged across the set of noisy networks. Use of exact matches is expected to be an extremely stringent test, especially in the case where edges are removed. This is depicted in the following plots.
Figure 3.29. Module overlap between PPI network and Noisy networks in Worm (C.Elegans). Modules were compared pairwise across networks for exact matches.

Figure 3.30. Module overlap between PPI network and Noisy networks in Fly (D.Melanogaster). Modules were compared pairwise across networks for exact matches.
Figure 3.31. Module overlap between PPI network and Noisy networks in Human (H.Sapiens). Modules were compared pairwise across networks for exact matches.

Figure 3.32. Module overlap between PPI network and Noisy networks in Yeast (S.Cerevisiae). Modules were compared pairwise across networks for exact matches.
Figure 3.31. Module overlap between PPI network and Noisy networks in Human (H.Sapiens). Modules were compared pairwise across networks for exact matches.

Figure 3.32. Module overlap between PPI network and Noisy networks in Yeast (S.Cerevisiae). Modules were compared pairwise across networks for exact matches.
Figure 3.33. Average modular overlap between decomposition of Worm (CE) PPI network and noisy networks, with identity threshold 0.9. Modules were compared pairwise.

Figure 3.34. Average modular overlap between decomposition of Fly (DM) PPI network and noisy networks, with identity threshold 0.9. Modules were compared pairwise.
Overlap of Modules between HS network and noisy networks with identity threshold 0.9

Figure 3.35. Average modular overlap between decomposition of Human (HS) PPI network and noisy networks, with identity threshold 0.9. Modules were compared pairwise.

Overlap of Modules between SC network and noisy networks with identity threshold 0.9

Figure 3.36. Average modular overlap between decomposition of Yeast (SC) PPI network and noisy networks, with identity threshold 0.9. Modules were compared pairwise.

With approximate matching, the behavior of the Meet/Min approaches that seen using a Pearson score on the edge weights. Though this clarifies the results to this point, the
behavior of the metrics must be determined at lower identity thresholds. Specifically, the relative behavior of the Meet/Min is dramatically different using approximate overlap; similarly, the relative behavior of the other metrics might change substantially with a small change in the identity threshold. To assess this, we look at the results of approximate modular overlap with identity thresholds of 0.8 and 0.7.

Figure 3.37. Average modular overlap between decomposition of Worm (CE) PPI network and noisy networks, with identity threshold 0.8. Modules were compared pairwise.
Overlap of Modules between DM network and noisy networks with identity threshold 0.8

Figure 3.38. Average modular overlap between decomposition of Fly (DM) PPI network and noisy networks, with identity threshold 0.8. Modules were compared pairwise.

Overlap of Modules between HS network and noisy networks with identity threshold 0.8

Figure 3.39. Average modular overlap between decomposition of Human (HS) PPI network and noisy networks, with identity threshold 0.8. Modules were compared pairwise.
Figure 3.40. Average modular overlap between decomposition of Yeast (SC) PPI network and noisy networks, with identity threshold 0.8. Modules were compared pairwise.
Figure 3.41. Average modular overlap between decomposition of Worm (CE) PPI network and noisy networks, with identity threshold 0.7. Modules were compared pairwise.

![Overlap of Modules between DM network and noisy networks with identity threshold 0.7](image)

Figure 3.42. Average modular overlap between decomposition of Fly (DM) PPI network and noisy networks, with identity threshold 0.7. Modules were compared pairwise.
Figure 3.43. Average modular overlap between decomposition of Human (HS) PPI network and noisy networks, with identity threshold 0.7. Modules were compared pairwise.

Figure 3.44. Average modular overlap between decomposition of Yeast (SC) PPI network and noisy networks, with identity threshold 0.7. Modules were compared pairwise.
As with approximate overlap at an identity threshold of 0.9, thresholds 0.8 and 0.7 show similar results: the Meet/Min performs extremely well relative to the other metrics.

By comparison, the Geometric, Jacard, and Hypergeometric are consistent across both tests: the Geometric performs relatively weakly when large numbers of edges are added, the Jacard performs poorly when edges are removed and intermediately when they are added, and the Hypergeometric performs well throughout. In one sense the performance of the Geometric is puzzling: according to the Pearson test, it performed relatively weakly when large numbers of edges were removed. By contrast, in the module overlap test it performed reasonably well when large numbers of edges were removed, but performed relatively less well when large numbers of edges were added.

As the Geometric and Hypergeometric were deemed optimal in comparison to detecting structure significant to randomized networks in the previous section, they are the final two choices to be decided amongst based on their performance in the presence of noise. Toward choosing an edge metric for further investigation, the Hypergeometric dramatically outperforms the Geometric especially when large numbers of edges are removed. For all subsequent investigations, we use the Hypergeometric as the edge metric for modular decompositions.

### 3.3 Biological Validation of Modules

Thus far, all validation has relied on statistical tests of each edge metric’s ability to capture significant topological information in each PPI graph. Since the intended endpoint is biological, the density and modularity must also be validated by biological criteria, since the presence of topologically-significant, but biologically-irrelevant
modules is not useful. Towards this end, we attempt to annotate the modules using Gene Ontology annotations, and determine whether modules are significantly enriched according to a wealth of gold-standard, hand-annotated biological information.

Additionally, we show an example from each PPI network highlighting the method's ability to correctly partition distinct functions.
Figure 3.45. Modules involved in the unfolded protein response in the worm PPI network: CE28071 (abu-11), CE06846 (lys-8), CE14080 (pqn-71; prion-like), CE07866 (abu-7), CE07866 (pqn-5; prion-like). Upon further decomposition, these proteins are correctly partitioned into the small complex involved in unfolded protein response.
Figure 3.46. Cell cycle modules in the fly PPI network. The large module includes a variety functions involved in the cell cycle: CG7405 (Cyclin H), CG10498 (cdc2), CG10308 (Cyclin J), CG11525 (Cyclin G), CG9183 (plutonium; DNA binding). These proteins are further isolated on further decomposition to yield the group of cell cycle regulatory proteins (Cyclins).
Figure 3.47. Signaling modules from the human PPI network. The large module contains many proteins involved in signaling pathways, including known oncogenes: HGNC:3430 (ErbB2/Her2-Neu), HGNC:3431 (ErbB3), HGNC:11283 (V-src Sarcoma), HGNC:6192 (Jak2), & HGNC:8804 (PDGFR). Subsequent decomposition results in the medium-sized
graph that involves signaling through the insulin-like receptor (ILR), and further decomposition yields the subset of ILR-signaling involved in phosphoinositol signaling: HGNC:8979 (PI3K-p85 subunit) and HGNC:9065 (Phosphoinositide phospholipase C gamma).
Figure 3.48. Modules from budding yeast (Saccharomyces Cerevisiae). The larger graph shows proteins involved in the cell division cycle. Note that the small set of nodes in the far right of the larger graph. Upon further decomposition this group becomes disjoint, leading the smaller graph pictured. The remaining part of the larger graph is the anaphase promoting complex, critical to mitosis. The smaller graph is the regulatory unit that controls the cell cycle, including the master cell cycle regulatory switch YBR160W (cdc28). Cdc28 is conserved across all known eukaryotes and is the central regulatory hug in cell division.

Currently, Gene Ontology provides the best way to assess large datasets with gold standard annotations\textsuperscript{66}. Though extremely useful for benchmarking subsets of data being analyzed, these ontologies are far from comprehensive.

Gene ontology annotations come in three flavors: 1) biological process (BP), 2) molecular function (MF), and 3) cellular compartment (CC). Although GO annotations represent a tremendous amount of information and an inordinate number of results of previous experiments, they primarily apply one annotation of each type to each gene. For our current purpose of validating our assignment of genes to modules, this is sufficient. However in the future, a dynamic picture of modularity would encompass genes switching functions as needed, and GO annotations might not prove useful for benchmarking modular assignment on a biological basis. For example, GO annotations provide a singular picture of each aspect of a gene: a single chemical activity, a single cellular localization, and involvement in a single process. As we begin to understand the many facets of each gene’s dynamic behavior, the utility of GO annotation in its current form will decrease.

To assess the biological validity of modular assignment of genes according to GO annotations, we check each module in the modular tree for enrichment of any GO category: BP, MF, or CC. Each comparison between a module M and an annotation A is
scored according to a cumulative hypergeometric test based on: the number of genes in yeast, the number of genes in M, the number of the genes with annotation A, and the size of the overlap between the genes in M and the genes in A. Due to the large number of modules and the large number of annotations, a substantial correction is required for multiple hypothesis testing; all P-values are Bonferroni corrected by the product of the total number of modules in the tree and the total number of annotations. Typically, each annotation group is separated, and P-values are adjusted by the product of the number of groups (e.g. modules) and the number of annotations of a given type (e.g. BP). However, we are using all annotations together and as such combine them into one large group, leading to a more stringent correction. For example, each gene has one possible BP, MF, & CC annotation. Thus the correction for multiple hypotheses should be the product of the number of all types of annotations (|BP| + |MF| +|CC|) and the number of genes. This normalization is applied to every P-value obtained by this process.

Almost half of the modules (49%; 562/1151) are enriched for at least one GO annotation (at least one GO annotation group [e.g. DNA damage/repair] from one of the GO categories: BP, MF, CC), defining enrichment as a P-value < 0.05 after Bonferroni correction. The following histogram shows the best GO annotation score attached to both genes and modules through this process. The P-values attached to genes are the smallest P-value attached any module that the gene is a part of.
Figure 3.49. Histogram of best GO annotation scores for Genes and Modules in Yeast (S. Cerevisiae).

The modules of primary interest are those with sizes between 10 and 100 genes, as these are large enough to explain large swaths of cellular function, while being small enough to provide real information at a scale smaller than the whole proteome. The following table illustrates several modules from yeast within the desired size range with significant enrichment in BP. Though they were checked for enrichment and included in the P-value correction, MF and CC annotations were largely ignored, since the purpose of using these modules later will focus on BP. This is due to the intended meaning of the annotations: BP is meant to capture cellular functions, whereas MF is meant to capture the chemical function of a single protein (e.g. enzymatic activity) and CC reflects where a
protein has been shown to localize in a single experiment. Though there are similar numbers of high scoring enrichments in each category, the BP annotations most closely resemble the information we are trying to capture through modular decomposition of the PPI networks.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Size</th>
<th>P-Value</th>
<th>Corrected P-Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoskeleton</td>
<td>94</td>
<td>$5.3 \times 10^{-20}$</td>
<td>$1.5 \times 10^{-13}$</td>
</tr>
<tr>
<td>RNA Polymerase II Activity</td>
<td>65</td>
<td>$1.5 \times 10^{-37}$</td>
<td>$5.6 \times 10^{-30}$</td>
</tr>
<tr>
<td>Ubiquitin-dependent Catabolism</td>
<td>39</td>
<td>$3.9 \times 10^{-55}$</td>
<td>$1.1 \times 10^{-48}$</td>
</tr>
<tr>
<td>Ribosome Biogenesis</td>
<td>88</td>
<td>$1.7 \times 10^{-77}$</td>
<td>$4.5 \times 10^{-71}$</td>
</tr>
<tr>
<td>DNA Repair</td>
<td>24</td>
<td>$1.2 \times 10^{-13}$</td>
<td>$3.2 \times 10^{-7}$</td>
</tr>
<tr>
<td>Splicing</td>
<td>56</td>
<td>$1.6 \times 10^{-50}$</td>
<td>$4.5 \times 10^{-44}$</td>
</tr>
<tr>
<td>v-SNARE activity</td>
<td>26</td>
<td>$1.8 \times 10^{-15}$</td>
<td>$4.8 \times 10^{-9}$</td>
</tr>
</tbody>
</table>

Figure 3.49: Gene Ontology (GO) Biological Process Enrichment among Seven Modules from S. Cerevisiae. *P-values adjusted for multiple hypothesis testing with a Bonferroni correction.

The results suggest an interesting bias in the modules found by this approach. Notably, the majority of highly enriched modules correspond to primordial functions shared by almost all eukaryotes, and virtually every other organism on the planet. This bias could reflect a number of possibilities: 1) the approach is biased toward finding primordial functions, 2) the dataset the approach is based on (the yeast PPI), is biased, 3) GO annotations contain more annotations of primordial functions than other functions, or 4) areas of the PPI network corresponding to primordial functions have higher densities.

Systematic error in the approach is difficult to assess without an additional systematic technique for comparison. Given that the decomposition relies purely of graph structure and topological information, and has no knowledge of primordial function, methodological bias seems an unlikely reason to observe a strong enrichment for primordial functions.
Systematic bias in the dataset is extremely likely for the following reasons: 1) primordial functions tend to be more essential for cellular function than other organism/genus specific functions, 2) due to their relative importance, primordial functions are vastly more studied, leading to greater elucidation of their underlying networks and a possible over representation in the PPI graph, and 3) furthermore, primordial functions are most known to be most enriched for disease genes in humans, suggesting even more bias in their study.

Similar to the systematic bias expected in the PPI network due to greater scientific interest in primordial function, GO annotations may very well have an identical bias. Thus, annotations in “younger” functions that could be species specific would be less likely, because relatively less of them exist in GO at this time. There is currently no way to measure the contribution of such a bias.

An additional source of systematic bias is the underlying true topology. If modules are in fact reused as cellular building blocks, we might expect that they have evolved to be denser the longer they have been in use. Considering the broad conservation of these groups, this is probably the true cause for the enrichment in primordial functions: modules that have existed longer have had more time to become much denser than newer modules. Unfortunately, substantially more high quality data is required to approach and conclusively answer this question.
3.4 Summary

In order to assess the validity of our approach to modular decomposition, we have tested the performance of four edge metrics: Geometric, Hypergeometric, Jaccard, and Meet/Min. The four metrics were compared using the following procedures:

1. Ability to detect modularity in PPI graphs relative to randomized networks that maintained degree distribution. In this test, the Geometric, Hypergeometric, and Meet/Min were all able to detect substantial differences between the PPI graph and the randomized networks. However, the Meet/Min performed best in areas of the network with low modularity, which are not of interest for our applications in later chapters. Thus, the Geometric and Hypergeometric were the strongest performing metrics in this test.

2. Robustness of edge scoring in the presence of noise, in the form of edges being randomly added or removed. Performance was assessed by comparing the distribution of edge weights, as scored by each metric, in the original PPI graph and the noisy graphs. Specifically, the Pearson Correlation Coefficient was calculated for edge weights on the intersection of edges between the PPI network and each noisy network. Amongst the four metrics, the Hypergeometric and Meet/Min were most able to produce similar edge weights in the presence of noise.

3. Robustness of the modular decomposition in the presence of noise. Since the determination of modules is our principle interest, we chose to look at how many of the modules from the decomposition of the original PPI graph are present in the decomposition of each noisy network, under each metric.
a. First, we attempted to check modular overlap using exact matches. For instance, for a module from the original network to be considered present in a noisy network, an identical module had to be present in the noisy network (the set of genes in the two modules had to be identical). Using exact matches, the Hypergeometric performed best. Most surprising was the poor performance of the Meet/Min. As this might have been caused by the stringency of exact matching, we performed a second test using approximate matches.

b. Second, we attempted to check modular overlap using approximate matches at a variety of thresholds (0.7, 0.8, & 0.9). Using approximate matches, the Hypergeometric and Meet/Min were again the strongest performers, as before in testing the distribution of edge weights.

3.5 Conclusions

These results demonstrate that substantial modularity exists in PPI graphs across species. This modularity is significant topologically (relative to background degree distribution and noisy input data) and biologically (functional enrichment). Thus, the module trees defined by this approach provide a scaffold on which to do further analyses of biological systems to investigate large-scale properties.
4. Modularity Applied to Analysis of High-Throughput Data

Introduction

With the decision to allocate billions of dollars to sequencing the human genome, came the expectation that this knowledge would benefit society, most likely through innovations in medicine. To date, very few changes in clinical care can be traced to elucidating the human genetic code. One of the roadblocks in translating genomic information has been how to leverage genetic data to both grasp the complexity of disease and build appropriately sophisticated therapeutics.

Ultimately, the application of genomic and other system-wide measurements to medicine will fall into two broad categories: 1) deciphering the etiology of complex disease and 2) determining the myriad responses of cells to small-molecules and tool compounds. In this section, we attack the latter problem using a large-scale view of the proteome to characterize the effects of a number of small-molecules. Many of the molecules under consideration are FDA-approved drugs, and provide an exciting platform on which to transform biological conclusions into medical understanding.

In the preceding section, we demonstrated that significant modularity exists in biological networks. Moreover we showed that this significance exists both at a topological level and a biological level, suggesting we can use the topology to further our understanding of biology. With this as our goal, we attempt to take the results of the previous section and apply them to contextualize and inform high-throughput screen data.

