

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Mapping Drug Chemistry from the Viewpoint of Small Molecule Metabolism

Permalink

<https://escholarship.org/uc/item/1c07f96j>

Author

Adams, James Corey

Publication Date

2009

Peer reviewed|Thesis/dissertation

Mapping Drug Chemistry from the Viewpoint of Small Molecule Metabolism

by

James Corey Adams

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2009
by
James Corey Adams

for my family,
who have made all things possible

Acknowledgements

This dissertation is mainly composed of two first-author manuscripts. The work presented in **Chapter 2** was performed under the guidance of Patricia C. Babbitt in collaboration with co-first author Michael J. Keiser. The manuscript has been submitted for publication to PLoS Computational Biology and is in review. It is reproduced here with permission from MJK, LB, DSL, OGW, and PCB. Additional material is available through an online resource at <http://sea.docking.org/metabolism>. The work presented in **Chapter 3** was also performed under the guidance of Patricia C. Babbitt.

To my advisor Patsy Babbitt for the enormous freedom and personal support that she has provided over the years, to our fantastic collaborators Mike Keiser and his advisor Brian Shoichet, and to all the Babbitt lab members, it has been an honor to work with you;

To K.T. Moortgat and the UCSF Center for BioEntrepreneurship, Douglas Crawford and QB3, Peter Mui, Rebecca Seal, and Rana Datta who help build the UCSF Innovation Accelerator, and the numerous other members of the UCSF entrepreneurial community who have inspired so many of us to translate our research from the lab to market;

To my family, my parents for their steadfast patience and encouragement, Gabrielle and April for taking the pressure off, my brother Jonathan for proving that it is not genetic, Jon and Lisa for riding and walking beside me across the rooftop of the

world, Ed for his easy humor and much-needed perspective, and especially my sister Noël always willing to wash away my woes with wine;

To so many friends over many years, especially E.L., A.Y., M.H., E.J., A.E.T., S.O., and A.P.W., dearer to my heart than I have ever found the words to say;

Thank you.

Abstract

Small molecule drugs target many small molecule metabolic enzymes in humans and pathogens, often mimicking endogenous ligands. The effects may be therapeutic or toxic, but are frequently unexpected. A large-scale mapping of the intersection between drugs and metabolism is needed to better guide drug discovery. To map the intersection between drugs and metabolism, we have grouped drugs and metabolites by their associated targets and enzymes using ligand-based set signatures created to quantify their degree of similarity in chemical space. The results reveal the chemical space that has been explored for metabolic targets, where successful drugs have been found, and what novel territory remains. Chemical similarity links between drugs and metabolites predict potential toxicity, suggest routes of metabolism, and reveal drug polypharmacology. To aid other researchers in their drug discovery efforts, we have created an online resource of interactive maps linking drugs to metabolism that enable easy navigation of the vast biological data on potential metabolic drug targets and the drug chemistry currently available to prosecute those targets. Thus, this work provides a large-scale approach to ligand-based prediction of drug action in small molecule metabolism.

Furthermore, this work challenges a fundamental dogma in modern molecular biology – the presumption that individual protein structural and chemical requirements are the dominant constraints in small molecule metabolic enzyme evolution. We directly test that assumption by weighing the absolute and relative constraints imposed by structural homology, metabolic pathway context, and transcriptional coregulation. We believe this work is the first to explicitly argue – from the molecular level perspective of

genomic data – that selection constrains enzyme evolution as much at the level of metabolic pathway organization as it does at the level of individual protein structure.

Table of Contents

| | |
|---|----|
| Chapter 1. Introduction for a General Audience..... | 1 |
| Chapter 1. Technical Introduction | 3 |
| Chapter 2. A Mapping of Drug Space from the Viewpoint of Small Molecule Metabolism | 7 |
| 2.1 Abstract | 8 |
| 2.1.1 Background | 8 |
| 2.1.2 Methodology/Principal Findings..... | 8 |
| 2.1.3 Conclusions/Significance | 8 |
| 2.3 Results | 12 |
| 2.3.1 Drug-metabolite links recapitulate known drug-target interactions | 12 |
| 2.3.2 Human drug “effect-space” maps detail interactions between drug classes and enzyme targets | 16 |
| 2.3.3 Species-specific effect-space maps for drug discovery in pathogens..... | 19 |
| 2.3.4 Case study: MRSA | 19 |
| 2.4 Discussion | 23 |
| 2.5 Conclusion..... | 25 |
| 2.6 Methods..... | 25 |
| 2.6.1 Compound sets | 25 |
| 2.6.2 Ligand sets..... | 26 |
| 2.6.3 Drug sets..... | 26 |
| 2.6.4 Set comparisons..... | 27 |
| 2.6.5 MRSA essentiality and synthetic lethal analysis..... | 27 |
| 2.7 Acknowledgments | 27 |
| 2.8 References | 37 |
| Chapter 3. Global Predictors for the Evolutionary Rates of Enzymes | 43 |
| 3.1 Abstract | 43 |
| 3.2 Background | 44 |
| 3.2.1 Introduction | 44 |
| 3.2.2 A Pairs-based Approach..... | 48 |
| 3.3 Results | 49 |
| 3.3.1 Expression Level Accounts for Largest Percentage of Evolutionary Rate Variance ... | 50 |
| 3.3.2 Evolutionary Rates Correlate among SCOP Superfamily Members | 51 |