In 2003, Rosetta Inpharmatics, a biotechnology firm, performed a series of High-Throughput Screens (HTS) to better understand the mechanisms of action of 78
compounds. Although the ultimate interest of Rosetta is developing the next generation of therapeutics, they chose to carry out their experiments in the unicellular fungus Saccharomyces Cerevisiae (budding yeast). On the surface, the choice of budding yeast as a platform to investigate the mechanisms of therapeutics is unexplainable. However, many of the proteins involved in disease processes are within primordial modules that are conserved across eukaryotes. This makes yeast an attractive system to perform high-throughput analyses, while maintaining relevance.

Since completion of these experiments, technologies have become available to perform similar investigations in cultured human cells. Thus, methods developed in yeast, and corresponding analytic techniques, can soon be translated to humans.

One tremendous advantage of working with yeast is the ability to generate collections of yeast strains with targeted mutations. Such a collection is at the heart of the Lum et al. (Rosetta) experiment: 3,500 yeast strains, each missing one of two copies of a given gene. So-called heterozygously deleted strains provide an Achilles Heel in the form of the deleted gene, with virtually no change in the baseline function of the cell.

The primary intent of Lum et al. was to use the engineered weakness as a means to identify the protein target of each small-molecule. Specifically, if protein A is the target of compound X, decreasing the amount of protein A should increase the efficacy of X relative to both cells with no deletion, and cells with a deletion in another gene. It is worth reiterating that all cells under consideration by Lum, were diploid (two copies of each gene), and each strain was missing a single copy of one gene. Later we will consider experiments performed in haploid strains.
Underlying the intent of using heterozygously deleted strains for target identification, is the belief that a given compound has one and only one target, and interacts only with that target within the cell. If this model is in fact true, for each molecule there would be a single strain that is far more sensitive to the presence of the molecule, and all other strains would have no change in growth. Though this known not to be true in practice, the degree of the deviation is staggering: for many molecules there are tens of strains that are sensitized, for some there are hundreds. This suggests a picture of drug or chemical biology analogous to the genetics of complex traits. We will make a necessary detour to explore the link between small-molecule phenotypes and complex traits, as the underlying connection is tied to the modular structure of biological systems.

Modern genetics is hinged on the pioneering work of Gregor Mendel, who showed that many traits in agricultural systems can be separately inherited. Almost a century later, the molecular underpinnings of his work were shown to be DNA and more specifically genes. By this time, it had been shown that a number of rare diseases could be explained by the dysregulation of a single gene. Though a monumental success for genetics as a science, the scope of these conclusions is limited: the vast majority of diseases cannot be explained at the level of a single gene (monogenic, Mendelian disorders).

For many years, diseases that did not demonstrate Mendelian characteristics were assumed to have a non-genetic etiology. A prime example is heart disease: conventional wisdom first attributed heart disease to spiritual causes, eventually giving way to a causative basis in lifestyle. Studies in the last two decades have shown conclusively, that
genetic contributions play as much of a role as lifestyle in the morbidity and mortality of heart disease.

Initial conclusions that common diseases such as hypertension, diabetes, and heart disease had major genetic contributions were met with skepticism. A modular picture of a cell suggests that a given disease, if viewed as a failing in a larger cellular function, would have many genetic perturbations that could give rise to the same phenotype. Specifically, if a module with tens to hundreds of genes were responsible for a specific function, and failure of that function led to the disease in question, any set of lesions leading to the failure of the module should lead to the disease. Thus, we would expect that if we viewed the collection of perturbations leading to a given disease, from a functional modular point of view, a complex phenotype might appear more focused. One of the clearest and simplest examples comes from studies on the genetics of Non-Small Cell Lung Cancer (NSCLC).

Over the past two decades, with the advent of molecular biology, accumulated data has shown that mutation in three genes is the driving force in this type of malignancy: Epidermal Growth Factor Receptor (EGFR), K-Ras, and b-Raf. Meanwhile, extensive characterization has shown that these proteins are consecutive steps in a signaling cascade involved in cellular growth. Moreover, mutations in this trio are found to be mutually exclusive in any given tumor, suggesting that a single perturbation to this signaling module is sufficient to make the necessary contribution to uncontrolled cell growth on the road leading to cancer.

In the preceding chapter, modular decomposition of the human PPI network showed that large signaling modules can be systematically separated, resulting in smaller
modules, whose constituency agrees with the current understanding of signaling in cancer. Specifically, each of the resulting submodules is important in a different type of malignancy. This very strongly suggests that viewing these cancers from a modular point of view is an accurate depiction of their underlying defects.

Extending the conclusions of growth signaling in cancer, we can postulate the following: cells are organized into modules centered around reusable functions, and any substantial disruption of the constituents produces the same functional consequences (the failure of the function accomplished by the entire module). Thus, just as genetic perturbations targeting genes in the same module should have similar functional consequences, compounds targeting genes in the same module should similarly produce the same effect. This hinges on the idea that by abrogating the function of a protein, a small-molecule can mimic the effects of a genetic perturbation that similarly disrupts protein function. Thus, the response of cells to small-molecules can be viewed as analogous to genetic changes, and might benefit from a modular depiction.

As genomic technologies continue to advance, so do the techniques for discovering causative or correlative genetic lesions in complex diseases. Not surprisingly, most of these techniques are extensions of techniques that were used to find the cause of monogenic disorders. Much like sequencing a genome, these studies provide a “parts list” for a given disease: they enumerate a set of genes that, when perturbed, lead to the gene in question. Though this provides the starting point for understanding, and diagnosing, these diseases, it is far from a mechanistic picture.

As a measure of how complex the phenotypic pattern of a given drug across the set of mutants was, Lum partitioned the compounds into three groups: Group I had
virtually no response across all mutants, Group II had a strong response in a very small number of strains (usually less than ten), and Group III had a strong response in a larger number of strains, but still less than the majority. These groups can be likened to Mendelian and Complex traits: Group II is the set of approximately “Mendelian Compounds” and Group III is the set of “Complex Compounds.” Given their goal of target identification, an inherently Mendelian concept, it is not surprising that most compounds in Group III, and the complex end of Group II were difficult to pin down.

### 4.1 Methods

If modularity provides a means to understanding genomic information and complex traits, we would expect recasting Lum’s results in a modular framework to shed light on the activity and mechanisms of many of the more complex compound phenotypes. To accomplish this, we perform the following experiment:

1) We extend the score from Lum (which assigns a score to each strain for each compound; this score can be thought of as a score for each gene-compound pair) to a modular score for each module-compound pair.

2) Score the set of modules produced in the preceding chapter, based on the yeast protein-protein interaction (PPI) network.

3) Construct randomized versions of Lum’s dataset, and score them using the modular score.

4) Compare the modular score across the original dataset to randomized scores to find modules significantly enriched for hits by a given compound.
Assessing these predictions is very difficult, but in some cases, where the mechanism of action of a compound is known, we can qualitatively benchmark our predictions.

The first step in exploring the Lum dataset from a modular point of view, is translating each gene-compound score into a module-compound score. As the Lum scores are reported as P-values, this can be done by simply summing over the log(P) scores of genes in a given module, to give an aggregate P-value for the entire module.

Since Rosetta uses custom microarrays for their experiments, it was not possible to obtain and analyze the raw data for other interesting effects in such pooling experiments. Specifically, a pooled experiment may show that some strains have tremendous growth advantages when subjected to certain compounds (engineered resistance), which can be viewed as synthetic suppression.

The second step is producing a modular score for each compound. To calculate this score, we use the set of modules obtained from the modular tree of the yeast PPI, giving each module a single score per compound, as in the module-compound score above. One caveat of using the entire modular tree involves the number of partitions of the genes inherent in the tree. For example, each level of the tree corresponds to a partition of the genes still remaining during the decomposition. Though this seems to carry substantial useless information, it actually adapts nicely to the task at hand. This is due to the unknown degree of complexity of the underlying mechanism. Consider the two following toy examples. First, suppose that a given compound targets a module consisting of 100 genes, the optimal partition would be the level of the tree that produced the most modules of size 100. Second, suppose the compound had targets in 10 modules, each of size 10; the correct partition would be dramatically different. Thus, with no a
priori knowledge of the complexity of the mechanism, it is best to test across a wide range of modular sizes. Thus, we score all modules in the modular tree of the yeast PPI.

Third, we construct randomized data to determine the significance of the modular scores. For each compound, the mapping from gene to score is altered, which maintains the overall complexity of the compound’s genome-wide profile. Specifically, by permuting the names of genes relative to their phenotypic score (for a given compound condition), the background score distribution is maintained, while disrupting potential modular information. This provides a stringent control for assessing the significance of modular enrichment in a given condition.

Finally, for each module-compound score we assign a Z-score for the compound relative to the distribution of module-compound scores from the random dataset. Due to the constraints on the random data set, these scores correct for both the size of the modules and the background distribution of scores for each compound. The Z-scores are derived from P-values (corrected for multiple hypothesis testing) through the inverse CDF. The P-values are uniformly distributed from 0 to 1, and the Z-scores are distributed according to standard normal.

Results

Since our goal is to provide information that clarifies large datasets (reduces the number of genes involved in a phenotype of interest) while explaining complex traits (more than a single gene), we will focus only on modules of size 10-1000, with special emphasis on modules of size 10-100.
For each compound-module pair, a Z-score is assigned. However, several modules are assigned high Z-scores for the majority of compounds. These non-specific modules may be due to Multi-Drug Resistance (MDR) effects, or haploinsufficiency, wherein the cell is slightly sick due to being reduced to one copy of the gene at the locus of interest.

4.2.1 Non-Specific Compound-Module Interactions

An additional possibility that could explain several modules having high Z-scores across many conditions is a bias in the compound collection. For instance, if the compound collection is focused around a common biological theme, one might expect the modules connected with that theme to be high scoring in these analyses. On the surface, the set of compounds chosen by Rosetta is incredibly diverse at a number of levels: type of chemistry (including small, simple “drug-like” molecules as well as incredibly complex natural products), therapeutic use (antineoplastic, antifungal, antihypertensive, etc.), etc. If such a bias exists, we expect it to be detectable by at the level of pairwise similarity between compound profiles.

To assess pairwise similarity, we used three metrics: 1) a normalized dot product (the dot product of two profile vectors, normalized to the number of common values), 2) Pearson Correlation Coefficient, and 3) Euclidean Distance:
Figure 4.1. Normalized histograms assessing compound similarity. All pairs of compounds were scored according to the three metrics.

These pairwise scores indicate that the compounds are not sufficiently similar to assume the non-specific effects are caused by a bias in the compound collection. This provides an interesting question to determine if this is an artifact or a cellular function that when perturbed leads to broad vulnerability to small-molecule or other treatments. Further investigations are required to resolve this question.

Including modules which don’t appear to be compound specific, all 78 compounds have at least one module with $Z > 2$. 

118
Figure 4.2. Recurrence of Modules within Target Size Range. Many modules had significant scores in the explanation for many compounds. Modules were partitioned into compound-specific (Recurrence < 10) and compound-non-specific (Recurrence >= 10).

According to the histogram, we separate modules into two groups based on their recurrence, R: compound-specific modules (R < 10), and compound-non-specific modules (R >= 10).

**Compound Specific Modular Explanations**

Restricting to compound-specific modules, we find that 27 compounds have at least compound-specific module with significant scores. If the recurrence cutoff is relaxed to R < 30, all 78 compounds have at least one modular explanation. Although
high recurrence can be explained by lack of specificity, the collection of compounds used by Lum is enriched for three types of compounds (antifungals, antibacterials, and antineoplastics) increasing the likelihood of high recurrence due to mechanistic similarity in spite of high molecular specificity of the compounds. These two situations cannot be distinguished with this approach.

One caveat, which is critical to the exploration of complex traits, is the expectation of multiple modular explanations. In such a case, limited coverage of the phenotypic space and the modular tree will limit predictions, possibly misrepresenting a compound with a complex phenotype as Mendelian. Given the current coverage of the phenotypic data (62% of yeast proteins) and the modular tree (53% of yeast proteins), we can predict a set of modules that contribute to a given phenotype without certainty of additional contributions.

To explore the ability of modular structure to explain complex traits, we focus on three of the four Group III compounds: Actinomycin D, Ara-CMP, and, Carbendazim. The fourth Group III compound, Pentamidine, did not yield any compound-specific modular predictions.

Actinomycin D is a common laboratory tool compound, with a long history as a chemotherapeutic. In a survey of more than 250,000 compounds by the National Cancer Institute (NCI), Actinomycin D was observed to have the broadest and most potent ability to kill tumor cells in culture. Currently, Actinomycin D is used in the treatment of rhabdomyosarcoma, Wilms’ tumor, and gestational trophoblastic neoplasia. Because of its substantial toxicity and narrow therapeutic window, as evidenced by the NCI study, Actinomycin D has limited clinical application.
Recurrence vs. Module Size for Modular
Explanations of Actinomycin D
(Size 10 < s < 100 & Score Z > 2)

Figure 4.3. Modular explanations of Actinomycin D phenotype in Yeast (S.Cerevisiae). The low-recurrence modules in the bottom half of the graph are of primary interest.

Actinomycin D had two primary modular explanations with attached significant GO terms. The first module contains 11 proteins and is enriched in biological process annotation for RNA elongation from the RNA polymerase II promoter (module-compound score: Z = 4.7, R = 5; module-GO annotation enrichment: P = 4.8 x 10^{-15}), and enriched in molecular function DNA-dependent ATPase activity (module-compound score: Z = 4.7, R = 5; module-GO annotation enrichment: P = 8.9 x 10^{-19}).

The second module, also of size 11, showed enrichment for a biological process annotation in 35S primary transcript processing (module-compound score: Z = 2.9, R = 5; module-GO annotation enrichment: P = 7.1 x 10^{-23}), and a molecular function annotation
3'-5' exoribonuclease activity (module-compound score: $Z = 2.9$, $R = 5$; module-GO annotation enrichment: $P = 1.2 \times 10^{-23}$).

The first module is in fact the known target of Actinomycin D, which is known to inhibit transcription by binding DNA and interfering with RNA elongation due to RNA polymerase. The second modular prediction does not correspond to a known mechanism, but is a somewhat plausible as a second effect for a molecule with known nucleotide binding properties.

Figure 4.4. Modular explanations of Ara-CMP phenotype in Yeast (S.Cerevisiae). The low-recurrence modules in the bottom half of the graph are of primary interest.
Ara-CMP is a tool compound, which was briefly used as a chemotherapeutic. Interestingly, none of its modular explanations indicated a function related to genomic stability, which is expected given its putative mechanism as a Nucleotide/Nucleoside Mimetic (NNM). However, its most significant modular prediction came from a 17-protein module enriched for chromosome segregation (module-compound score: $Z = 4.0$, $R = 8$; module-GO annotation enrichment: $P = 4.6 \times 10^{-25}$). The finding of chromosome segregation, although not surprising for a chemotherapeutic in general, seems moderately unlikely in the setting of an NNM. This discrepancy could be explained in terms of a link between genomic instability and the concomitant inability of correct mitotic processes, though substantial evidence remains is still lacking.

![Recurrence vs. Module Size for Modular Explanations of Carbendazim](image)
The low-recurrence modules in the bottom half of the graph are of primary interest. Carbendazim, a broad-spectrum antifungal, had two unrelated modular explanations. The first module contained 14 proteins and was enriched for biological process annotation in peroxisomal organization and biogenesis (module-compound score: \( Z = 5.8, R = 9 \); module-GO annotation enrichment: \( P = 1.3 \times 10^{-16} \)), and was enriched in molecule function for protein binding (module-compound score: \( Z = 5.8, R = 9 \); module-GO annotation enrichment: \( P = 7.9 \times 10^{-13} \)).

The second module contained 13 proteins and was enriched in biological process annotations for regulation of cell growth (module-compound score: \( Z = 2.1, R = 8 \); module-GO annotation enrichment: \( P = 1.5 \times 10^{-11} \)). The mechanism of Carbendazim is not well understood, nor is a target known in yeast. Thus, further experimental studies are required to validate these predictions.

**Conclusions**

The ability to decipher the multi-faceted mechanisms of action of lead compounds and developmental therapeutics remains a primary challenge in translating genomic discoveries into clinical breakthroughs. We have attempted to recast this problem in terms of cellular modules, hoping to further disseminate mode-of-action information from the data of Lum et al.

Focusing on the most complex phenotypes, those of Group III compounds, we find that using a module tree as a hierarchial map of the yeast PPI networks substantially organizes and enhances phenotypic data, providing modular explanations for many of the
small-molecule phenotypes under investigation. In the case of Actinomycin D, this approach leads to prediction of the correct mechanism. Ultimately, to be useful, this approach requires datasets closer to complete coverage of the proteome of the organism of interest. More importantly, this type of experiment must be ported to cultured human cells, which is now becoming possible with the advent of arrayed and pooled RNAi approaches.

Given the ability of the modular decomposition to correctly separate signaling modules relevant to different tumor types (previous chapter), in the human PPI network, this approach could simplify and explain many complex phenotypes. Additionally, as the approach appears equally facile at simplifying small-molecule phenotypes, it may provide a way to bridge complex biological function with complex small-molecule responses. In combination with a systematically elucidated picture of the human proteome, these approaches could radically increase the pace of drug discovery.
5. Modular Alignment and Conservation

Every day, the vast amounts of genomic sequence data increase by leaps and bounds. These sequences allow new perspectives, questions and approaches to characterize and understand biological phenomena. Perhaps the most powerful approach in the field of genomics, is the comparison of genes across species. Determining how species have changed or not, provides a picture of how evolution has reshaped and optimized each organism. Additionally, comparative genomics provides a solid picture of function and evolutionary constraints.

Over the last two billion years, eukaryotes have branched off from primitive prokaryotes, and eukaryotes have split into fungi and animals. Since then, a number of interesting changes have taken place in how organisms evolve. Comparing budding yeast to any animal genome shows that the vast majority of yeast’s ~5,700 genes were present in some form in the last common eukaryotic ancestor\textsuperscript{38}. This suggests that the yeast genome has not changed dramatically in gene content. By comparison, animals have genomes two to three times larger than yeast (current estimates put humans at ~22,000 genes\textsuperscript{35}). Although this increase in genome size seems substantial, it appears strikingly low in light of the thousands of cells types must animals consist of.

Quite in contrast to the number of genes each organism possesses, the sizes of genomes are increasing. Yeast has a genome of approximately twelve million bases, more than seventy percent of which is composed of genes. On the other hand, humans have a genome of three billion bases, but less than one percent codes for proteins. Until sequencing on the human genome, the vast majority of intergenic, non-coding sequence
was assumed to be garbage strewn in an evolutionary wasteland. Comparative analyses have since demonstrated that many of these areas are strictly conserved across humans, and across mammals as well. Subsequent investigations have demonstrated that these areas are important for regulation of gene expression. This suggests that changes in the genome are taking place not at the level of growing or radically changing the parts, but in making minor changes in how genes are used.

These observations immediately raise two questions: is the rate of increase in number of genes decreasing, and if so, are organisms becoming more reliant on changing the wiring between genes and modules? The second question is difficult to answer without an extremely large number of genomes with precise identification of genes and gene number. Mounting evidence suggests that the answer to the second question is yes, leading to additional questions relating the organization of genes to modules. For instance, are intra-modular connections evolutionarily maintained more so than inter-modular connections? And are the mechanisms for tuning within modules different from those between modules? For example, if connections within modules are inflexible, are changes within modules made at the level of gene expression, whereas changes between modules are made at the level of protein-protein interactions?

Here, we make a first attempt to reformulate standard questions in comparative genomics in terms of modules rather than single genes. In so doing, we look for conserved modules that are represented across diverse sets of species to determine the importance of prevalence of each module.

The standard approach in comparative genomics, is to observe how a given gene is conserved, or not, across organisms. Here, we extend this approach to comparisons of
modules from four different species: yeast, fly, worm, and human. Just as comparative genomics gives insight into function at a single-gene level, we show that comparison of modules elucidates how modules can change. Ultimately, it would be possible to use results on modular evolution to determine how entire proteomes evolve, as there are several important issues currently beyond the reach of comparative genomics investigations at the gene-by-gene level.

5.1 Approach

Beginning with the module trees constructed in Chapter 3, we attempt pairwise modular alignments, followed by an extension to modules conserved across multiple species. Finally, we conclude with a preliminary attempt to align sections of the modular tree as an attempt to look at super-modular structure in the network. Our analysis will consist of three parts:

1. Alignment of pairs of modules from different species. These alignments will be performed for all modules from the PPI networks of yeast, fly, worm, and human.
2. Building off of pairwise alignments, construct alignments of modules across multiple species.
3. Use alignment of modules to construct a picture of similarities/differences at the level of the whole proteome, as covered by the PPI. This will be accomplished by looking for conservation of edges in the module tree.

5.1.1 Approach: Pairwise Alignment of Modules
All approaches to align networks thus far have focused on subgraph isomorphism, or some related problem. Given the noise levels in these data, this approach is likely to work on small modules, at the level of single protein complexes, but is unlikely to scale well to large modules spanning entire functions. Mapping between nodes in different graphs is then accomplished using sequence similarity between the proteins the nodes represent. To sidestep weaknesses in isomorphism caused by aberrant data, we separate the use of topology and sequence conservation. Specifically, if the partition that defines the modules is fixed before comparison across species, the topology within each module can be ignored for the pairwise alignment. Thus we construct a pairwise score for Module A from Species X to Module B in Species Y as follows:

1. Perform All vs. All BLAST between sequences in Module A and sequences in Module B.
2. Sum over log(E) values from BLAST alignments.
3. Normalize the score to the size of Module A and the size of Module B, as well as distribution of sequences in Species X and distribution of sequences in Species Y.

As in Chapter 4, the normalization of each score must compensate for both module size and some score applied to the module, which is the result of a background score distribution. Similarly here, we normalize by constructing a set of randomized modules for Module A and a set of randomized modules for module B. Each set of randomized modules comes from a randomization of all proteins that are contained within at least one module in the module tree that each module is derived from. Specifically, each randomized module is derived from a permutation on the names of proteins within the tree. Pairs of modules, one from each randomized set, are used to construct a distribution
of scores. Each randomized module is only scored once, against a single randomized module from the other species; this prevents biases in the randomized score distribution. Based on the randomized score distribution, each pairwise score is normalized to a Z-score; this process correctly adjusts for module sizes and distribution of sequences in a given organism’s proteome. For further analyses, the Z-score is used as the pairwise module alignment score.

One drawback of using the approach from chapter 4, is the favoring of smaller modules over larger modules. This is an artifact of how the randomizations are performed. For example, suppose we are attempting to align the root node of a given tree. Assuming there are no other connected components (only a single tree for the species in question), all of the randomized set of proteins from the node label permutation (permuting the names of the proteins), will be identical to the original set. In the case where the root node is aligned to a small module in the other species, randomization of the other species will partially correct for this effect. However, when comparing two large modules, it is virtually impossible to get a high scoring alignment. The only circumstances in which two large modules will produce a significant alignment is when there are multiple trees for each species (corresponding to multiple connected components in the starting PPI). Correcting this bias, while maintaining favorable properties of this scoring system is an area we are actively investigating.

5.1.2 Approach: Alignment of Modules Across Multiple Species

A natural extension of pairwise module scores, is scores across modules from multiple species. Using the pairwise scores derived in the previous part, composite Z-
scores can be derived naturally in a number of ways given a set of modules, each from a distinct species. We will refer to aligned modules as “metamodules.” However, one important question emerges to anchor the validity of these constructions. Suppose we have three modules A, B, & C from distinct species, with a very strong pairwise alignment between A & B, but very weak alignments between A & C and B & C. Optimally, we want to choose either an alignment of A,B,&C or A,B, not both. Thus the question becomes, how do we compare sets of metamodules of size n to all sub-metamodules of size n-1? This suggests against summing Z-scores, or averaging Z-scores, as the collection of Z-scores makes these distributions not comparable. Thus, we normalize the sum of Z-scores according to $m^{1/2}$ (where m is the number of pairwise module Z-scores). Using this approach, modules of different sizes can be compared to determine the correct number and identity of modular constituents in each metamodule. With this extension it becomes straightforward to construct the correct metamodule to represent conservation of functional building blocks across species. We provide the following approach:

1. Starting with pairwise scores between all pairs of modules across n species, we construct all metamodules of size n (a module from each species in every metamodule). Since the number of metamodules of size n is expected to scale as the product of the number of modules in each species, all scores of $Z < 1$ are removed. For a given module to become part of a metamodule, it must have alignment edges with weight $Z \geq 1$, to all other modules within in the metamodule, leading to a complete graph within each metamodule.
2. In the previous step, a number of modules will not be included in metamodules of size \( n \). These modules are combined into metamodule of the largest possible size, such that they have significant alignments (\( Z \geq 1 \)) to all other modules within the metamodule.

3. For each metamodule of size \( k \), all \( k \) sub-metamodules of size \( k-1 \) are constructed and scored.

4. Normalized Z-scores of the original metamodule of size \( k \) are compared to \( k \) the sub-metamodules of size \( k-1 \). If the normalized Z-score of the original metamodule is greater than the Z-scores of all of the sub-metamodules, the metamodule is retained. Otherwise, it is decomposed into the \( k \) submetamodules.

5. This process is repeated until all metamodules have higher scores than their submetamodules, or are reduced to size 2 (for which there is no \( k-1 \) comparison).

As there is no guarantee that larger modules will be favored over smaller modules, there is the possibility that all metamodules will decompose into submetamodules of size 2. However, this is extremely unlikely given that the scoring metric, which is an extension of sequence similarity, has some transitive properties.

5.1.3 Approach: Super-Modular Structure from a Module Tree

After metamodules have been constructed, it is possible to ask whether metamodules can be combined into a metamodule tree. This would allow an approximate global alignment. This work is preliminary at this point, but opens several interesting questions, including: if modules are conserved across species, is their organization conserved as well, and are there a core set of building blocks which are reused across
networks? The simplest approach to constructing a metamodule tree is the following. For each pair of metamodules to be joined by an edge in the metamodule tree, we designate one the parent node $P$, and the other the child node $C$. Then, we construct an edge between them if and only if, each pair of modules from a given species $S$ (which we will call $s_p$ and $s_c$, to designate the modules from species $S$ in the parent node $P$, and the module from species $S$ in the child node, respectively) has an edge between them in the module tree $S$, and $s_p$ is the parent of $s_c$ in the module tree of $S$. There are inherent drawbacks to such a stringent approach. We are currently considering other approaches.

5.2 Results

5.2.1 Results: Pairwise Alignment of Modules

We constructed alignments between all pairs of modules across species from yeast, fly, worm, and human, with the number of alignments ($Z \geq 1$) listed below:

<table>
<thead>
<tr>
<th>Species</th>
<th>DM</th>
<th>HS</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>1688</td>
<td>4501</td>
<td>3777</td>
</tr>
<tr>
<td>DM</td>
<td></td>
<td>15321</td>
<td>13066</td>
</tr>
<tr>
<td>HS</td>
<td></td>
<td></td>
<td>36578</td>
</tr>
</tbody>
</table>

Figure 5.1. Number of pairs of modules across species, yielding $Z$-scores greater than 1. All scores are normalized to each pair of modules and each species' sequence constituency.

Immediately notable is the enormous disparity in the number of pairwise alignments between different pairs of species. This disparity is not surprising given that the number
of alignments should scale as the product of the number of modules in each network. The following histograms show the distribution of scores in each alignment:

![Histograms of Pairwise Module Alignment Z-scores](image)

Figure 5.2. Histograms of pairwise module comparisons between pairs of four species: CE (worm), DM (fly), HS (human), and SC (yeast).
Figure 5.3: Example of a metamodule. RNase MRP/RNase P from Saccharomyces Cerevisiae (budding yeast) and Homo Sapiens (human). These RNases are broadly conserved across eukaryotes. Interactions from the PPI network are shown in blue and red, respectively; BLAST alignment edges are shown in black.

The number of modules and parameters of each interactome are listed below:

<table>
<thead>
<tr>
<th>Species</th>
<th>Modules</th>
<th>Nodes (interactome)</th>
<th>Edges (interactome)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>94</td>
<td>2800</td>
<td>4453</td>
</tr>
<tr>
<td>DM</td>
<td>282</td>
<td>7000</td>
<td>22142</td>
</tr>
<tr>
<td>HS</td>
<td>1016</td>
<td>7508</td>
<td>20974</td>
</tr>
<tr>
<td>SC</td>
<td>1151</td>
<td>5118</td>
<td>31156</td>
</tr>
</tbody>
</table>

Figure 5.4. Number of modules in module-tree representation of each network. Also, number of nodes and edges in each network/interactome.

To validate these results relative to a gold standard, similar to the Biological Validation of modular decompositions earlier, we utilized Gene Ontology (GO) annotations. Since
our technique is reliant on sequence similarity across species, we begin by removing GO terms derived from sequence similarity (ISS), electronic annotation (IEA), or not disclosed (ND). Additionally, to increase the ability to compare across species, it is necessary to use annotations that are as species non-specific as possible. Thus, we use the GO-Slim terms (a flattened, reduced set of GO terms derived from the GO hierarchy). In order to score the similarity in annotation across species, we combine and extend two techniques (the technique we employed earlier and a technique used previously, known as a coherence score):

1. As with modular decomposition in Chapter 3, we score the significance of each GO category for each GO annotation type for each module separately. Thus, for each module, we obtain a list of GO annotations that a given module is enriched for. As before, the enrichment score is computed based off a hypergeometric test taking into account the size of the network, the size of the module, the size of the GO annotation category (how many genes in species S have annotation A), and the number of genes in the module that have annotation A. The score is normalized using a Bonferroni correction. Due to the use of GO-slim categories, this a much smaller normalization due to a greatly reduced set of tests. Enrichments were considered specific for $P < 0.01$ after Bonferroni correction.

2. In order to compare two modules, we construct a coherence score. The coherence score is computed by iterating through all of the annotations that two modules are both enriched for, and the fraction of genes in each module that fall into one of the shared annotation categories. For example, suppose we are looking at modules A from Species X and B from Species Y. After looking for individual
enrichment in Biological Process, we find that A is enriched for mRNA processing and splicing, and B is enriched for snoRNA synthesis and splicing. Since the only annotation shared by the two modules is splicing, the coherence for this pairs of modules is the ordered pair (a,b), where “a” is the fraction of proteins in A that have a GO annotation of splicing, and “b” is the fraction of proteins in B that have a GO annotation of splicing.

3. We perform this analysis for all three types of GO categories: Cellular Compartment (CC), Molecular Function (MF), and Biological Process (BP). Coherence scores were performed for all pairwise modular alignments between C.Elegans (worm), D.Melanogaster (fly), and S.Cerevisiae (yeast). For technical reasons, it is difficult to perform this analysis on the human network, due to different identifiers being used across datasets and databases. First, we show the number of pairwise alignments that showed coherence scores greater than zero in tabular form. This is followed by scatter plots that show the ordered pairs of each coherence score.

<table>
<thead>
<tr>
<th>Species A</th>
<th>Species B</th>
<th>BP</th>
<th>CC</th>
<th>MF</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>DM</td>
<td>646 (38%)</td>
<td>420 (25%)</td>
<td>812 (48%)</td>
<td>1688</td>
</tr>
<tr>
<td>CE</td>
<td>SC</td>
<td>1376 (36%)</td>
<td>770 (20%)</td>
<td>1684 (45%)</td>
<td>3777</td>
</tr>
<tr>
<td>DM</td>
<td>SC</td>
<td>4829 (37%)</td>
<td>4179 (32%)</td>
<td>1376 (11%)</td>
<td>13066</td>
</tr>
</tbody>
</table>

Figure 5.5. Number of aligned module pairs with positive coherence scores for alignments between modules from worm (CE), fly (DM), and yeast (SC). Number of pairs (and fraction of total number of alignments) that had positive coherence scores are listed for each type of GO annotation: Biological Process (BP), Cellular Compartment (CC), and Molecular Function (MF).
The coherence score, extended in this form is expected to be extremely stringent, which we observed immediately, with less than half of all pairwise alignments having a positive coherence. There are two important points to note: 1) the total number of alignments reported above is the number of alignments with $Z > 0.5$, thus it is a filtered set and expected to be enriched for better alignments, and 2) coherence scores are not expected to have small positive values. The second expectation is a fallout of how there are calculated. Specifically, for a given module to be enriched for a certain annotation requires a substantial fraction of its proteins to have that annotation. Otherwise, it would not be considered enriched even before comparison to other module across species.

Figure 5.6. Scatter plot of coherence scores for pairwise alignment of modules between worm (CE; x-axis) and fly (DM, y-axis) under each type of GO annotation.
Figure 5.7. Scatter plot of coherence scores for pairwise alignment of modules between worm (CE; x-axis) and yeast (SC, y-axis) under each type of GO annotation.
Figure 5.8. Scatter plot of coherence scores for pairwise alignment of modules between fly (DM; x-axis) and yeast (SC, y-axis) under each type of GO annotation.

We provide one additional measure of the behavior of the coherence score in the above plots. To assess the average behavior, we compute the average of the Euclidean distance of each point to the origin in each of the above plots.

<table>
<thead>
<tr>
<th>Species A</th>
<th>Species B</th>
<th>BP</th>
<th>CC</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>CE</td>
<td>DM</td>
<td>1.19</td>
<td>1.19</td>
<td>1.37</td>
</tr>
<tr>
<td>CE</td>
<td>SC</td>
<td>1.10</td>
<td>1.19</td>
<td>1.28</td>
</tr>
<tr>
<td>DM</td>
<td>SC</td>
<td>1.17</td>
<td>1.25</td>
<td>1.33</td>
</tr>
</tbody>
</table>
Given the floor on the coherence scores, the average distance measures above are meant not for assessment of their absolute values, so much as comparison to one another. For instance, the MF annotation group has the highest scores throughout, which is apparent on visual inspection of the scatter plots. Conversely, the BP annotation group has the lowest scores. This could be due to true relative strengths and weaknesses of the approach, but is more likely due to a bias in the annotations or how they were applied. Overall, we see that the alignment procedure yields a reasonable number of high coherence alignments, as assessed through GO annotations.