| | |
|---|----|
| 3.3.3 Evolutionary Rates Correlate in Adjacent Metabolic Network Enzymes | 52 |
| 3.3.4 Expression Levels Correlate in Adjacent Metabolic Network Pairs – But Not Transcriptional Module Pairs | 55 |
| 3.3.5 The Independence of Network Context and SCOP Superfamily | 56 |
| 3.4 Discussion | 58 |
| 3.4.1 The Independence of SCOP Superfamily and Network Context from Expression Level | 60 |
| 3.4.2 The Independence of SCOP Superfamily from Network Context | 61 |
| 3.4.3 SCOP Superfamily | 62 |
| 3.4.4 Metabolic Network Context | 64 |
| 3.5 Conclusion..... | 65 |
| 3.6 Methods..... | 66 |
| 3.6.1 Evolutionary Metrics | 66 |
| 3.6.2 Biochemical Network..... | 67 |
| SCOP | 68 |
| 3.6.3 Transcriptional Modules..... | 68 |
| 3.6.4 Statistical Methods | 69 |
| 3.7 Acknowledgments | 70 |
| 3.8 References | 77 |
| 3.9 Epilogue References..... | 83 |
| Appendix A. The Chemical Diversity of Drugs and Metabolites..... | 84 |
| A.1 Chemical similarity among drugs and among metabolites..... | 84 |
| A.2 Topological differences between drug and metabolic space | 86 |
| Appendix B. MRSA Growth Assays | 93 |
| B.1 Rationale..... | 93 |
| B.2 Experimental Design..... | 94 |
| B.3 Discussion..... | 94 |

List of Figures

| | |
|--|----|
| Figure 2.1 Similarity Ensemble Approach (SEA)..... | 28 |
| Figure 2.2.A Effect-space map – Nucleoside reverse transcriptase inhibitors (NRTIs)... | 29 |
| Figure 2.2.B Effect-space map – Dihydrofolate reductase (DHFR) inhibitors..... | 29 |
| Figure 2.2.C Effect-space map – Thymidylate synthase (TS) inhibitors..... | 29 |
| Figure 2.3 Effect-space map showing chemical similarity between drugs and metabolites in MRSA..... | 30 |
| Figure 2.4 Essential/synthetic lethal map of MRSA metabolism..... | 31 |
| Figure 3.1 Principal component analysis (PCA) of evolutionary rate (dn) in small molecule metabolic enzymes yields two highly significant components..... | 71 |
| Figure 3.2 Pearson correlation between small molecule metabolic enzyme pairs according to constraint: metabolic network, SCOP superfamily, and transcriptional module..... | 72 |
| Figure 3.3 Pearson correlation between small molecule metabolic enzyme pairs with orthogonal constraints: metabolic network only, SCOP superfamily only, metabolic network and SCOP superfamily combined..... | 73 |
| Figure A.1.A Distribution of MDDR drug set links to MetaCyc reaction sets..... | 89 |
| Figure A.1.B Distribution of MetaCyc reaction set links to MDDR drug sets..... | 89 |
| Figure A.1.C Distribution of chemical similarity links within MDDR..... | 90 |
| Figure A.1.D Distribution of chemical similarity links within MetaCyc..... | 90 |
| Figure A.2.A MetaCyc metabolic network..... | 91 |
| Figure A.2.B MDDR drug network..... | 91 |
| Figure A.2.C MRSA metabolic | 91 |
| Figure A.3 Quantifying chemical diversity..... | 92 |
| Figure B. 1 MRSA growth inhibition by AOAA and MFT..... | 95 |

List of Tables

| | |
|--|----|
| Table 2.1 Metabolic enzyme targets and their best links to MDDR..... | 32 |
| Table 2.2 Selected best hits between MetaCyc reaction sets and MDDR drug sets..... | 33 |
| Table 2.3 Selected links between human metabolic reactions and current drugs..... | 34 |
| Table 2.4 Selected links between MDDR drug classes and human folate and pyrimidine metabolism..... | 35 |
| Table 3.1 Principal component analysis yields multiple variables that together explain variation in metabolic enzyme evolutionary rates..... | 74 |
| Table 3.2 Expression level measures correlate in adjacent metabolic network pairs..... | 75 |
| Table 3.3 Pathway context and structural superfamily independently correlate with metabolic enzyme evolutionary rates..... | 76 |

Chapter 1. Introduction for a General Audience

All humans, plants, and animals use enzymes to metabolize food for energy, build and maintain the body, and get rid of toxins. Drugs used to clear infections or cure cancer often target enzymes in bacteria or cancer cells, but the drugs can interfere with the proper function of human enzymes as well. Recent studies have mapped drugs to enzymes and many other targets in humans and other organisms, but have not focused on metabolism. **Chapter 2** presents a new method to predict what enzymes drugs might affect based on the chemical similarity between classes of drugs and the natural chemicals used by enzymes. We have applied the method to 246 known drug classes and a collection of 385 organisms (including 65 National Institutes of Health Priority Pathogens such as malaria, anthrax, and the plague) to create maps of potential drug action in metabolism. We also show how the predicted connections can be used to find new ways to kill pathogens and to avoid unintentionally interfering with human enzymes.

Some enzymes change slowly over time, while others evolve rapidly. **Chapter 3** investigates what determines this evolutionary rate. The blueprints for enzymes and other proteins are found in genes. How much a gene is turned on (i.e. expression level) is known to be a major predictor of how fast it will evolve. However, many other potential predictors also exist. How hard is the enzyme's job? How important is the job – can you live without it? How stable is the enzyme in stressful environments? Think, for example, of the wildly colorful and heat-loving bacteria growing happily in the scalding pools of Yellowstone National Park. Previous

studies have focused on these conventional predictions of how rapidly an enzyme will evolve over time.

In **Chapter 3** we challenge this fundamental dogma of molecular biology. Many enzymes work together in sequences called metabolic pathways. Together they produce essential molecules, such as the building blocks of DNA and proteins, or to get rid of dangerous toxins. Because enzymes work together as a group, we hypothesize that an enzyme's metabolic pathway neighbors help predict an enzyme's evolutionary rate. This idea of "group selection" has been extremely controversial in evolutionary biology for decades. Group selection usually refers to groups of animals such as herds of antelope or schools of fish. However, here we apply the concept to groups of enzymes that together contribute to metabolic fitness. We present for the first time evidence that molecular level group selection is at least as important as individual enzymes in predicting evolutionary rate.

Chapter 1. Technical Introduction

1.1 Drugs and metabolism

The breathtaking chemical diversity of biological systems pales in comparison to the breadth of all carbon-based chemistry. The estimated number of possible carbon-based molecules with molecular weight under 500 daltons, the range of most common metabolites, exceeds 10^{60} . Yet the substrates and products of small molecule metabolism, for all organisms studied to date, number only on the order of 10^4 . These molecules represent a small but highly active slice of biological chemistry. In the search for new drugs, chemical similarity to these metabolites frequently may suggest an increased likelihood of *in vivo* perturbation, and ultimately therapeutic effect. The chemical similarity may be broad and non-specific, or indicate direct molecular mimicry of endogenous compounds. Examples of these anti-metabolites include antibacterial sulfonamides, nucleoside reverse transcriptase inhibitors used in antiviral therapy, and antineoplastics such as methotrexate and 5-fluorouracil. A wealth of biological knowledge about metabolism allows further filtering through vast chemical diversity to target only essential enzymes and their related drug chemistry. For example, genetic studies in numerous organisms have revealed lethal and synthetic lethal gene knockouts that could be targeted based upon chemical similarity to endogenous substrates. Orthologs to these same therapeutic targets from pathogens can be sources of toxicity when accidentally targeted in humans. For these reasons, the intersection between drug chemistry and small molecule metabolism and drug chemistry merits comprehensive analysis to better guide drug discovery.

Development of a method to predict the interaction between drugs and small molecule metabolic enzymes would help avoid drug toxicity in humans and facilitate the discovery of new drugs against pathogens. In **Chapter 2**, we present a new approach to predicting drug action by comparing the chemical similarity between known drug classes and the endogenous substrates and products of metabolic reactions. The results reveal the chemical space that has been explored for drugs against metabolic targets, where successful drugs have been found, and what unexploited territory remains. Specifically, based solely upon ligand bond topology, we predict the likely array of metabolic drug targets in humans for each drug class from a large collection of drug classes reported in the patent literature. We also apply our approach to a large collection of model organisms and NIH Priority Pathogens. Using this information, we created an online resource of interactive metabolic maps to explore the vast biological data on potential metabolic drug targets and the drug chemistry currently available to prosecute those targets. These drug “effect space” maps can be used by other researchers to navigate the connections between known drug classes and metabolism. To demonstrate how our method may be applied to drug discovery against emerging pathogens, we present a case study in MRSA that integrates synthetic lethal analysis with our ligand-based chemical similarity approach to recover many of the known drug targets in MRSA.

JCA conceived of, designed, and performed the experiments mapping drugs to metabolism, analyzed the data, and wrote the paper; MJK developed the initial SEA method, designed experiments mapping drugs to metabolism, and developed the computational tools used in this study, LB performed MRSA experiments under the guidance of HFC, DSL performed the essentiality and synthetic lethal analyses under the guidance of OGW, and PCB directed and supervised the research.

1.2 Evolution and Metabolism

What determines the evolutionary rate of a gene? Many hypotheses have been offered, but the precise determinants remain elusive. Expression level is a dominant factor, but additional properties include functional class, stability, dispensability, and pleiotropy. Integrated analyses that both demonstrate significant correlations with evolutionary rate and distinguish the proportion of variance a given factor may explain constitute the next step in further elucidating the principles of protein evolution. **Chapter 3** outlines just such an approach in revealing the predictors of evolutionary rate among small molecule metabolic enzymes. In addition to gene expression, we hypothesize that two key selectable properties of metabolic systems – structural similarity in homologs of dissimilar function and the output of metabolic pathways – represent important determinants that uniquely distinguish metabolic pathways from other biological systems. First, investigation of structural homologs of dissimilar function allows dissection of the structural constraints on enzyme evolution from overlapping constraints on their functional roles in metabolic pathways. Second, the chemical conversion of a substrate, via multiple reactions within a metabolic pathway, into a biologically useful product provides the primary readout of metabolic fitness.

Chapter 3 challenges a fundamental dogma in modern molecular biology – the presumption that individual protein structural and chemical requirements are the dominant constraints in small molecule metabolic enzyme evolution. That assumption is directly tested by weighing the absolute and relative constraints imposed by structural homology, metabolic pathway context, and transcriptional coregulation. We believe this work is the first to explicitly argue – from the molecular level perspective of genomic data – that selection constrains enzyme

evolution as much at the level of metabolic pathway organization as it does at the level of individual protein structure.

Chapter 2. A Mapping of Drug Space from the Viewpoint of Small Molecule Metabolism

JC Adams^{1*}, MJ Keiser^{2*}, L Basuino³, HF Chambers³, DS Lee^{4,5,6}, OG Wiest⁷, PC Babbitt^{8†}.