One additional attractive opportunity for assessing our results comes from gene expression experiments. Unlike the PPI networks that give rise to our modules, gene expression experiments are tied to the conditions under which they were generated. Thus, if one were to perform the same experiment on yeast cells and cultured human cells, capture the gene expression profiles, and compare them, it would be possible to find modules across species that are functionally related. The expected product of such an analysis is a connection between groups of genes across species; this would indicate functional similarity as viewed through the lens of a specific perturbation. Unfortunately, to date, there are not sufficient matched sets of experiments across species to support such a strategy. As the cost of these technologies decreases, it will become possible to construct this experiment to view modularity from a different point of view.

5.2.2 Results: Alignment of Modules Across Multiple Species

Given the number of modules derived from each species, there are approximately 30 billion possible metamodules of size four at the beginning of our iterative procedure.
Removing all alignment scores for which $Z < 1$, the starting set of metamodules numbers just under 3 million total:

<table>
<thead>
<tr>
<th>Size of Metamodule</th>
<th>Number of Metamodules of that size in starting set</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1,200</td>
</tr>
<tr>
<td>3</td>
<td>1,128</td>
</tr>
<tr>
<td>4</td>
<td>2,760,718</td>
</tr>
</tbody>
</table>

Figure 5.5. Number of Metamodules of sizes 2-4 after filtering out weak alignments ($Z < 1$), before the beginning of iterative procedure.

Starting with the above set of metamodules, the iterative procedure is applied, yielding the following new distribution of metamodules:

<table>
<thead>
<tr>
<th>Size of Metamodules</th>
<th>Number of Metamodules of that size in final set</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1,485,222</td>
</tr>
<tr>
<td>3</td>
<td>1,457,022</td>
</tr>
<tr>
<td>4</td>
<td>2,275,044</td>
</tr>
</tbody>
</table>

Figure 5.5. Number of Metamodules of sizes 2-4 after filtering out weak alignments ($Z < 1$), after iterative procedure.

Most striking is the finding that the number of metamodules of size 4 does not change dramatically. Although this could represent true conservation of the majority of modules across species, it could also result from the following: a bias in the interactome (a sampling bias due to areas of research interest), a bias in the construction of modules leading the same types of modules being chosen across species, or, most likely, a bias in the scoring scheme.
Specifically, normalizing the Z-score is meant to make metamodules of different size easily comparable. However, if this procedure does correctly normalize, certain sizes of metamodules might be favored over others. For example, the table above suggests that large metamodules are being favored over smaller metamodules. The most likely contributor to this occurrence is the bias in the scoring toward small modules. As mentioned earlier, the scoring introduces a slight bias causing small modules to have higher scores than large modules. Any bias will be magnified during the construction of metamodules since small modules are not only the highest scoring, but also the most prevalent, almost by definition. These effects are difficult to deconvolute, as certain sizes of modules might be more strongly conserved than others. As our scoring system improves, it might become possible to distinguish more general trends in conservation from artifact.

5.2.3 Results: Super-Modular Structure from a Module Tree

Preliminary investigations into conservation of module trees across these four species have yielded very small sets of edges in metatrees constructed out of metamodules of size 4. The degree distribution within the metatree is extremely inhomogeneous, dominated by a small number of large metamodules, connected to a large number of small metamodules. Although this is most likely due to problems with the scoring, that was observed through the large number of metamodules of size four, it may also be due to overly stringent constraints in how the tree is constructed.

We are currently investigating ways to search for the optimal set of metatrees, by iteratively moving metmodules between size k and k-1. This would provide a global
means to determine optimal local and global structure, instead of focusing entirely on the constituency of the metamodule by itself.
6. Discovering Modularly Targeted Small-Molecules

Much as modularity provides a different viewpoint from which to view a cell, discovering modularly targeted molecules requires a different approach. The predominant method used to discover new lead compounds relies on *in vitro* biochemical approaches used on a massive scale. Here, we propose an orthogonal approach that relies on genetic methods in place of conventional biochemical techniques.

Genetics was born out of the ultimate experiment in faith: manipulating a system of interest with virtually no knowledge of its components. Though this presents a daunting prospect, a century of substantial breakthroughs being made through genetic analysis highlights the virtues of this approach. Moreover, the unbiased, top-down viewpoint of genetics can often provide new and interesting avenues into a problem, far beyond that originally envisioned. We attempt to exploit these benefits toward the discovery of new small-molecules centered around a module of interest.

No tool in the geneticist’s arsenal is more powerful than the ability to generate mutations. Fundamental discoveries, such as the elucidation of cell division cell control, were accomplished starting with generation and characterization of mutants. As with most things, if one is good, two is better. In the case of genetics, mutations are often generated around a phenotype of interest. Suppose we are interested in the ability to grow in the presence of kryptonite. We might take cells that are able to grow in the presence of kryptonite, introduce mutations, and see which lose this ability. By discovering the location of the mutations, we could determine the critical part, whose loss of function led to the inability to grow in the presence of kryptonite. In theory, by doing this enough times we could saturate the entire set of cellular constituents required for
kryptonite tolerance. In a way analogous to genome sequencing, this would provide us with a “parts list,” but no knowledge of how these parts interact to contribute to kryptonite tolerance. One might naturally abandon genetics at this point, and try to characterize the biochemistry of these parts. Blind in our devotion to genetics, we will continue on using purely genetic techniques.

Starting with cells that have lost their kryptonite tolerance through mutation, we again subject them to mutations and look for those that revert and are again able to tolerate kryptonite. One could guess that the mutation that originally led to loss of kryptonite tolerance might be precisely reversed, but it turns out to be rarely the case. Across decades of genetics papers, reversion due to reversal of the original mutation is almost never observed. Thus, we now have a set of mutants, with two mutations apiece (we make the simplifying assumption that only one mutation is generated at each step, which can be controlled in practice to yield this result).

More importantly, we have a relationship: a mutation in A (the mutation that caused loss of kryptonite tolerance) can be suppressed by a mutation in B. Doing this exhaustively, we can develop large mechanistically distinct classes of mutants. We can also begin to organize our groups as follows: we can see which member of class B can suppress which members of A. This allows us to partition each group, providing a new layer of information with no knowledge of the underlying parts (we are ignoring the technical details of comparing mutations in different cells; suffice it to say this is well within the abilities of basic genetics). To further add to our knowledge of these relationships, we can look for mutations that exacerbate the initial mutation: we could find a set A’ of mutants that have reduced, but not completely abrogated, kryptonite
tolerance. We then look for a group of mutations in a class B’, that lead to complete loss of kryptonite tolerance. In this way we can elucidate many relationships related to our trait of interest.

Unfortunately, we cannot engineer mutations in patients to fix genetic predispositions to disease. Fortunately, we can find molecules that mimic the mutations, that might then be ported to clinical use. Ignoring the billions of dollars, and diminishing odds of clinical success in the latter step, we focus on how to use small-molecules to perform genetic experiments toward discovering new lead compounds.

First, we require a phenotype more frequently observed than loss of kryptonite tolerance. We choose gain of function through Phosphatidylinositol-3-Kinase (PI3K) signaling. PI3K is required for a crucial step in cell growth signaling that is observed to be dysregulated in cancer, diabetes, and schizophrenia. PI3K is responsible for the phosphorylation of PIP2, resulting in PIP3; the reverse reaction is catalyzed by PTEN phosphatase. Increase in PI3K function, or more commonly loss of PTEN function, is observed in a wide variety of cancers, as well as in type II diabetes. Interestingly, loss of flux through this signaling pathway is linked to schizophrenia. We now describe a chemical genetic approach to target this pathway.

Decades of searching for molecules targeted at PI3K have yielded a number of compounds that inhibit it in culture systems, but fail to make the transition into the clinic. Perhaps the most prevalent example is the natural product Wortmannin:
Since our goal is not specifically to find an inhibitor of PI3K, but to decrease effective flux through PI3K signaling, we first attempted to find molecules capable of suppressing Wortmannin. To accomplish this, we screened 2,500 known bioactive compounds to find any that might be capable of suppressing the growth inhibitory effects of Wortmannin.

This screen was performed in the budding yeast Saccharomyces Cerevisiae, with the ubiquitous laboratory strain BY4742. Cells were grown in Complete Synthetic Media + Wortmannin at a concentration that completely inhibits growth (293nM). Growth was monitored at 48 and 72 hours. Three molecules suppressed Wortmannin under these conditions, one of which retested successfully: U73122:
Figure 6.2: U73122, a bioactive compound capable of suppressing Wortmannin. U73122 is known to inhibit PLC-gamma.

As the mechanism of U73122 was known, it provided an immediate hypothesis to explain how it suppresses Wortmannin. U73122 is known to target PLC-gamma, which breaks PIP3 down to IP3 and diacyl-glycerol (DAG), one of the pathways downstream of PI3K function. This suggested that by blocking one pathway downstream of PI3K, flux through another pathway is increased, a decrease in which is the basis of Wortmannin’s effect.

Further confirmation of this hypothesis is still being investigated.

Additional experiments into U73122/Wortmannin phenotypes produced a surprising result: U73122 inhibited cell growth in isolation, but was suppressed by Wortmannin.

![Growth Profiles of Wortmannin and U73122](image)

**Growth Profiles of Wortmannin and U73122**

- Control
- Wortmannin
- U73122
- U73122 & Wortmannin

Figure 6.3. Growth profiles at 48 hours for BY4742 in the presence or absence U73122 and Wortmannin. Optical Density (600nm) is proportional to cell number in log phase growth.

This provided a very attractive way to find Wortmannin-like molecules: screen for suppressors of U73122. Not only did this provide an additional avenue from which to
investigate this signaling pathway, it provided a cleaner compound. Not only has Wortmannin demonstrated an inability to make a useable therapeutic, it is extremely difficult to use in laboratory investigations. Specifically, significant relative experimental variation prevents consistent screening conditions. Luckily, U73122 proved to be extremely reliable in producing a consistent phenotype at a fixed dose. Thus we elected to conduct a second screen looking for suppressors of U73122 and Wortmannin.

6.1 Screen of Commercial Compounds for U73122 & Wortmannin

Modifiers in Budding Yeast

To enlarge our search space, we screened ~75,000 commercial compounds from three libraries: Chembridge Microformat (~50,000 compounds), Enamine (~10,000 compounds), and Maybridge I/II/III (~15,000 compounds). Each library compound was screened under three conditions: in the presence of U73122, in the presence of Wortmannin, and a control condition without either compound. This allowed us to assess the bioactivity of each compound as well as its ability to modify the growth inhibitory phenotypes of U73122 and Wortmannin. All experiments were done in Complete Synthetic Media (CSM) at room temperature (25C) in 384 well microtiter plates, with a working volume of 45uL. On each 384 well plate, 64 wells were used as controls, not exposed to library compounds, and provide a plate control as well as contributing to the mock treatment distribution. These wells should not be confused with the control wells that were exposed to library compounds but not U73122 or Wortmannin. To clarify, we will refer to them as mock-controls and library-controls, respectively. Wortmannin was used at a dose of 293nM, and U73122 was used at a dose of 3.5uM; both compounds
were diluted from 5mg/ml stock solutions in DMSO, into CSM. Growth was assessed using a SpectraMax HT spectrophotometer to measure Optical Density at 600nm (OD600). OD600 is a standard way to measure yeast growth, and has been shown to be directly proportional to cell number during log phase growth. For our purposes, absolute growth numbers are not necessary; we are primarily interested in relative growth across conditions. The following histograms show the growth profiles at 48 and 72 hours:

![OD600 Histogram (48 hours)](image1)

![OD600 Histogram (72 hours)](image2)

Figure 6.4. Normalized histogram of (mock-)control wells at 48 & 72 hours. Note: these wells were not exposed to library compounds. They provided a mock distribution for comparison.

Immediately evident is the difference between the two compounds: U73122 continues to inhibit the controls at 72 hours, whereas Wortmannin fails to completely inhibit the
majority of wells at 48 hours. One way to improve Wortmannin’s ability to inhibit growth is simply to increase the concentration. Unfortunately, this is difficult in the context of a suppression screen, since raising the dose of the inhibitory stimulus may limit its ability to be suppressed. Plates were stratified according to the growth of control wells: plates exposed to U73122 or Wortmannin were only used for subsequent investigation if the mock-control wells showed no growth at 48 hours. This substantially limited the number of library compounds that were candidates for Wortmannin suppression. However they were useful in the context of finding Wortmannin enhancers (compounds that increase the ability of Wortmannin to inhibit cell growth; bioactive compounds were not considered as Wortmannin enhancers).

Figure 6.5. Normalized histogram of wells exposed to library compounds, at 48 & 72 hours.
The histograms for compound wells show a few notable features relative to the control wells: 1) there are numerous bioactive compounds (as denoted by the larger left-sided tail of the histograms in the upper two plots, indicating these wells [library-controls] did not grow at either time point), 2) there is a very slight increase in the number of wells that grew in the presence of U73122, suggesting that some library compounds were able to suppress U73122 at both timepoints, and 3) despite Wortmannin’s ability to completely inhibit growth in mock-control wells, there is a marked increase in growth at both timepoints suggesting that suppression is still observable (and can be differentiated, albeit at lower confidence). As we were limited in the number of compounds we could request for reorder (750), we focused on plates where the mock-controls did not grow at 48 hours. We had three primary groups to look for:

1. U73122 Suppressors. These are compounds that rescued growth in the presence of U73122. Mock-controls had OD600 values of less than 0.04 at 72 hours. The choice of using the mock-control growth at 72 hours, rather than 48 hours, provides a more stringent assessment of plate dosage and hence suppression capability of the library compound. Due to the restriction on retest number, this was chosen because of the large number of plates that passed filtering (based on mock-control growth) and the large number of library compound wells that showed suppression at 48 hours. Amongst plates that passed filtering, wells were ranked according the OD600 at 48 hours in the presence of U73122, and the top 500 wells were chosen for retesting.
2. Wortmannin Suppressors. These are compounds that rescued growth in the presence of Wortmannin. Mock-controls had OD600 values of less than 0.04 at 48 hours. Compared to the U73122 condition, a much smaller number of plates passed the filtering step for the Wortmannin condition. Thus, the stringency of the mock-control growth value was reduced to no growth at 48 hours (rather than 72 hours). Library compounds from plates that passed filtering were ranked according to OD600 in the presence of Wortmannin at 48 hours. The top 200 were chosen for retesting.

3. Wortmannin Enhancers. These are compounds that inhibited cell growth in the presence of Wortmannin, but did not inhibit cell growth in the absence of Wortmannin. Plates that failed filtering for Wortmannin suppression were considered for Wortmannin enhancement. Wells that showed no growth at 72 hours were considered, and were ranked by the growth of mock-controls (since well without signs of growth cannot be ranked, they were ranked by the growth inhibition of the mock-controls which is a proxy for Wortmannin dose on a given plate. The lower the Wortmannin dose, the more suggestive of Wortmannin enhancement). The top 50 were chosen for retesting.

The decision to search for Wortmannin enhancers was largely practical: due to the difficulty of achieving correct dosing in a suppression experiment in general, and when using Wortmannin in particular, it was not possible to get consistent inhibition at a level that might allow for suppression. Thus, the plates that would otherwise be ignored, due to growth of mock-control wells, were utilized to study Wortmannin enhancement. An analogous decision prevented the study of U73122 enhancement: very few mock-control
wells grew across the entire experiment, completely nullifying the possibility of measuring U73122. The decision to pick substantially more Wortmannin suppressors at a less stringent threshold, instead of more Wortmannin enhancers at a more stringent threshold, was made despite a large number of plates not passing filtering for Wortmannin suppression (as evidenced by the large number of mock-controls that grew at 48 hours in the presence of Wortmannin). The reasons were two-fold: 1) Wortmannin suppression was deemed more interesting than Wortmannin enhancement due to its ability to shed light on unexplored aspects of PI3K biology, and 2) synthetic suppressors are substantially more likely to retest than synthetic enhancers.

The latter claim can be understood through the simplistic logic that it is easy to inhibit growth or outright kill cells through cumulative damage. Thus, difficulty in distinguishing true synthetic lethality (specific) from cumulative damage (non-specific) in a primary screen, can lead to extremely low retest rates. Since the purpose of synthetic genetic, and synthetic-chemical genetic, experiments is to expose connections between coupled cellular activities, suppression provides a much more attractive prospect for the following reason: the ability to rescue a cell from growth inhibition is very specific to the source of inhibition. In numerous studies we have observed extremely high retest rates in synthetic suppression screens (>90%), and extremely low retest rates (<20%) in synthetic lethal screens.