¹Graduate Program in Pharmaceutical Sciences and Pharmacogenomics, University of California, San Francisco, CA 94158-2550

²Graduate Program in Bioinformatics, University of California, San Francisco, CA 94158-2517

³San Francisco General Hospital, University of California San Francisco, San Francisco CA 94110

⁴Center for Complex Network Research and Departments of Physics, Biology, and Computer Science, Northeastern University, Boston, MA 02115

⁵Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115

⁶ Department of Natural Medical Sciences, Inha University, Incheon 402-751, Korea

⁷Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, 46556

⁸Departments of Biotherapeutic Sciences and Pharmaceutical Chemistry and California Institute for Quantitative Biosciences, University of California, San Francisco, CA 94158-2550

* Co-first author

† Corresponding author

Patricia C. Babbitt, corresponding author
University of California, San Francisco
Mission Bay QB3 Building,
Suite N508E
1700 4th Street, Box 2550
San Francisco, CA 94143-2550
(CA 94158 for direct delivery by courier)
Tel +1 (415) 476-3784
Tel Assistant +1 (415) 514-4261
Email: babbitt@cgl.ucsf.edu

2.1 Abstract

2.1.1 Background

Small molecule drugs target many core metabolic enzymes in humans and pathogens, often mimicking endogenous ligands. The effects may be therapeutic or toxic, but are frequently unexpected. A large-scale mapping of the intersection between drugs and metabolism is needed to better guide drug discovery.

2.1.2 Methodology/Principal Findings

To map the intersection between drugs and metabolism, we have grouped drugs and metabolites by their associated targets and enzymes using ligand-based set signatures created to quantify their degree of similarity in chemical space. The results reveal the chemical space that has been explored for metabolic targets, where successful drugs have been found, and what novel territory remains. To aid other researchers in their drug discovery efforts, we have created an online resource of interactive maps linking drugs to metabolism. These maps predict the “effect space” comprising likely target enzymes for each of the 246 MDDR drug classes in humans. The online resource also provides species-specific interactive drug-metabolism maps for each of the 385 model organisms and pathogens in the BioCyc database collection.

2.1.3 Conclusions/Significance

Chemical similarity links between drugs and metabolites predict potential toxicity, suggest routes of metabolism, and reveal drug polypharmacology. The metabolic maps enable interactive navigation of the vast biological data on potential metabolic drug targets and the drug chemistry

currently available to prosecute those targets. Thus, this work provides a large-scale approach to ligand-based prediction of drug action in small molecule metabolism.

2.2 Introduction

Drug developers have long mined small molecule metabolism for new drug targets and chemical strategies for inhibition. The approach leverages the “chemical similarity principle” [1] which states that similar molecules likely have similar properties. Applied to small molecule metabolism, this principle has motivated the search for enzyme inhibitors chemically similar to their endogenous substrates. The approach has yielded many successes, including antimetabolites such as the folate derivatives used in cancer therapy, and the nucleoside analog pro-drugs used for antiviral therapy. However, drug discovery efforts also frequently falter due to unacceptable metabolic side-effect profiles or incomplete genomic information for poorly characterized pathogens [2,3,4].

With the recent availability of large datasets of drugs and drug-like molecules, computational profiling of small molecules has been performed to create global maps of pharmacological activity. This in turn provides a larger context for evaluation of metabolic targets. For example, Paolini et al. [5] identified 727 human drug targets associated with ligands exhibiting potency at concentrations below 10 μ M, thus creating a polypharmacology interaction network organized by the similarity between ligand binding profiles. Keiser et al. [6] organized known drug targets into biologically sensible clusters based solely upon the bond topology of 65,000 biologically active ligands. The results revealed new and unexpected pharmacological relationships, three of which involved GPCRs and their predicted ligands that were subsequently confirmed *in vitro*. Cleves et al. [7] also rationalized several known drug side effects and drug-drug interactions based upon three-dimensional modeling of 979 approved drugs. However, despite the clear rationale and past successes in applying ligand-based approaches to drug

discovery, global mapping between drugs and small molecule metabolism, the goal of this study, has been hindered by both methodological challenges and incomplete genomic information. The relatively recent availability of metabolomes for numerous organisms allows a fresh look on a large scale [8,9,10,11,12,13].

In this work, we link the chemistry of drugs to the chemistry of small molecule metabolites to investigate on a large scale the intersection between small molecule metabolism and drugs. The Similarity Ensemble Approach (SEA) [6] was used to link metabolic reactions and drug classes by their chemical similarity, measured by comparing bond topology patterns between sets of molecules. Two types of molecule sets are used in this work. The first comprises drug-like molecules known to act at a specific protein target, and the second comprises the known substrates and products of an enzymatic reaction. While this approach is complementary to target and disease focused methods [5,14,15,16,17,18,19,20,21,22,23], neither protein structure nor sequence information is used in the comparisons. Thus, these links provide an orthogonal view of metabolism based only upon the chemical similarity between existing drug classes and endogenous metabolites.

To provide the results in the context of metabolism, drug “effect-space” maps were also created. For each of the 246 drug classes investigated in this work, effect-space maps enable visualization of the chemical similarities between drugs and metabolites painted onto human metabolic pathways, allowing a unique assessment of potential drug action in humans. In addition, to aid target discovery in pathogens, 385 species-specific effect-space maps were created to show the predicted effect-space of currently marketed drugs painted onto metabolic

pathways representing target reactions in model organisms and pathogens. Examples of these maps are provided below and their applications in predicting drug action, toxicity, and routes of metabolism are discussed. To enable facile exploration of the drug-metabolite links established by this analysis, interactive versions of both sets of maps are available at <http://sea.docking.org/metabolism>.

Finally, we consider the role of these effect space maps in addressing the challenge of new drug target discovery. We have integrated our results with those from essentiality and synthetic lethal analysis and applied this information retrospectively to assess metabolic drug targets in methicillin-resistant *Staphylococcus aureus* (MRSA). MRSA is a major pathogen causing both hospital- and community-acquired infections that, like many other hospital acquired infections, is resistant to at least one of the antibiotics most commonly used for treatment [24,25,26,27,28]. A recent meta-analysis shows that MRSA causes more fatalities in the US than HIV, underscoring the urgent need for new treatments [29].

2.3 Results

2.3.1 Drug-metabolite links recapitulate known drug-target interactions

To evaluate such the chemical similarity between drug classes and metabolic reactions, links between sets of metabolic ligands and sets of drugs were generated according to SEA (**Figure 1**) [6]. The similarity metric consists of a descriptor, represented by standard two-dimensional topological fingerprints, and a similarity criterion, the Tanimoto coefficient (Tc).

Expectation values (E) were calculated for each set pair by comparing the raw scores to a background distribution generated using sets of randomly selected molecules (see **Methods** for further details). To represent metabolic ligand sets, the MetaCyc database, which includes enzymes from more than 900 different organisms catalyzing over 6,000 reactions, was used [12]. The substrates and products of each enzymatic reaction were combined to form a reaction set, each of which was required to contain at least two unique compounds. Ubiquitous molecules called common carriers, which frequently play critical roles in reaction chemistry but do not distinguish the function of a specific enzyme, were removed, leaving a total of 5,056 reactions involving 4,998 unique compounds. To represent drugs, a subset of 246 targets of the MDL Drug Data Report (MDDR) collection, which annotates ligands according to the targets they modulate, was used [30]. These sets contain 65,241 unique ligands with a median and mean of 124 and 289 ligands per target, respectively. Overall, 246 drug versus 5,056 reaction set comparisons involving 1.39×10^9 pairwise comparisons were made.

Although drugs and metabolites typically differ in their physiochemical properties, significant and specific similarity links nonetheless emerged. Using SEA at an expectation value cutoff of $E = 1.0 \times 10^{-10}$, a previously reported cutoff for significance [6], 54% (132 of 246) of drug sets link to an average of 43.7 (median = 10) or 0.9% of metabolic reactions. None of the remaining 46% (114 of 246) of drug sets link to any metabolic reaction sets. For instance, while the α -glucosidase drug set links to the α -glucosidase reaction ($E = 1.00 \times 10^{-51}$), the thrombin inhibitor drug set does not link significantly with any metabolic reaction. The thrombin inhibitor drug set targets the serine protease thrombin, which does not participate in small molecule metabolism, but rather plays a role in the coagulation signaling cascade. Similarly, 67% (1,790

of 5,371) of reaction sets do not link to any MDDR drug set at the expectation value cutoff of $E = 1.0 \times 10^{-10}$, but those that do hit an average of 2.8 (median = 2) or 1.1% of drug sets. This is strikingly similar to the 0.9% of metabolic sets that an average drug set hits. For instance, the metabolite set for the valine decarboxylase reaction, which is not an MDDR drug target, does not link significantly to any drug sets, but the retinal dehydrogenase reaction set links, as expected, to the retinoid drugs at $E = 3.05 \times 10^{-98}$. For full results, see **Supplemental Data**.

To determine the recovery rate of known drug-target interactions, it was hypothesized that chemical similarity between MetaCyc reaction sets and corresponding MDDR drug sets could specifically recover the known drug-target interactions. The 246 MDDR drug set targets include 62 enzymes that could be mapped to MetaCyc via the Enzyme Commission (EC) number [31] describing the overall reaction catalyzed [32]. The results show that all 62 reaction sets for these targets link to at least one MDDR drug set. The majority of best hits (42 out of 62) were found at expectation values of $E = 1.0 \times 10^{-10}$ or better (**Table 1**). At expectation values better than $E = 1.0 \times 10^{-25}$, 61% (19 of 31) of best hits recover either the specific known target or another enzyme in the same pathway. Examples of specific compounds linked by this analysis are given in **Table 2** for a selected group of these best-scoring hits.

Other links recovered off-pathway hits, which often reflect known polypharmacology that is well-documented. For example, the glycylamide ribonucleotide formyltransferase (GART) inhibitor drug set hits both the GART reaction set ($E = 1.55 \times 10^{-82}$) and the off-pathway but pharmacologically related antifolate target dihydrofolate reductase (DHFR) ($E = 1.02 \times 10^{-134}$). Other off-pathway hits reflect biological connections, or physical connections,

between targets. For example, the adenosine deaminase reaction set links to the A₁ adenosine receptor agonist drug set ($E = 7.69 \times 10^{-159}$) (**Table 1**) capturing the known interaction between A₁ adenosine receptors and adenosine deaminase on the cell surface of smooth muscle cells [33]. Considering only the stringent case of exact matches based on EC numbers, a Mann-Whitney rank-sum test (also referred to as the U-test) shows that the expectation values for links between reaction sets and drug sets of known drug target enzymes were significantly better than the expectation values for links to reaction sets of non-target enzymes, i.e., 62 known enzyme targets were recovered in a background of 4,920 non-target “other” enzymes at a statistical significance of $P = 2.01 \times 10^{-6}$.