All combined, 750 molecules were chosen for further characterization. It is worth noting that many more molecules could be followed up. Retesting additional molecules would also allow a more meaningful assessment of structure-activity relationship (SAR), as the commercial libraries are enriched for certain types of chemistries. With a broad
retest, it may be possible to identify specific families of chemistries that are linked to each of the phenotypes being investigated.

6.2 Retesting Hits from Commercial Compound Primary Screen

To further evaluate and confirm the activity of the 750 compounds chosen in the primary screen, a retest was initiated. Since the primary screen was only able to partially resolve the potency of each compound for a given phenotype, the stringency of the retest was substantially higher than the primary screen. Specifically, concentrations of both U73122 and Wortmannin were increased by 50%. Retests were conducted as follows:

1. 1uL of each library compound was robotically pipetted into the well of a 384 well plate (compound stock plate).

2. 49 uL of CSM was added to each well. Each solution was mixed by hand using a 16-channel Finn Pipette.

3. 3 x 15uL aliquots were removed from each well and separated into three separate wells (one for each condition/assay plate: library-control, U73122, and Wortmannin).

4. 30uL of CSM was added to the 15uL from the previous step, bringing the working volume to 45uL. In each 384-well assay plate, half of the wells were used as mock-controls to provide a more substantial control distribution and a better indication of fluctuations in condition-dose/plate-to-plate variability.

5. To the remaining 5uL in the compound stock plate, 45uL of CSM was added.

6. The 50uL in the compound stock plate was separated into two 25uL aliquots (one for each condition: U73122, and Wortmannin).
7. 20uL of CSM was added to each well from the previous step, bringing the
working volume to 45uL. As in step 4), half of the wells of each 384-well assay
plate were used as mock-controls.

8. OD600 values were measured at 48, 60, 72, & 84 hours.

Steps 1-4 were meant to exactly reproduce the conditions of the original experiment with
more mock-controls, and increased dosage in the U73122 and Wortmannin conditions.
Steps 5-7 were meant to simulate a low dose of library compound that should have been
approximately 10-fold lower than the original experiment. The following two scatter
plots show data from the retest experiment focused around the phenotype of primary
interest (U73122 Suppressors, Wortmannin Enhancers; this is “Wortamannin-like”
phenotype):

![Scatter plot](image-url)
Figure 6.6. High Dose. Growth in the presence of U73122 at 48 hours (Y-axis) vs. Growth in the presence of Wortmannin at 84 hours (X-axis). Each point represents a specific library compound. “Wortmannin-like” molecules are expected in the upper left. Timepoints were chosen for maximum stringency.

Figure 6.7. Low dose. Growth in the presence of U73122 at 48 hours (Y-axis) vs. Growth in the presence of Wortmannin at 84 hours (X-axis). Each point represents a specific library compound. “Wortmannin-like” molecules are expected in the upper left. Timepoints were chosen for maximum stringency. Note: single molecule in upper left showing dramatic “Wortmannin-like” phenotype at low dose. This constellation of conditions is the most stringent test for this phenotype.

Despite the increase in stringency, 542/750 (73%) compounds retested under the original experimental conditions and 52/750 (7%) compounds retested under the low dose conditions. Though the low-dose retests are encouraging, it should be noted that these wells might simply have retested under low dose because the compound was not
completely mixed and did not distribute evenly. We are currently in the process of characterizing these molecules further.
7. Systematic Interpretation of Genetic Effects Using a Scaffold of Molecular Interactions

Genetic interaction, in which two mutations have a combined effect not exhibited by either mutation alone, is a powerful and widespread tool for establishing functional relationships between genes. Numerous screens have generated a set of over 1,400 genetic interactions in yeast. We demonstrate that by integrating these data against a large network of known protein-protein, protein-DNA, and biochemical interactions, it is possible to reveal physical pathway mechanisms responsible for the observed genetic relationships.

7.1 Introduction

A major, ongoing challenge in the field of genetics is to interpret observed genetic interactions in a mechanistic, physical cellular context \(^{91-93}\). Genetic interactions consist of two major varieties: synthetic lethal interactions, in which mutations in two nonessential genes are lethal when combined; and suppressor interactions, in which one mutation is lethal but combination with a second restores cell viability. Screens for genetic interactions have been used extensively to shed light on pathway organization in model organisms \(^{91-94}\), while in humans, genetic interactions are critical in linkage analysis of complex diseases \(^{95}\) and in high-throughput drug screening \(^{96}\). For species such as yeast, recent experiments have defined large genetic networks cataloguing thousands of such interactions \(^{94,97,98}\).

Explanations for genetic interactions typically involve some combination of physical interactions or events, including protein-protein binding, transcriptional cascades, and/or biochemical reactions. Auspiciously, large databases of physical interactions have also
recently become available. Direct physical interactions among proteins (protein–protein interactions) are revealed by systematic two-hybrid\textsuperscript{99-102} or immunoprecipitation (IP) studies\textsuperscript{93,94}, while physical interactions between protein transcription factors and their promoter binding sites (protein→DNA interactions) are rapidly determined using chromatin IP in conjunction with DNA microarrays\textsuperscript{95}. Metabolic pathway databases store another type of physical interaction, consisting of biochemical reactions linked together by the metabolic substrates they consume or produce.

Previously, we and others found significant correlated structure between the genetic and physical networks for yeast—many physical and genetic interactions are co-incidental\textsuperscript{98}, and proteins with many interactions in the physical network are likely to also have many interactions in the genetic network\textsuperscript{96}. Such correlation suggests that it may be possible to model observed synthetic lethal and suppressor relationships using networks of physical interactions—and in so doing, to systematically elucidate the ‘wiring diagram’ of signaling, regulatory, and metabolic circuitry underlying cell function. Here, we propose and evaluate one such approach based on a probabilistic model of the physical network. Previously, probabilistic models have proven powerful in biological sequence analysis and other bioinformatic domains; a similar physical interaction model has been recently applied to explain gene expression changes in response to network perturbation\textsuperscript{97}. Here, we construct a model of physical interactions to optimally explain the observed synthetic lethal and suppressor effects. In addition, we experimentally confirm several new genetic interactions predicted by the model relating to cell cycle.

7.2 Methods

7.2.1 Overview: Modeling Physical Pathway Explanations
Classically, synthetic lethal and suppressor interactions between pairs of genes have been explained either via a single physical pathway directed from one gene to the other (Fig. 7.1a), or by a pair of redundant pathways directed from each gene to a common downstream recipient\(^9\) (Fig. 7.1b).

**Figure 7.1: Physical pathway explanations for genetic interactions.** Within a single physical pathway, the additive effect of mutating two genes in combination can lead to a synthetic lethal interaction (between genes \(g_1\) and \(g_2\); panel A), or else a lethal mutation in one gene can be suppressed by a mutation in a second gene downstream (B). For parallel physical pathways, non-lethal mutations in each pathway can combine to produce synthetic lethality (C), or else a mutation in one pathway can bypass a lethal mutation in the other to restore viability (D). Effects of mutations are represented by dark gray nodes and dotted arrows. Depending on the context, mutations may denote single point mutations, larger disruptions, or complete gene deletions.

Formalizing these rules in the context of the physical interaction network (a directed graph), we define an explanatory path (e-path) to be any path of physical interactions for which there exists a non-empty set \(\Omega\) of downstream nodes reachable from both path endpoints. For example, the paths \(A \rightarrow B \leftarrow C\) and \(A \rightarrow B \rightarrow C\) are e-paths, but \(A \leftarrow B \rightarrow C\) is not. Since this definition includes the degenerate cases in which either endpoint is in \(\Omega\),
e-paths are a high-level representation of both the parallel and single pathway explanations shown in Fig. 1. We say that a node pair \((a,b)\) spans the e-path if some node in \(\Omega\) is reachable from both \(a\) and \(b\) (single directed path) or else is reachable from \(a\) and \(b\) along separate paths that are non-overlapping (parallel paths). We assume that genetic interactions between spanning nodes occur with high probability \(\alpha\) and otherwise occur with background probability \(\beta\), where \(\alpha \gg \beta\). Thus, each e-path partitions the set of all node pairs into two sets—those that are predicted to genetically interact and those that are not.

### 7.2.2 Formal Model and Scoring

We formulate a genetic interaction model that associates a probability of genetic interaction \(p_i\) with each node pair \(i\), given a set of e-paths and the parameters \(\alpha\) and \(\beta\):

\[
p_i = 1 - (1 - \alpha)^k (1 - \beta)^{N-k}
\]  

(1)

where \(N\) is the total number of e-paths and \(k_i\) is the number of e-paths for which pair \(i\) is predicted to genetically interact. Thus, \(p_i\) is the probability that at least one e-path causes a genetic interaction at pair \(i\).

We wish to find the set of e-paths for the model that maximizes the probability of the observed data set of positive and negative genetic interactions. A positive observation indicates that the node pair has been experimentally tested and determined to be a synthetic lethal or suppressor relationship, while a negative observation has been tested and determined to have no such genetic relationship. Using Bayes’ Rule:

\[
\Pr(M \mid D) = \frac{\Pr(D \mid M) \cdot \Pr(M)}{\Pr(D)} \approx \Pr(D \mid M) \cdot \Pr(M)
\]

(2)
where $M$ and $D$ denote the model and data, respectively. $Pr(D)$ is constant since all models are conditioned on the same set of observed genetic interaction data. $Pr(M)$ represents the probability that the e-paths in the model exist, that is, that the physical interactions $i$ in the union of all e-paths in $M$ are real and that the remaining physical interactions $j$ are not real:

$$Pr(M) = \prod_{i \in M} e_i \prod_{j \in M} (1 - e_j)$$  \hspace{1cm} (3)$$

Probabilities of interaction $e_i$ and $e_j$ are taken from relevant prior literature and depend on whether the physical interaction is of type protein-protein, protein-DNA, protein-reaction, or reaction-reaction (see Results and Fig. 2a).

Finally, $Pr(Data | Model)$ is computed as:

$$Pr(D | M) = \prod_{i \in D^+} p_i \prod_{j \in D^-} (1 - p_j) \prod_{k \in D^0} [w \cdot p_k + (1 - w) \cdot (1 - p_k)]$$  \hspace{1cm} (4)$$

The probabilities $p$ are those associated with the genetic interaction model as described above. $D^+$ denotes the subset of $D$ that are observed positive genetic interactions, $D^-$ the subset that are observed negative, and $D^0$ the remaining node pairs that have not yet been observed positive or negative for genetic interaction. For each unobserved node pair, we compute the expected probability contribution to $Pr(D | M)$ given the background probability $w$ of genetic interaction for node pairs chosen at random. We set $w = 0.0078$, estimated from the systematic analysis by Tong et al.\textsuperscript{98} which revealed 291 genetic interactions among 37,376 node pairs tested (double gene knockouts).

**7.2.3 Search Procedure**
Two searches were performed: first, to identify a set of e-paths as candidate explanations for the observed genetic interactions; and second, to find the particular subset of these e-paths that optimized the model score $\Pr(D|M)$. In the first search, an $A^*$ algorithm\textsuperscript{18} was used to identify the shortest and near-shortest e-paths connecting node $a$ to $b$ for each genetically interacting gene pair $(a,b)$. The $A^*$ heuristic function was chosen as the distance of the shortest path from node $p \rightarrow b$ as precomputed using ALLPAIRS-SHORTEST-PATHS\textsuperscript{18}; note that this heuristic does not distinguish among e-paths vs. non e-paths, so that the shortest e-path connecting $(a,b)$ may not be the shortest path overall. Searches were terminated for paths longer than three edges or after a maximum duration of 500 seconds.

In the second search, an optimal-scoring subset of e-paths was chosen for the model using a greedy hill-climbing procedure as follows:

1) Initialize the model with one e-path chosen at random and compute $\Pr(M | D)$.

2) For each e-path $e$, compute $\Pr(M' | D)$ where:

   $M' = M - e$ if $e \in M$

   $M' = M + e$ if $e \notin M$

3) If $M' > M$ for some $e$, choose the new model $M$ as the $M'$ with maximal probability score and go to step 2).

4) Otherwise, record $M$ as a local maximum and re-initialize at step 1).

The model and associated search procedure were implemented in a plug-in to the network visualization and analysis package Cytoscape; the plug-in is included with Supplementary Information.
7.3 Results

7.3.1 Construction of the Genetic and Physical Interaction Networks

We assembled a genetic interaction network from two primary data sets (Fig. 2a). The first consisted of 664 synthetic lethal and 490 suppressor interactions culled from the literature and catalogued in the MIPS database. The second data set was derived from a systematic screen pairing each of nine gene deletions with each of ~4,700 gene deletions possible in yeast, resulting in 291 gene pairs with observed synthetic lethal interactions and a remaining 38,126 pairs screened for which synthetic lethal interactions were observed to be absent. The single genetic network synthesizing these data consisted of 5,124 protein nodes linked by 39,518 edges of type suppressor, synthetic-lethal, or non-synthetic-lethal.

We assembled a corresponding physical interaction network, consisting of 5,453 protein and 463 biochemical reaction nodes connected together by edges of four types: 15,116 protein-protein interactions, 5,869 protein-DNA transcriptional interactions, 878 protein-reaction relationships, and 1,743 reaction-to-reaction transformations (Fig. 2a). Protein-protein pairwise interactions were obtained by downloading the entire contents of the DIP database as of June, 2003. These included data from high-throughput co-immunoprecipitation and two-hybrid experiments as well as low-throughput information mined from the literature. Protein-DNA interactions were obtained from a recent large scale chromatin-immunoprecipitation study in yeast (interactions with \( p \leq 0.001 \)); these linked protein transcription factors to the proteins whose genes were directly bound by these transcription factors. Biochemical reaction pathways were obtained from KEGG; to encode reaction pathways in the network, reactions were
defined as nodes and neighboring reactions having at least one metabolic product or substrate in common were linked to one another (reaction↔reaction edge). In addition, proteins in the network with enzymatic activity were linked to all reactions that they were known to catalyze (protein→reaction edge).
A. GENETIC NETWORK
- Type: Source: Number:
  - Suppressor: MIPS: 1,200
  - Synthetic lethal: MIPS: 1,109
  - Synthetic lethal: SCA: 291
  - NOT synth. lethal: SCA: 18,000

PHYSICAL NETWORK
- Type/Direction: Source: Num.: Prob.
  - Protein-protein: DIP: 1,500, see S. cerevisiae
  - Protein-DNA: CHE: 5,390
  - Protein-DNA: XIG: 5,000, 0.9
  - Reaction: EFG: 5,000, 0.9

UNIVERSE OF ePATHS

NETWORK MODEL
(H1, H2, ... , Hn)

PREDICTIONS
(1, 2, 3, ... , n)

Probabilities of genetic interaction for each gene pair

B. Diagrams illustrating genetic and physical interactions

C. Types of interactions:
   - [i] Transcriptional cascade
   - [ii] Transcriptional coregulation
   - [iii] Protein interaction chain
   - [iv] Biochemical reaction

GENETIC
- synthetic lethal → suppressor →

PHYSICAL
- protein-protein → protein-DNA → protein-reaction → reaction-reaction →
Figure 7.2: Overview of the network modeling approach. [a] Genetic and physical networks are integrated using the Cytoscape platform: data sources, numbers of interactions, and interaction probabilities (physical only) are given for each network. After generating the universe of available physical pathway explanations (ePaths), the model is built on the subset of ePaths that optimally explains the genetic network. [b] The optimal scoring model, bird’s eye view; [c] Examples of physical pathways elucidated in the model, categorized into four types of explanations [i]-[iv].

Although particular reaction-reaction or protein-protein edges could be directed, as in the case of a non-reversible reaction or where one protein phosphorylates another, in the absence of additional information we considered that information could flow in either direction and therefore modeled each interaction \((a,b)\) of these types as two reciprocally directed edges in the network \((a \rightarrow b \text{ and } b \rightarrow a)\). Also, for compactness, metabolites were not explicitly represented, and reactions sharing only general cofactors such as ATP, \(H_2O\), or NAD were not linked (a complete list of excluded cofactors is given in Supplementary Information).

99.5% of nodes (5962 out of 5993) in the physical network fell into a single large connected component (in which every pair of nodes is connected through a path of sufficient length). The genetic network had 74% of nodes (723 out of 972) in the largest connected component. The physical and genetic networks showed high overlap: in 94.4% of genetic interactions (1289 out of 1365 non-redundant genetic interactions), both interactors also encoded proteins in the physical network.

7.3.2 Explaining Genetic Relationships with Physical Interaction Paths

With the integrated physical/genetic network as a basis, we used our network modeling approach to identify the set of physical interaction pathways which best explained the observed genetic interactions. An initial search of the network identified
6,628 explanatory physical paths of length 1, 2, or 3 for which the path endpoints were genetic interactors. As shown in Table 1, this number was approximately 4.7 times greater than that expected if the nodes in the genetic network were randomly permuted (eliminating functional correspondence between the physical and genetic networks while preserving their topologies; mean±stdev = 1415.6±168.3 over 10 random permutations¹). Taken together, the physical paths were found to explain 529 out of the 1365 observed genetic interactions (vs. 168.3±1.6 at random), with 21% of physical paths explaining more than one genetic interaction (vs. 14.2±0.004% at random). Thus, a significant and large number of physical paths were found to explain the observed genetic interactions, and physical paths were enriched in their ability to explain multiple genetic interactions in a single path.

Table 7.1: Number of genetic interactions (and significance) by type and length

<table>
<thead>
<tr>
<th>PATHWAY TYPE / LENGTH</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>ALL LENGTHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein interaction chain</td>
<td>69</td>
<td>154</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>(1.2±1.0)</td>
<td>(25.1±7.2)</td>
<td>(0.0±0.0)</td>
<td>(0.0±0.0)</td>
<td>(0.0±0.0)</td>
<td>(25.6±7.3)</td>
</tr>
<tr>
<td>Biochemical reaction</td>
<td>0</td>
<td>36</td>
<td>24</td>
<td>79</td>
<td>269</td>
<td>303</td>
</tr>
<tr>
<td></td>
<td>(0.0±0.0)</td>
<td>(0.2±0.4)</td>
<td>(2.7±1.5)</td>
<td>(39.2±9.1)</td>
<td>(183.0±26.7)</td>
<td>(190.1±25.7)</td>
</tr>
<tr>
<td>Transcriptional regulation (UNI-directed)</td>
<td>7</td>
<td>13</td>
<td>86</td>
<td>280</td>
<td>552</td>
<td>573</td>
</tr>
<tr>
<td></td>
<td>(0.6±1.3)</td>
<td>(5.0±3.5)</td>
<td>(28.4±16.5)</td>
<td>(166.7±32.8)</td>
<td>(433.0±41.7)</td>
<td>(457.9±42.3)</td>
</tr>
<tr>
<td>Transcriptional regulation (BI-directed)</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>34</td>
<td>66</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>(0.0±0.0)</td>
<td>(0.2±0.4)</td>
<td>(1.5±2.5)</td>
<td>(15.9±12.3)</td>
<td>(44.7±21.4)</td>
<td>(45.3±21.8)</td>
</tr>
<tr>
<td>ALL TYPES</td>
<td>76</td>
<td>196</td>
<td>109</td>
<td>316</td>
<td>599</td>
<td>684</td>
</tr>
<tr>
<td></td>
<td>(1.8±1.5)</td>
<td>(30.1±9.3)</td>
<td>(61.4±17.6)</td>
<td>(195.7±32.1)</td>
<td>(484.5±43.6)</td>
<td>(514.8±44.2)</td>
</tr>
</tbody>
</table>

¹ The nine most highly-connected nodes in the genetic network were used as ‘baits’ in the synthetic lethal screen by Tong et al. and may have been chosen for screening based on prior knowledge of interactions in the physical network. To conservatively correct for any hidden biases or dependencies between the genetic and physical networks introduced by this experimental design, these nine nodes were not permuted during network randomization.
For each genetic interaction, we assumed that no more than one or a few explanatory paths could be related to the true mechanism of interaction. Accordingly, we formulated a probabilistic scoring function to evaluate a set of explanatory paths by the quality of physical interactions along each path and by the numbers of observed positive genetic interactions the paths correctly explain and negative genetic interactions they correctly reject. As described further in Methods, the scoring function is governed by two key parameters: \( \alpha \), the probability that a physical path causes genetic interaction for a node pair it explains, and \( \beta \), the (much lower) probability that a physical path causes genetic interaction for a node pair it fails to explain. For a given choice of parameters, a search is performed to find the set of explanatory paths that maximizes the scoring function; this optimal subset is called "The Model". We found that the set of paths incorporated in the model was identical over 10 searches begun from different entry points—thus, it is likely that the search is returning the global optimum for the present parameters and input data considered.

Model performance was assessed over a range of \((\alpha,\beta)\) values using standard five-way cross-validation as detailed in Fig. 3. A maximum sensitivity of 6.9% was obtained by including all 6,628 available explanatory paths in the model. However, as shown in Fig. 3, sensitivity does not scale linearly with model size, such that 50% of maximum sensitivity can be obtained for models incorporating less than 5% of explanatory paths. By comparison, if model predictions were random, the expected sensitivity would be approximately 0.78%, estimated as the overall frequency of positive genetic interactions observed in the synthetic-lethal screen by Tong et al.\(^8\) (291 positives out of 38,126 gene pairs screened).
Figure 7.3: Model performance over a range of parameters. Five-way cross-validation was performed for combinations of model parameters $\alpha=(0.01, 0.1, 0.2)$ and $\beta$ in the range ($10^{-12} \leq \beta \leq 10^{-4}$). In cross-validation, the set of observed positive and negative genetic interactions was randomly partitioned into five subsets of equal size (39,518 / 5 ≈ 7,904 interactions each), from which four were designated as training data and one as test data. Model sensitivity was computed as (true positive predictions / all observed positives) among the positive and negative genetic interaction data comprising each cross-validation test set, averaged over the five test sets. The scatter plot shows sensitivity vs. total number of physical paths in the model at each parameter combination. Adjacent values in the $\beta$ series are connected by lines along fixed values of $\alpha$. Although the best sensitivities were achieved at small $\beta=1 \times 10^{-12}$, the highest sensitivity-to-paths ratio was for parameters ($\alpha=0.1; \beta=5 \times 10^{-7}$).

Model parameters were fixed at ($\alpha=0.1; \beta=5 \times 10^{-7}$) in order to maximize the ratio of model sensitivity to model size, i.e., to correctly predict positive interactions in the test set while incorporating as few physical paths as possible. At these parameters, the optimal scoring model included 257 paths explaining 293 genetic interactions, and it made 219 new genetic interaction predictions (see below). Model sensitivity and specificity were 5.2% ($p<0.0005$) and 99.9%, respectively. Because such a small fraction of node pairs were genetic interactors, model specificity was also extremely high for a
random model; sensitivity is the more significant measure of model performance in this case.

7.4 Discussion

7.4.1 The Physical Pathway Model Associates New Functions with Genetic Interactions.

To better understand the cellular roles represented by the model, we associated functional annotations from the Gene Ontology (GO) database with the proteins involved in each genetic interaction and physical pathway explanation\(^103\). Of the 507 genetic interactions with physical explanations in our model, we found that 207 occurred between genes with identical annotations in the GO database. Thus, some insight into these genetic interactions is provided by looking at the GO functions alone. Although the remaining genetic interactions were between genes with different GO annotations, 160 were associated with particular physical pathways in the model whose genes were collectively enriched for specific GO functions (\(p<0.001\) using the hypergeometric test). Thus, integration of the physical and genetic networks allows us to associate functions with a much greater number of genetic interactions than would be possible by looking at the genetic network alone.

Complete GO annotations for each genetic interaction and associated physical explanation(s) in the model are given in Supplementary Information.

The above analysis left 140 genetic relationships, which could not be associated with a functional category. We hypothesized that some of these might represent links between two distinct but functionally related cellular processes acting in a complementary fashion.
(also known as extrinsic interactions). Indeed, over the 140 gene pairs, we found two significant functional pairings: genes involved in cellular import paired more frequently than expected with genes involved in vacuolar or vesicular/Golgi transport ($p<4.3 \times 10^{-5}$), and the same type of relationship was found for RNA transcription paired with cell death or cell rescue ($p<1.2 \times 10^{-4}$).

7.4.2 Protein Interaction Chains are the Predominant Explanation for Genetic Interactions.

As shown in Fig. 2c, explanatory paths could be classified into one of four types depending on the interactions they contained: (i) metabolic pathways having one or more biochemical reactions; (ii) protein interaction chains consisting of protein-protein interactions only; (iii) transcriptional cascades having one or more protein-DNA interactions in series; and (iv) transcriptional co-regulation consisting of protein-DNA interactions in parallel. Of these, we found that protein interaction chains were by far the predominant type of explanatory path in the model: of the 6,628 explanatory paths found in the initial search, 4,709 were protein interaction chains (the breakdown for the remaining paths was 132 metabolic reactions, 501 transcriptional cascades, and 501 transcriptional co-regulation). Of course, these frequencies are highly dependent on the prior probabilities (false positive rates) assigned to each type of interaction. However, this trend towards protein interaction chains is also reflected in the literature—of 50 randomly-sampled genetic interactions catalogued in the Yeast Protein Database, all 34 for which the mechanism of interaction was known were attributable to protein-protein interactions.
7.4.3 Genetic Interaction Predictions fill in Gaps of Knowledge in the Existing Genetic Network.

The genetic interaction predictions increased the overall connectivity of the genetic network in functionally significant ways. Specifically, to be useful, genetic interactions must span functional groups, relating higher-level information. Although the network of observed genetic interactions initially consisted of 87 regions (connected components), the genetic interaction predictions made by the model frequently bridged two regions of common function such that 21 of these regions collapsed to a total of eight. Surveying for significance in terms of gene ontology, we found 7 of the 8 to have associated p-values of $p<0.001$ (with Bonferroni correction for multiple hypothesis testing).
7.4.4 Significant Enrichment for Essential Proteins Downstream of Explanatory Paths.

If explanatory paths in the physical interaction network are an accurate model of genetic interactions, $\Omega$ nodes at the intersection, would be expected to be enriched for essential genes. In a comparison of downstream proteins, we find that 23% of $\Omega$’s are essential in model paths, compared to 9% in randomized paths ($p < 0.001$).

7.5 Conclusion

As data generation becomes ever more systematic and high-throughput, the burden is placed on computational methods and models to systematically distill hypotheses from the data at a comparable rate. Here, we have shown that genetic relationships may be reconciled with, and at a high level explained by, a model built from pathways through a physical interaction network. We have developed a generalized tool for performing this search and for assembling the pathways most likely to explain the genetic data into a physical model. This model provides hypotheses for the physical mechanisms underlying the observed synthetic lethal or suppressor phenotypes and can be validated by way of the new genetic relationships it implies. Computational models of this type will be crucial for modeling pathway function in model organisms, in which thousands of genetic interactions have been observed but unexplained, and for the study of diseases and other phenotypes arising from genetic linkage or drug interactions in humans.
8. Dissection of Complex Traits in S.Cerevisiae Using Small-Molecules

We demonstrate that natural variants of the yeast *Saccharomyces cerevisiae* can be a model system for the study of complex traits, which are encoded by multiple genes with graded effects and are exemplified by many human diseases. In contrast to artificial single-gene knockout libraries of *S. cerevisiae* widely used to study individual gene function, we used 314- and 1932-member libraries of mutant strains generated by meiotic recombination to study the cumulative, quantitative effects of natural mutations on phenotypes induced by 23 small molecules. This approach reveals unexpected relationships between calcium ion homeostasis and respiratory processes. The combination of natural variants of yeast and small molecules that modulate evolutionarily conserved cellular processes can enable a better understanding of the general features of eukaryotic complex traits.

8.1 Introduction

Most genetic traits, including common human diseases such as cancer and schizophrenia, are complex, i.e. are encoded by allelic variants of at least two different genes whose individual contributions vary from small to large, i.e. have graded effects. Current theories regarding complex traits range from one extreme, the common disease/common variant model, which assumes that common alleles at a few loci underlie complex traits, to the other, the genetic heterogeneity model, which assumes that rare alleles at many loci underlie complex traits (1). Here we use natural genetic variation, as opposed to artificial mutations, in the budding yeast *Saccharomyces cerevisiae* as a means to study complex traits – specifically, the response of cells to small molecules. The use of small molecules to reveal complex traits was first suggested by studies of
metabolism in *Neurospora* over 60 years ago \(^{115}\). More recently, *S. cerevisiae* has been exploited as a means to study global mRNA levels as a complex trait \(^{116,117}\), among others \(^{118,119}\).

### 8.2 Methods

We used 23 small-molecule perturbagens (SMPs) (Table S1), each of which differentially inhibits the growth of one or more of three diploid yeast strains: S1029, YAG040, and XHS123\(^7\) (Haggarty *et al.* manuscript in preparation). S1029 is a homozygous diploid, isogenic to the commonly used laboratory strain S288c, the only *S. cerevisiae* strain with a complete genome sequence. YAG040 is also a homozygous diploid, derived from a clinical isolate found 20 years ago in the lung of an AIDS patient in San Francisco\(^{120}\). XHS123 is the heterozygous diploid formed by mating haploid spores derived from S1029 and YAG040. XHS123 and YAG040 were constructed with artificial dominant drug resistance markers, kanMX and hphMX, respectively, both of which serve as standards of Mendelian, single-gene traits.

The 23 SMPs are chemically diverse with respect to molecular descriptors such as molecular weight, cLogP, and the number of rotatable bonds\(^{121}\); likewise, they differ in their reported mechanisms of action. They are evenly divided between natural products and synthetic products, and seven are FDA-approved drugs. Many have been studied exclusively in the context of mammalian cells, for example, the central nervous system-active compounds flunarizine and pimozide. These are shown here to inhibit potently the growth of yeast in a genotype-dependent and heritable manner.

We generated and then perturbationally profiled libraries comprising 314 and 1932 random XHS123 meiotic segregants in order to characterize the complexity of
resistance or sensitivity to each SMP. We sporulated XHS123 and manually arrayed segregants from dabs of single colonies grown on agar plates into individual wells of 384-well plates containing rich media. In tandem and as a control, we sporulated YAG040, wherein both alleles at each locus are identical. Population-level growth of each segregant was measured in duplicate and as a function of cell density using a spectrophotometer (OD$_{600}$), at two time points post-inoculation in clear-bottom, 384-well plates containing rich media incubated at room temperature with one of the SMPs (Fig. 1A,D,E).

Duplicate measurements were averaged and arranged in a matrix, $S$, consisting of an ordered array of rows and columns. Each row ($x_i$; where $i = 1$ to $m$) corresponds to growth in the presence of a particular SMP; each row ($y_j$; where $j = 1$ to $n$) corresponds to a segregant. Accordingly, an element $(m, n)$ of $S$ encodes information about segregant $m$ for phenotypic descriptor $n$. In addition to the XHS123 segregants, we included the parental (S1029 and YAG040) and hybrid strains (XHS123).

8.3 Results

Generally, growth inhibition by a small molecule segregates either in a Mendelian (i.e., monogenic) or a complex (i.e., polygenic) fashion. As expected for a Mendelian trait, $1/2$ of YAG040 were resistant to hygromycin B, and $1/2$ were sensitive. Moreover, a fraction of hygromycin B-resistant segregants was sensitive to the mating pheromone alpha factor, which arrests haploid MATa cells (Fig. 1 B). Similarly, $1/2$ of XHS123 segregants were resistant to G418 but unable to grow without supplemental uracil, or vice versa (Fig. 1 C).
We next determined whether growth inhibition by each SMP segregates in a Mendelian or complex fashion. Raw OD_{600} measurements of 314 segregants were rank-ordered and compared to the control Mendelian trait, resistance to G418. We observed that rank-ordered distributions of monogenic traits exhibited a steep transition centered at the halfway mark, dividing the population into two phenotypic groups, one fully sensitive, the other fully resistant (Fig. 2A,B,F).

Growth inhibition by most SMPs was found to be complex. Only the segregation of resistance to flunarizine, a Ca^{2+} channel blocker, appears to be monogenic. This result was also borne out in tetrad analysis, where a 2:2, resistance/sensitivity segregation was observed. However, at a second time point, \( \frac{1}{4} \) segregants remained sensitive while \( \frac{3}{4} \) became resistant, suggesting a more complex genetic underpinning to resistance to flunarizine: two sensitizing alleles, one fast-acting (48 h), the other slow-acting (72 h) (Fig. 2B). In addition to being complex, growth inhibition by the majority of SMPs also has a kinetic component. The response to rapamycin exemplifies the variegated temporal and quantitative nature of complex growth inhibition by a small molecule (Fig. 2C).

As another control, we observed both Mendelian and complex growth inhibition that was not induced by a SMP, but rather by constituents of stress media: growth in maltose-containing media was Mendelian, while growth in synthetic wine must (SWM), which combines pH, non-ionic osmotic, and nitrogen-source stresses, and rich media containing 0.9M NaCl, which induces ionic osmotic shock, was complex (Fig 2D,E,F).

We estimated the complexity of growth inhibition by each SMP: the approximate number of genetic determinants that underlie resistance (or sensitivity). In order to gain greater statistical resolution, we examined the 1932-member library of random XHS123
meiotic segregants and focused on a narrow subset of SMPs. The ranked-order
distributions of MATα segregants in the presence of G418, H2O2, tunicamycin, and
doxorubicin illustrate a series of four complex traits underlain by increasing numbers of
genetic determinants (Fig. 2A). If we assume that the fraction of the resistant segregants
approximates $\frac{1}{2^n}$, where $n$ equals the number of underlying genetic determinants at a
given kinetic snapshot, then a single gene underlies resistance to G418, 2 underlie
resistance to H2O2, 3 underlie resistance to tunicamycin, and 4 underlie resistance to
doxorubicin. The number of genetic determinants that underlie resistance fell within this
range in the case of all 23 SMPs.