The predicted links recapitulate many known drug-target interactions and suggest new hypotheses about drug-target interactions. One such new prediction involves the phospholipase A₂ (PLA₂) inhibitor drug class. The substrates and products of PLA₂ recapitulate its known link to the PLA₂ inhibitor drug set ($E = 9.82 \times 10^{-26}$), however, the sterol esterase reaction returns an even better score against the PLA₂ inhibitor set ($E = 3.18 \times 10^{-44}$) (**Table 1**). Although this predicted pharmacological relationship has, to our knowledge, not been previously documented, the result is consistent with the known biological relationship between PLA₂ and sterol esterase. Both enzymes are secreted by the pancreas and require phosphatidylcholine hydrolysis to facilitate intestinal cholesterol uptake [34]. Thus, this link suggests that therapeutic agents directed against PLA₂ may also inhibit sterol esterase, perhaps even more strongly than their intended target.

2.3.2 Human drug “effect-space” maps detail interactions between drug classes and enzyme targets

To present links between small molecule metabolites and drugs in the context of their known (and potential) metabolic targets, metabolic “effect-space” maps for currently marketed drugs were generated for each of the 246 drug classes investigated in this work. These maps enable visualization of the chemical similarities between drugs and metabolites painted onto human metabolic pathways, illustrating potential interactions between an individual drug class and specific metabolic enzymes in humans. Examples include the nucleoside reverse transcriptase, dihydrofolate reductase, and thymidylate synthase inhibitors which target pyrimidine nucleotide metabolism and biosynthesis of the essential coenzyme folate (**Figure 2 and Table 3**). Using the canonical human metabolic pathways from HumanCyc [35], a subset of the BioCyc [12] database collection, reactions in each metabolic network have been colored according to their similarity to known drug classes (**Figure 2**). While **Table 1** presents only the top link for each of 62 enzyme targets in MetaCyc against the 246 MDDR drug classes, the networks in **Figure 2** detail all significant hits for selected drug classes against the pyrimidine and folate pathways. Interactive versions of these maps, one for each of the 246 drug classes included in our analysis, are available online (see below).

It has previously been shown that chemical similarity between known drugs often suggests novel drug-target interactions [5,6,7,14]. Consistent with these observations, effect-space maps such as those shown in **Figure 2** can also be used to exploit chemical similarities between drugs and metabolites to indicate potential routes of drug metabolism and toxicity [3,11,36,37]. For example, the nucleoside reverse transcriptase inhibitors (NRTIs) used in HIV

therapy are administered as pro-drugs. The effect-space map reflects this route of NRTI metabolism leading to viral inhibition. The top three hits yielded by the NRTI drug set queried against human metabolism – thymidine kinase ($E = 3.48 \times 10^{-26}$), thymidylate kinase ($E = 7.48 \times 10^{-28}$), and deoxythymidine diphosphate kinase ($E = 1.54 \times 10^{-24}$) (reaction numbers 2, 3, and 4 in **Figure 2A**, additional results in **Table 3A**) – successively phosphorylate the NRTI pro-drugs into the pharmacologically active NRTI triphosphates [38,39]. The viral reverse transcriptase enzyme then incorporates the fully phosphorylated NRTIs into the growing DNA strand, thereby terminating transcription of the viral DNA. In this example, considerable toxicity mitigates the therapeutic value of inhibiting viral DNA transcription since the phosphorylated NRTIs directly inhibit human nucleotide kinases and mitochondrial DNA pol- γ . They also may be incorporated by pol- γ into the growing human mitochondrial DNA strand, and once incorporated are inefficiently excised by DNA pol- γ exonuclease [40]. Thus, the effect-space map illustrates both the route of metabolism and a mechanism of toxicity for NRTIs in humans.

Drug effect-space maps also offer a broad glimpse of potential human metabolic interactions predicting new “polypharmacology”. From the ligand perspective, “drug polypharmacology” refers to a single drug or drug class that hits multiple targets. For example, dihydrofolate reductase (DHFR, reaction number 7 in **Figure 2**) uses NADPH to reduce 7,8-dihydrofolate to tetrahydrofolate. Antifolate drugs inhibit DHFR, and, as expected, the DHFR drug set recovers the DHFR reaction substrates and products as the top similarity hit in human metabolism ($E = 1.46 \times 10^{-82}$) (**Figure 2B**, **Table 3A**, **Table 4**). However, at least 20 other reactions also use folate coenzymes in human metabolism [41,42,43]. Accordingly, SEA finds additional links between the DHFR drug set and established antifolate targets outside the

pyrimidine and folate biosynthesis pathways such as serine hydroxymethyltransferase (SHMT, $E = 2.68 \times 10^{-44}$), phosphoribosyl-aminoimidazole-carboxamide formyltransferase (AICAR transformylase, $E = 2.21 \times 10^{-39}$), and phosphoribosyl-glycinamide formyltransferase (GART, $E = 2.21 \times 10^{-39}$) (**Table 3A**). The effect-space maps in **Figure 2** illustrate the results from **Table 3A** and **Table 4** in a single view, illustrating drug polypharmacology with respect to critical metabolic pathways.

Alternatively, from the target perspective, “target polypharmacology” may refer to a single target being modulated by multiple classes of drugs. For instance, thymidylate synthase (TS) is another classic antifolate target that uses a folate coenzyme to methylate deoxyuridine phosphate, generating deoxythymidine phosphate [44,45,46,47]. As expected, the TS reaction links to known antifolate drug classes such as GART inhibitors ($E = 4.76 \times 10^{-73}$) and DHFR inhibitors ($E = 1.91 \times 10^{-48}$) (**Tables 3B and 4**). However, TS is also effectively inhibited by uracil analogs such as fluoropropynyl deoxyuridine, which is not a folate, but rather a pyrimidine analog. Accordingly, the TS reaction also links to reverse transcriptase inhibitors, which include fluoropropynyl deoxyuridine and additional pyrimidine analogs such as azidothymidine (AZT) ($E = 5.68 \times 10^{-11}$) (**Table 4**). The target polypharmacology of the thymidylate synthase enzyme is mirrored by the drug polypharmacology of the thymidylate synthase inhibitors. The TS inhibitors link not only to the reactions of deoxyribonucleotide biosynthesis including thymidylate synthase ($E = 2.54 \times 10^{-75}$), but also the GART ($E = 1.50 \times 10^{-60}$) and DHFR ($E = 1.96 \times 10^{-123}$) reactions (**Figure 2C and Table 3**). Thus, SEA recapitulates the known polypharmacology of TS. Effect-space maps illustrate and clarify these pharmacological relationships.

2.3.3 Species-specific effect-space maps for drug discovery in pathogens

The great diversity of metabolic strategies, pathways, and enzymes present in humans, model organisms, and pathogenic species constitutes not only a major challenge, but also an opportunity for drug discovery. To visualize the drug-metabolite links identified in this work in a drug discovery context, species-specific effect-space maps were created for each of 385 organisms from the BioCyc Database Collection. Target reactions existing in common and differentially between each of these species and humans are shown in these metabolic maps. As with the human effect-space maps, this set of species-specific effect-space maps is available in interactive form online (see below). A static version of the effect-space map for MRSA is shown in **Figure 3**. To illustrate how these maps may be used to integrate different types of drug and biological data in a metabolic context (**Figure 4**), we present a case study on MRSA.

2.3.4 Case study: MRSA

In the following retrospective case study, we demonstrate how drug-metabolism maps may be used to guide drug-target discovery. As described for **Figure 2**, each node in the MRSA network (**Figure 3**) represents one reaction set, the substrates and products of a single metabolic reaction. Edges connect the reactions according to canonical BioCyc MRSA pathways. Each reaction in the network has been colored according to the expectation value of the best link between the reaction set and any of the 246 MDDR drug sets. Lighter colored nodes have higher expectation values indicating less drug-like reaction sets, while darker colored nodes indicate more drug-like reaction sets. To provide therapeutic context, reactions that are also present in human metabolism have been faded, indicating that drug sets targeting these enzymes in MRSA

may have the undesirable potential to inhibit the human enzymes as well. As with the other organisms represented in our online map set, most reactions in the MRSA subset have little chemical similarity to any MDDR drug set. Although 74% of the 469 MRSA metabolic reactions have measurable similarity to at least one MDDR drug set, only 36% of these links had expectation values of $E = 1.0 \times 10^{-10}$ or better. Several complete pathways of diverse chemical classes, including shikimic acid, phospholipid, peptidoglycan, teichoic acid, and molybdenum cofactor biosynthesis, lack links to any drug set at all. Only 18 of the 469 MRSA metabolic reactions are already known to be drug targets in MDDR. Fourteen of these are represented in **Figure 3** (as diamonds), but all 18 of these also appear in humans. Human orthologs to the MRSA enzymes that catalyze these shared reactions would likely be vulnerable to the same inhibitors, putting drugs that target these reactions at risk for toxicity.

The approach described in this paper may be applied to predict both drug action and toxicity at specific enzyme targets. However, successful modulation of the target may not alone be sufficient to kill the pathogen due to redundant pathways for the formation of critical metabolites. Therefore, we used the metabolic maps to integrate essentiality and synthetic lethal data, and thereby provide a more nuanced picture of the potential for drug discovery targeting metabolism in MRSA. Using flux balance analysis of the metabolic network, the essential enzymes and metabolites that are required for the formation of all necessary biomass components of an organism can be identified [48,49]. Such an analysis has recently been performed by several of the authors for 13 strains of *Staphylococcus aureus*, including the methicillin-resistant N315 strain [50]. In short, the metabolic network was reconstructed from the

genome to include all reactions and the essentiality of a given enzyme was then assessed by the effect of the removal of that enzyme on biomass production in an ideally rich medium.

From the flux balance analysis, 39 predicted essential reactions could be mapped to our dataset (**Figure 4**). Several of these reactions have been successfully targeted by currently marketed drugs. For example, antifolate targets DHFR ($E = 1.02 \times 10^{-134}$), thymidylate synthase ($E = 2.54 \times 10^{-75}$), and dihydrofolate synthase ($E = 1.35 \times 10^{-70}$) all score strongly against MDDR drug sets. However, while species-specific antifolates do exist, many antifolates such as methotrexate used in cancer therapy cause severe toxicity [43]. To avoid such toxicity, 14 of the 39 essential MRSA reactions that are also present in humans were excluded from further consideration. All but two of the remaining 25 essential reactions were accounted for by four pathways: folate, tryptophan, histidine, and shikimate biosynthesis. These essential reactions include well-validated (but as yet unsuccessful) drug targets such as shikimate kinase in chorismate biosynthesis as well as several novel targets [51].