After establishing that sensitivity to a small molecule can be a complex trait, we
addressed whether the same genetic determinants were responsible for resistance to all
SMPs, or whether a unique set of genes selectively confers resistance to each SMP.
Several key observations affirm the latter. First, we observed unlinked segregation of
resistance to flunarizine and to G418 (Fig. 1C) in the 314-member library of segregants,
as well as in tetrad dissections. Second, two-dimensional hierarchical clustering (2-D
HAC) of the 314-member library revealed that recombination had both randomly and
independently re-assorted the genetic determinants underlying resistance or sensitivity to
each SMP (Fig. 3A). In order to illustrate this point more clearly, we compared the
responses of the parental and hybrids and some nearest neighbor segregants at higher
magnification (Fig. 3B). Upon closer inspection, the parental and hybrid responses to
nearly each of the 23 SMPs differed (Fig. 3B). Remarkably, however, there were
instances where recombination preserved the parental and hybrid resistances and/or
sensitivities, except at one or a few positions (Fig. 3B). For example, one of the two
segregants that is a nearest neighbor of YAG040, C01, inherited resistance to tunicamycin, an inhibitor of protein glycosylation, while the other, D11, did not (Fig. 3B). This result is true for the majority of SMPs, and argues that a unique set of genetic determinants selectively confers resistance to each SMP.

Similar results were obtained using principal component analysis (PCA). PCA yields principal components, the first three of which captured nearly half (46.5%) of the variance in the dataset (supplemental). PCA was performed on both the segregants and the SMPs, yielding ‘biological’ (Fig. 3D) and ‘chemical’ (Fig. 4A) spaces, respectively. In contrast to 2D-HAC, PCA sharpened the focus on those SMPs that comprise two clusters, the modulators of Ca$^{2+}$ homeostasis and of respiratory processes, while at the same time rendering a visualization that reaffirmed random and independent assortment of the various underlying genetic determinants.

A two-dimensional plot of the first two principal components, PCA1 and PCA2, reveals four diffuse clusters of segregants: the major contributors to PCA1 are the genetic determinants underlying resistance to the modulators of Ca$^{2+}$ homeostasis, while the major contributors to PCA2 are the genetic determinants underlying resistance to G418. When the third principal component, PCA3, whose major contributors are the genetic determinants underlying resistance to the modulators of respiratory processes, was included on a third axis, the resulting biological space revealed distributed points with no significant clumping.

The response of the 314 segregants to the remaining SMPs comprises over half of the total variance in the dataset. In other words, every segregant was different both qualitatively and quantitatively with respect to growth inhibition by every SMP.
Therefore, each segregant is a unique genetic reagent, analogous to a single-gene knockout, which can be studied further using established genomic and proteomic techniques.

Next, we examined the SMPs for which the underlying genetic determinants of resistance or sensitivity appeared to be the same. 2-D HAC and PCA rendered four clusters of SMPs, three of which contained SMPs that have been described previously as targeting the same proteins or pathways in cells ("functional analogs"). The first cluster consists of flunarizine, pimozide, calcimycin, curcumin, and penitrem A, all modulators of calcium ion homeostasis (Fig. 4B). For example, flunarizine, indicated for migraines, and pimozide, indicated for Tourette’s syndrome, are reported to block the mammalian T-type Ca\textsuperscript{2+} channel, while calcimycin is a calcium ionophore naturally produced by *Streptomyces chartreusensis* that causes a rapid accumulation of intracellular Ca\textsuperscript{2+}. Curcumin, a naturally occurring small molecule from *Curcuma longa*, is reported to increase the level of Ca\textsuperscript{2+} accumulation by inhibiting Ca\textsuperscript{2+}-ATPase, and has recently been suggested as a treatment for cystic fibrosis (20). Penitrem A, a small molecule produced by bacteria belonging to the genus *Penicillium*, inhibits Ca\textsuperscript{2+}-activated K\textsuperscript{+} channels.

The second cluster consists of menadione, FCCP, LY83583, and 1,9-pyrazoloanthrone, two known and two putative modulators of respiratory metabolism. Menadione, also known as vitamin K3, is an inducer of oxidative stress. FCCP is a proton ionophore that diminishes the mitochondrial membrane potential. LY83583 and 1,9-pyrazoloanthrone are inhibitors of soluble guanylate cyclase and JNK II kinase, respectively, though the former has not been cloned in yeast, while the latter has not been characterized in yeast. Overall, our results suggest that this procedure also enables
insights into mechanisms of small molecules since small molecules with unknown mechanisms of action can be grouped together with those having known mechanisms.

The third cluster contains two antagonists of a master regulator of calcium-mediated signaling, calmodulin CMD1: chlorpromazine and E6-berbamine. Chlorpromazine is an antipsychotic therapeutic reported to exerts its clinical effects via inhibition of calmodulin, though it has also been described as D2 dopamine receptor antagonist. E6-berbamine is a derivative of berbamine, an isoquinoline alkaloid produced by plants of the genus Berberis and reported to be an inhibitor of calmodulin.

The fourth cluster contains the remaining SMPs, all of which are reported to inhibit growth by distinct mechanisms of action: rapamycin, tunicamycin, manumycin, anisomycin, and doxorubicin, modulators of the TOR proteins, protein glycosylation, farnesyltransferase activity, translation, and topoisomerase II, respectively.

We determined the degree of linkage between functional analogs on the basis of a hypergeometric p-value linkage test between pairs of traits. In its simplest form, we hypothesize that phenotypic linkage occurs when the same gene or group of genes confers resistance or sensitivity to multiple compounds. We also determined the converse of phenotypic linkage, phenotypic anti-linkage. One explanation for phenotypic anti-linkage is that reciprocal segregation of two functionally opposite alleles of the same genetic locus or group of loci (e.g. kanMX::URA3) is occurring.

We observed significant phenotypic linkage between SMPs in each of the three aforementioned clusters of functional analogs. One of the strongest linkages (r=0.893) in the entire dataset was observed between flunarizine and calcimycin: 144 (46%) are resistant to calcimycin, 94 (30%) segregants are resistant to flunarizine, 93 of which are
co-resistant to both (p-value $6 \times 10^{-43}$); an overlap of 43 (14%) is expected by chance
\[46\% \times 30\% = 14\%\] (Fig. 4C). Strong linkage ($r=0.414$) also exists between menadione and
FCCP, two modulators of respiratory metabolism: 84 (27%) segregants are resistant to
menadione, 34 (11%) segregants are resistant to FCCP; of this 34, 27 are also resistant to
menadione (p-value $1 \times 10^{-11}$), when only 9 (3%) are expected. The two modulators of
calmodulin are phenotypically linked ($r=0.622$): 114 (36%) are resistant to
chlorpromazine, 39 (12%) are resistant to E6-berbamine, 35 of which are co-resistant to
both (p-value $2 \times 10^{-13}$); 14 (4%) are expected by chance.

Resistance to flunarizine and calcimycin, two modulators of Ca$^{2+}$ homeostasis,
appears to be phenotypically linked to wortmannin and anisomycin, a PI-3 kinase
inhibitor and translational inhibitor, respectively, neither of which impinges directly on
Ca$^{2+}$ homeostasis. There is precedent for such a connection, though no such connection
between natural variations in genes in these cellular processes has been reported (21).

Two blocs of resistance stand out as anti-linked: dominant resistance to 3
compounds that affect respiratory metabolism that is encoded by YAG040-derived
determinants, and recessive resistance to 2 compounds that affect Ca$^{2+}$ homeostasis that
is encoded by S288c-derived loci (Fig. 3B). We are unaware of anything about Ca$^{2+}$
homeostasis in the YAG040 background that can explain its sensitivity, and vice versa
for S1029. However, S1029, as an S288c-derived strain, harbors a mutation in the gene
$HAPI$, which results in constitutively low levels of cytochrome c$^{125}$, while YAG040
encodes the wild-type copy of $HAPI$. The sensitivity of S1029 to modulators of
respiratory metabolism is consistent with it having a fragile respiratory apparatus.
34 segregants are resistant to FCCP, 144 are resistant to flunarizine, but only 4 are co-
resistant to both (p-value 3.5x10^4) when 16 are expected by chance (Fig. 4C).
Simultaneous treatment of segregants with both FCCP and flunarizine gave an identically
low overlap, indicating that perturbation of Ca^{2+} homeostasis is epistatic to perturbation
of respiratory processes. Mutually exclusive resistance is corroborated by PCA: just as
G418 and CSM-URA inhabit opposite corners of chemical space, so do FCCP and
flunarizine (Fig. 4A). This anti-linkage is validated by extensive literature that documents
the dependence of mitochondrial function on calcium^{126}, and suggests that Ca^{2+}
homeostasis plays a role in the evolutionary adaptation of *S. cerevisiae* to different
environments.

8.4 Conclusions

This work illustrates a flexible, scalable, low-cost, chemical-genetic method for
revealing complex traits in budding yeast. Cumulatively, the SMPs we used perturb a
wide swath of cellular processes but, individually, the genes with which they interact are
distinct, except in the case of functional analogs whose common reported mechanisms of
action we validated. The extensive characterization of natural genetic variation with small
molecules serves as a counterpoint to the use of *de novo* or artificial mutations in a fixed
genetic background. The combination of natural variants of yeast and small molecules
that modulate evolutionarily conserved cellular processes may have several applications:
(i) the identification of mechanisms of action of small molecules; (ii) delineation of the
roles of different alleles in drug sensitivity ("pharmacogenomics")^{127}; (iii) the
classification and characterization of genetic traits (Mendelian vs. complex); (iv) and the
genetic bases of complex traits (human disease models).
Notes

1. Parental diploid strains S1029, an S288c derivative (MATa/MATagal2/gal2 hoD::natMX4/hoD::natMX4), YAG040 (HO/hoD::hphMX4 MATa/MATa), and XHS123 (URA3/ura3D::kanMX4), all of which were generous gifts from John H. McCusker, UNC, Chapel Hill, NC.

2. The kanMX resistance cassette encodes a phosphotransferase that confers resistance to the aminoglycoside G418 (geneticin), while replacing the S288c-derived copy of URA3, a gene involved in uracil biosynthesis; the hphMX resistance cassette encodes a phosphotransferase that confers resistance to the aminoglycoside hygromycin B, while replacing one copy of the HO gene, the endonuclease responsible for mating-type switching.

3. See the NCI-sponsored public database ChemBank (chembank.med.harvard.edu) for complete structural annotations and mechanisms of action.

4. We used the reciprocal segregation of the kanMX and URA3 markers as a positive indicator of recombination, and discarded products of sporulation that were neither G418^SEN/URA+ nor G418^RES/URA-. We arrived at 314 segregants in the smaller library because at least 314 of 380 (83%) products of sporulation passed the recombination test. We arrived at 1932 segregants in the larger library because at least 1932 of 2944 (66%) products of sporulation passed the recombination test. The discarded products of sporulation are either contaminating diploids (G418^RES/URA+) or G418^SEN/URA-.

5. XHS123 is HO/ho, which means we expect approximately 50% of segregants to be haploid (ho) and 50% diploid (HO) as a result of self-diploidization. We find no significant effect of ploidy (5) on growth, except a subtle growth advantage to homozygous diploid recombinants that is undetectable by the 48hrs time point. For added stringency, we considered 375 MATa haploid segregants in our estimation of polygenic degree on the basis of sensitivity to the mating pheromone alpha factor.

6. PCA consists of a linear transformation of an original system of axes formed by the n- (or m-dimensions of the matrix of molecular descriptors. This transformation takes the form of a multidimensional rotation that preserves Euclidean distances. The directions of rotation are determined by considering the normalized covariance matrix D as a linear operator and computing a set of eigenvectors and corresponding eigenvalues that satisfy the eigenvalue equation: \( D\Psi = \Lambda\Psi \). The resulting eigenvectors (\( \Psi_n \)) of the matrix \( \Psi \) form a set of new, linearly independent axes, called principal components (PCs), each of which corresponds to an axis of the n-dimensional ellipsoid spanning the multivariate distribution of the original data. The corresponding eigenvalues (\( \lambda_1-\lambda_n \)) of matrix \( \Lambda \) account for progressively smaller fractions of the total variance in the original data. See I. T. Jolliffe. Principal Component Analysis. New York: Springer, 2002.
**Figure 1.** Overview of the generation and pertubational chemical profiling of random meiotic segregants of XHS123. (A) Two haploid yeast strains – in this study, derivatives of S1029 and YAG040 – are mated in order to generate a F1 heterozygous diploid, which is subsequently sporulated, yielding F1 segregants. Averaged, quantitative growth measurements of random segregants in multiple conditions is arranged into a matrix and computationally analyzed. (B) Segregation of a single Mendelian trait illustrated by a heatmap (red indicates saturating growth, or resistance; green indicates no growth, or sensitivity) of clustered $OD_{600}$ measurements of 358 YAG040 segregants grown in rich media containing 300$\mu$g/mL of the antibiotic hygromycin B or 10 $\mu$M alpha factor. (C) Segregation of two unlinked Mendelian traits illustrated by a heatmap of clustered $OD_{600}$ measurements of 314 XHS123 segregants grown in complete synthetic media lacking uracil (CSM-URA); rich media containing 200$\mu$g/mL G418; rich media containing 50$\mu$g/mL flunarizine. (D) Digital optical scan of clear-bottom, 384-well plate containing 366 random YAG040 segregants. (E) 3x6 cross-section of plate in (D); empty wells and wells containing saturating amounts of yeast are contrasted.

**Figure 2.** Rank-ordered distributions of XHS123 segregants demonstrate that growth inhibition by small molecules is polygenic, kinetic, and quantitative. (A) 375 XHS123 MATa segregants (17) are plotted on a one-dimensional graph, where the x-axis denotes the fraction of the segregant population and the y-axis denotes normalized $OD_{600}$ values: a rich media containing 200$\mu$g/mL G418; b0.03% $H_2O_2$; c 200ng/mL tunicamycin; d 25pg/mL doxorubicin. Hashed vertical lines appear at 1/2, 3/4, 7/8, and 15/16 marks. (B) Rank-ordered distribution of 314 XHS123 segregants in 50pg/mL flunarizine. The red curve corresponds to measurements taken at 48hrs; the blue curve corresponds to measurements taken at 72 hrs. The chemical structure of flunarizine appears in upper left corner of the graph. (C) Rank-ordered distribution of 314 XHS123 segregants in ng/mL rapamycin. The red curve corresponds to measurements taken at 48hrs; the blue curve corresponds to measurements taken at 72 hrs. The chemical structure of rapamycin appears in the upper left corner of the graph. (D) Rank-ordered distribution of 314 XHS123 segregants in rich media supplemented with 0.9M NaCl. (E) Rank-ordered distribution of 314 XHS123 segregants in synthetic wine must. (F) Rank-ordered distribution of 314 XHS123 segregants in rich media containing 2% maltose.

**Figure 3.** Dimensionality reduction by 2-D HAC and PCA demonstrates random and independent assortment of genetic determinants underlying resistance or sensitivity to each SMP. (A) Two-dimensional clustergram of a 314 x 23 matrix; each column is a small molecule; each row is a segregant; red indicates saturating growth, or resistance; green indicates no growth, or sensitivity. (B) Small-molecule complex trait profiles of S1029 ("S"), XHS123 ("X"), YAG040
("Y"), and two XHS123 segregants, D11 and C01, which are closely related to YAGO40. (C) Two-dimensional PCA plot of 314 random XHS123 segregants. PCA1 is to the first principal component segregant; PCA2 is the second principal component. S1029, YAGO40, and XHS123 are indicated by black lines. (D) Three-dimensional PCA plot of 314 random XHS123 segregants. PCA1, PCA2, and PCA3 are the first, second and third principal components, respectively. S1029, YAGO40, and XHS123 are indicated by black lines. Green dots are closer to viewer, red dots recede into the page.

**Figure 4.** Small molecules that target related cellular processes ("functional analogs") are phenotypically linked and co-cluster in chemical space. (A) Three-dimensional PCA plot of 25 SMPs. F1 corresponds to factor 1; F2 corresponds to factor 2; F3 corresponds to factor 3. Green tacks include controls and functional analogs; asterisks indicate compounds with accompanying structures: YBA50=penitrem A; YBA17=amiodarone; YBA16=anisomycin; YBA5=curcumin; YBA2=calcimycin. (B) Structures of a group of modulators of Ca\(^{2+}\) homeostasis. (C) Venn diagram depicting phenotypic linkage and anti-linkage. The red circle corresponds to growth in 50μg/mL flunarizine; the blue circle corresponds to growth in 25μg/mL FCCP; the green circle corresponds to growth in 25μg/mL calcimycin. The numbers of in the overlapping region correspond to the observed number co-resistant segregants. The numbers in parentheses corresponds to the expected number of co-resistant segregants. P-values of intersections are indicated with arrows.

This work was done in collaboration with Ethan Perlstein, and under the joint supervision of Bonnie Berger and Stuart Schreiber. Figures are not attached due to size, and can be obtained upon request.
9. Future Directions

Many of the future directions in this thesis are at the convergence of mathematics and medicine. Arguably, the clinical benefits of approaches similar to those espoused here, will be realized in the field of targeted, tailored therapeutics. In order to start making guesses about the future of clinical agents, it is necessary to take a look back at the successes and failures that gave rise to the medications currently in use. Fortunately, many of the roadblocks and pitfalls that have slowed the progress of drug development in the last century show signs of abating with the aid of new technologies and approaches.

Approaches to treatment, management, and cure of human disease come in many forms: surgical interventions, vaccination, treatment with small-molecules and proteins, and even targeted exposure to radiation. Although genetics and genomics may tangentially affect other areas (e.g. through better patient stratification for surgical interventions), their impact will most likely be felt most profoundly in the use of small-molecule and protein therapeutics. Thus, we begin with a history of modern chemotherapy.

9.1 History of Chemotherapy

The development of chemotherapy is often attributed to Paul Erlich and Sahachiro Hata for the use of arsphenamine (an arsenic containing compound) in the treatment of syphilis, starting in 1909. Erlich hypothesized that a molecule could be found with antibiotic effects. By screening hundreds of arsenic-containing compounds, Erlich and Hata found a number capable of inhibiting microbial growth, the most potent of which was arsphenamine. Not only was this the first chemotherapeutic, the approach pioneered
by Erlich is the foundation on which all current screening approaches are based. Today, chemotherapy primarily refers to the use of small-molecules in the treatment of cancer. In the latter half of the twentieth century, new chemotherapeutics have led to curative strategies to a number of malignancies, including: childhood acute lymphoblastic leukemia, Hodgkins’ and non-Hodgkins lymphoma, and gestational choriocarcinoma.

The modern era of cancer chemotherapeutics began during World War II. Soldiers accidentally exposed to mustard gas were found to have low white blood cell counts. The hypothesis at the time was that white blood cells were preferentially targeted due to their high growth rate relative to other tissues. Extending this logic, it was surmised that mustard gas (and more generally, the so-called “nitrogen mustards”, a more general class of compounds structurally similar to mustard gas) might be used in the treatment of cancers of the white blood cells. Early experimental success in a murine lymphoma model by Alfred Gillman and Louis Goodman further supported the use of nitrogen mustards as cancer therapies. Unfortunately, testing in humans demonstrated only a transient, but substantial, effect on tumor size. Although nitrogen mustards did not provide a curative strategy for lymphoma, it began a search for compounds capable of selectively targeting and killing fast growing cells: cytotoxic chemotherapies.

Another class of promising chemotherapeutics was found through observation of a similar unexpected occurrence. Folic acid supplements, when given to children with Acute Lymphoblastic Leukemia (ALL), were found to accelerate the course of the disease. Unlike the serendipitous discovery of a potential therapeutic, the hunt for an inhibitor of folic acid metabolism is considered the first example of rational drug design. Since the metabolism of folate, and intermediary metabolism as a whole, were only
beginning to be elucidated, it was inefficient to look for a molecule capable of blocking production of folate. Instead, Sidney Farber of Children’s Hospital and Harvard Medical School, and Harriett Kilte and Yellapragada Subbarao of Lederle Laboratories, synthesized a number of folate-analogues meant to compete with the native molecule in binding to enzymes requiring folate as a substrate or cofactor. Aminopterin and amethopterin showed significant effects, inducing the first recorded cases of remission in leukemia. However, as with nitrogen mustards, the effects were transient, and though the path forward was clear, new molecules were required. Although, amethopterin (commonly called methotrexate) is not curative in ALL in isolation, it is still used in a variety of contexts: as part of combination chemotherapeutics in the treatment of ALL, and more recently in the treatment of a number of auto-immune diseases such as ankylosing spondylitis, inflammatory bowel disease (Crohn’s disease and Ulcerative Colitis), psoriasis, psoriatic arthritis, and rheumatoid arthritis. Interestingly, a decade after Farber’s discovery, methotrexate was found to be curative in the treatment of gestational choriocarcinoma, a rare, but aggressive malignancy of trophoblastic cells of the placenta. Previous to treatment with methotrexate, choriocarcinoma was a rapidly fatal disease, most often due to early hematogenous spread to the brain and kidneys. Today, methotrexate in combination with other chemotherapeutics (etoposide, actinomycin D, cyclophosphamide, and oncovin) has produced a cure rate of >95% in choriocarcinoma. In fact, even advanced cases with substantial metastases carry a good prognosis.

Farber’s results were met with scorn and disbelief upon publication in 1953. One of the few to believe Farber was Joseph Burchenal at Memorial Sloan Kettering Cancer
Center in New York City. Burchenal, working in conjunction with Farber, reproduced
the results using methotrexate in 1955. Attempting to generalize the approach that had
found methotrexate, and find molecules with longer-lasting remissions and curative
potential, Burchenal began to look at other targets. Working closely with George
Hitchings and Gertrude Elion of Burroughs Welcome, who produced new analogues to a
wide assortment of metabolites, Burchenal began to search for molecules that could
interfere with any process required for cell growth and proliferation: alkylating agents to
damage DNA, nitrogen mustards, and nucleotides. Though skeptical of Farber’s
findings, many others began to look for more promising chemotherapeutics targeted at
folate synthesis and other related metabolic targets. The brilliant approach employed by
Burchenal and his collaborators allowed them to look far more broadly than their
competition (some estimate that they investigated as many as 100-200 metabolites):
instead of synthesizing analogues from scratch, they started with the target molecule of
interest and modified it using a number of simple reactions. Not only could they look at a
much larger number of targets, they were also able to produce many more molecules per
target. Additionally, the intricacies of purine metabolism and catabolism had recently
been elucidated by Greenberg, Buchanan, and Kornberg. Early success with 2,6-
diaminopurine validated their approach. However, 2,6-diaminopurine proved too toxic
for use in humans. To reduce the side effects, Hitchings and Elion modified the structure
of 2,6-diaminopurine to produce 6-mercaptopurine (6-MP). The first clinical studies of
6-MP showed full remission of one third of patients with ALL, with another ten percent
showing partial remission. The FDA approved 6-MP soon after.
Despite the dramatic impact of 6-MP and methotrexate on approaches to treating cancer, they were not curative. Soon after the discovery of 6-MP, chemists at Eli Lilly discovered that alkaloids in an extract from Madagascar Periwinkle were capable of inhibiting tumor growth. This group of compounds was referred to as the “vinca alkaloids” (derived from Vinca Rosea, the plant’s proper name), including vincristine, was later shown to disrupt microtubule polymerization thereby inhibiting cell division. The next major step in cancer treatment came in 1965, when James Holland, Emil Freireich, and Emil Frei began to use methotrexate, 6-MP, vincristine, and prednisone in combination. They surmised that cancer treatment was analogous to the treatment of bacterial infections with antibiotics: after prolonged courses using a single agent, the bacteria became resistant. By combining multiple antibiotics with different mechanisms of action, consider success had been achieved. Conceivably, this was mediated through the bacteria’s ability to generate resistance to one form of inhibition, but relative inability to become multiply resistant before the infection was cleared. The POMP regimen, as it came to be known, has led to cure rates greater than 80% in ALL.

Since the advent of combination chemotherapy, many new agents have been developed. Additionally, adjuvant chemotherapies have greatly increased the survival and cure rates when used in support of surgical strategies in breast, colon, and lung cancer.

Despite the success of combination cytotoxic chemotherapies in the treatment of ALL, and later in Hodgkin’s disease and non-Hodgkin’s lymphoma, many malignancies remain largely incurable. And though the 30 years following the work of Holland, Freireich, and Frei showed relatively few new avenues in cancer treatment, the last ten
years have completely changed the landscape of drug development. Since the initial use of chemotherapeutics, many of the underpinnings of cancer, as both a germ-line and a somatic genetic disease, have been uncovered. From this wealth of information, a number of promising drug targets and associated inhibitors emerged. Just as the first use of nitrogen mustards gave a first glimpse of the promise of cytotoxics, Gleevec, Iressa, and Herceptin demonstrate how much remains to be done (change).

Chronic myelogenous leukemia (CML) was the first malignancy found to have a clear genetic underpinning. In 1960, Peter Nowell and David Hungerford showed that a reciprocal translocation between chromosome 9 and 22 was present in leukemic cells from patients with CML. As both were from Philadelphia, Pennsylvania, the genetic defect in CML came to be known as the Philadelphia chromosome. Toward a functional understanding of this finding, subsequent work demonstrated that the Abl kinase on chromosome 9 was fused to the Breakpoint Cluster Region on Chromosome 22 giving rise to an abnormal protein product (BCR-Abl), that retained Abl’s tyrosine kinase activity. Unlike native Abl kinase, BCR-Abl is constitutively active, due to the replacement of Abl’s first exon, with sequences from the BCR gene. leading to its oncogenic effects. After it was shown in experimental systems that BCR-Abl formation was necessary and sufficient to cause CML, it became a very attractive target for which to find a new inhibitor. Before molecules targeted at BCR-Abl were discovered, CML was treated with a combination of bone marrow transplantation (BMT) and interferon therapy.

In the mid-1990's chemists at the drug company Novartis discovered a number of molecules targeted at the Plate-Derived Growth Factor Receptor (PDGF-R) kinase and
Abl kinase. The first molecules discovered, 2-phenylaminopyrimidines, were later optimized and derivatized in the search for more potent PDGF-R inhibitors. However, the most potent molecules were found to be dual PDGF-R and Abl inhibitors.

Eventually, one lead molecule was chosen for clinical study: Imatinib (Gleevec). Brian Druker, an oncologist at Oregon Health Sciences University, performed the first clinical investigation into the use of Imatinib for the treatment of CML. Early studies indicated that not only did Imatinib inhibit PDGF-R and BCR-Abl, it also showed similar activity in inhibition of c-Kit, a target known to be important in Gastrointestinal Stromal Tumors (GIST). Druker found that Imatinib could inhibit from 92-98% of growth in BCR-Abl positive cells in vitro, with minimal inhibition of normal cells. Unlike standard chemotherapies, reliant on quickly killing malignant cells during short cycles of treatment due to the substantial side effects (elaborate somewhere), Druker demonstrated that suppression of BCR-Abl cells would require continued inhibition. Specifically, in vitro, BCR-Abl positive cells could be rescued from apoptotic cell death after temporary exposure to Imatinib. This suggested a dramatic departure from the short burst formulation of chemotherapy, toward a chronic, well tolerated, orally available treatment.

This was further supported by data from a murine CML model, wherein mice requires three treatments of Imatinib per day to show benefit (yielding cure rates of 87-100%), and once a day treatments were largely ineffective. Following toxicity testing in further animal studies, Druker began phase I/II clinical trials in patients resistant to interferon therapy. Not only were no dose-limiting toxicities observed, all patients showed complete hematological responses (normal myeloid counts), and many showed cytological remission (absence of Philadelphia Chromosome on cytogenetic analysis).
Unfortunately, as with ALL, treatment of CML with Gleevec provides only temporary relief, and most patients' disease begins to progress approximately 18 months after they start on Gleevec. A second generation of inhibitors is now becoming available to treat patients no longer responsive to Gleevec: including Dasatinib and PD184352. Interestingly, Desatinib is known to have a wider set of targets than just BCR-Abl, including a number of other kinases. In fact Desatinib has substantially more off target effects than Gleevec. Though initially viewed as a lower quality inhibitor due to these off target effects, evidence is mounting that broad spectrum kinase inhibitors may actually be more effective in the long run. This effect has been likened to combination therapies, but mechanistic explanations are still nebulous at best.

Since the early work by Druker, Gleevec has been widely regarded as the model for the development of targeted therapies. Despite considerable effort, it has been difficult to reproduce this success in other malignancies. The predominant explanation is the nature of the underlying disease that separates many liquid tumors (hematological malignancies) from solid tumors. Liquid tumors are often characterized by single transforming events: as in CML, a single translocation leads to disease, presumably in the absence of any other genetic insult. Moreover, this lesion, is necessary and sufficient, and is observed in every patient with CML. This singular constellation makes the choice of drug target very simple. However, in most solid tumors there is little consensus as to the causative lesion or set of lesions, making the choice of drug target extremely difficult. This problem is very much analogous to the difference between rare, monogenic, Mendelian diseases such as cystic fibrosis or sickle-cell anemia, and common diseases with complex genetic underpinnings such as hypertension or myocardial infarction: with
the former, the causes have been clearly defined, potentially focusing therapeutic strategies, however with the latter, there is little to work with in the way of high-confidence therapeutic targets. In many ways, the somatic alterations critical to the (beginning) of cancer, may provide a springboard from which to understand complex traits. Liquid tumors (like mendelian genetics), solid tumors (complicated somatic genetics), and complex traits (rewrite). One of the first steps toward targeted therapies in solid tumors, came through a series of unexpected findings in Non-Small Cell Lung Cancer.

A variety of mutations in the epidermal growth factor receptor (EGFR) have been documented in a number of different malignancies: amplification, overexpression, and constitutive activation. Gefitinib, a first generation EGFR inhibitor designed by AstraZeneca, began clinical use for treatment of Non-Small Cell Lung Cancer (NSCLC) in 2002. Although overall efficacy was limited, a number of patients experienced dramatic responses. Further studies showed that the group of Gefitinib responders shared certain traits, some of which were uncommon in patients with NSCLC. Specifically, response rates were higher in women than men, Japanese patients than in a European-derived population, nonsmokers, and patients with adenocarcinoma (as opposed to large and squamous cell carcinoma). The dramatic disparity prompted two groups to search for a genetic explanation.

Using similar sequencing-based approaches, the groups of Daniel Haber and Matthew Meyerson both demonstrated that activating mutations in EGFR correlated perfectly with patient response to Gefitinib. Though there are many interesting technical details involved in these discoveries, the conclusions are overwhelming. Although these
properties were discovered serendipitously, Gefitinib was capable of allele-specific action.

On the surface, this example does not appear novel: there are numerous cases of small molecules whose side effects are tied to a particular allele – for example, 6-mercaptopurine (6-MP). However, what truly differentiates Gefitinib, is not simply its genotype specific effect, but that this specificity is achieved through a molecularly specific interaction. Specifically, in cases such as 6-MP the effect is mediated through intrinsic deficiencies in enzymatic function leading to slower processing of the molecule. In contrast, the concept of Gefitinib might be extended: by creating a molecule targeted at each allele observed in the clinic, each individual be treated with a medication tailored to their illness. Unfortunately, this has not yet been achieved, but to fully realize the benefit of such next generation treatments will require a change in the mechanisms meant to translate promising leads into clinical successes.

9.2 Targeted Trials

With knowledge of the mutations that confer sensitivity to Gefitinib, it is now possible to determine immediately whether or not a patient will respond to Gefitinib treatment. This suggests a dramatic departure from the current approach to clinical trials. Specifically, current clinical trials are meant to distinguish the relative benefit of a new treatment when compared to the standard of care. If the effect is small, clinical trials required to detect these differences can required tens of thousands of participants to be sufficiently powered to make robust conclusions. The result is a therapeutic development process that can take decades and cost hundreds of millions to billions of dollars. Just as
Gefitinib highlights the promise of targeted therapeutics, so does it highlight the weaknesses of current clinical trial structure.

Due to terrible prognosis for lung cancer patients, Gefitinib was approved contingent on eventual completion of a Phase III clinical trial. As clinicians were beginning to notice the surprising effects in certain subpopulations, Gefitinib was failing in Phase III trials. This can be simply explained in the following way: if only 1% of lung cancer patients have a Gefitinib-sensitizing mutation, Gefitinib will be ineffectual in 99% of cases. Though this suggests it makes a poor replacement for the standard of care, it also hides the efficacy within a small subgroup of patients. In contrast to this situation, consider the situation where patients are stratified in a trial according to their “mutation status.” To demonstrate beyond doubt statistically, that molecules such as Gefitinib work within their target group requires not tens of thousands of people, but less than one hundred!

9.3 Modularly Targeted Therapeutics

Accomplishing this new paradigm in drug development requires one dramatic departure long before a molecule reaches the clinic: the ability to prospectively find molecules targeted at the correct cellular moieties. This will require not only better methods to determine molecular targets, but also the means to survey targets more broadly.

Decades of research in large pharmaceutical companies and small biotech firms alike, has focused on a handful targets. Moreover, these targets have been studied largely at the exclusion of other untested targets. Our investigations to this point suggest that
many perturbations to a given function can yield similar outcomes. Thus, the scope of targets around a disease phenotype of interest can be dramatically enlarged. Not only will this increase the number of targets, it will open the door to rational discovery of combination therapies. For example, molecules targeted at the same module would provide a natural starting point for developing a novel multi-component treatment. Hopefully this approach would dramatically increase the pace of drug discovery.
Bibliography:


95. Sham P. Shifting paradigms in gene-mapping methodology for complex traits.