Only four remaining reaction sets hit against MDDR drug sets with expectation values better than 1.0×10^{-10} . Of these, MRSA's uroporphyrin-III methyltransferase reaction, in the cobalamin synthesis pathway, was the most similar of all its metabolic reactions to the MDDR drug sets. The reaction set and its matching drug sets (including top hits S-adenosyl-homocysteine hydrolase (SAMH) inhibitors ($E = 1.58 \times 10^{-204}$), adenosine (A2) agonists ($E = 8.11 \times 10^{-140}$), and adenosine (A1) agonists ($E = 7.70 \times 10^{-133}$)) all contained close analogs of the common carrier S-adenosyl-homocysteine (SAH). This common carrier was not filtered because it occurred in the network at a frequency below the threshold set in **Methods**. Despite the

ubiquity of this cofactor in essential human reactions and the concern for side-effects [52], SAH analogs have long been recognized as potent methylation inhibitors and potential antimicrobials [53]. Most other enzymes in the cobalamin pathway also use SAH as a cofactor and thus link prominently to MDDR (**Figure 4**) although they themselves are not predicted essential. The remaining three top hits among essential reactions, histidinol dehydrogenase ($E = 4.22 \times 10^{-16}$), tryptophan synthase ($E = 1.31 \times 10^{-19}$), and anthranilate phosphoribosyltransferase ($E = 4.27 \times 10^{-15}$) show more modest chemical similarity to MDDR drug sets. While these links may suggest potentially useful drug targets, the expectation values fall outside the more stringent empirical cutoff of 1.0×10^{-25} for exact or same pathway matches noted in **Table 1**.

Due to the lack of essential MRSA reactions with strong links to MDDR, synthetic lethal reactions that could be targeted in combination were investigated as new drug target alternatives. Again using predictions from flux balance analysis [50], 19 synthetic lethal enzyme pairs were mapped to MRSA (**Figure 4**). All but one of the synthetic lethal pairs, aroenate and prephenate dehydrogenase in tyrosine precursor biosynthesis, had at least one orthologous enzyme also present in humans. Of all the potential targets without human orthologs, whether essential or synthetic lethal, the aroenate dehydrogenase reaction set found the strongest links to MDDR drug sets ($E = 4.80 \times 10^{-28}$). Interestingly, both aroenate and prephenate dehydrogenase are catalyzed by the *tyrA* enzyme in the same active site, making *tyrA* essential for MRSA survival [54]. Furthermore, aroenate dehydrogenase was also predicted to be synthetic lethal with aspartate aminotransferase (AAT). Based upon the SEA results, combined with the MRSA synthetic lethal analysis, we found *tyrA* and AAT to be the targets most accessible to current drug chemistry. Consistent with this blind prediction, numerous inhibitors for both enzymes,

such as m-fluorotyrosine for tyrA and aminooxyacetate for AAT, have been previously reported in the literature [55,56,57,58,59]. Species-specific maps such as MRSA's (**Figure 3**) complement flux-balance analysis and other approaches, enabling systematic analysis of new strategies for attacking many pathogenic organisms.

A compilation of all of the metabolic network maps generated in this study is available at <http://sea.docking.org/metabolism>. These include interactive versions of the human effect-space maps shown in **Figure 2**, one for each of the 246 MDDR drug classes analyzed in this work, and 385 species-specific maps such as that shown in **Figure 3**. The species-specific maps were generated from the BioCyc database public collection, a compendium of 385 model organisms and pathogens whose genomes have been sequenced and their metabolomes deciphered. Of these, 65 have been designated as Priority Pathogens by the National Institute of Allergy and Infectious Diseases (NIAID) and include *Bacillus anthracis*, *Brucella melitensis*, *Cryptosporidium parvum*, *Salmonella*, SARS, *Toxoplasma gondii*, *Vibrio cholerae*, and *Yersinia pestis* [60]. Browse and similarity search tools are also provided, allowing exploration of the metabolic reaction sets and current drug classes used in this work, as well as comparison to user-defined custom ligand sets. These interactive tools enable facile exploration between the vast biological data on potential metabolic drug targets in these organisms and the drug chemistry currently available to prosecute those targets.

2.4 Discussion

A key product of this study is the construction of drug-metabolite correspondence maps that provide a more contextual picture of predicted drug action in human metabolism than has been previously available. Two aspects of these maps deserve particular emphasis. First, despite the differences in physiochemical properties of most drugs and small molecule metabolites, numerous links arise between drugs and metabolism. Viewed in the context of metabolic networks, the pharmacological relationships predicted by these links become biologically sensible. Moreover, retrospective analysis shows that these connections are biologically and pharmacologically significant. The links capture known polypharmacology and may predict potential targets that have been previously unrecognized. They reveal the relevant chemotypes previously explored in drug development and provide new tools to further interrogate the links between drugs and metabolism. Second, by integrating chemical and biological information with the metabolic context, our approach provides metabolome-wide exploratory tools to guide target identification and indicate routes of drug metabolism and toxicity.

With respect to the coverage of drug links across small molecule metabolism that this study provides, we note that the SEA method relies solely upon the chemical similarity of ligands to establish links between drug sets and reaction sets. Based on these links, and the biologically sensible connections shown in the results, we infer that a particular drug class may act on a certain target. However, drugs may also act against a target without resembling the endogenous substrate, for example, by allosteric regulation. The SEA method, as applied here to the substrates and products of metabolic reactions, does not capture these additional drug-target links. To estimate the frequency of such cases that fall beyond the scope of this study, the results reported in **Table 1** provide some answers. Of the 62 known enzyme targets in MetaCyc, 42

(68%) the substrate/product metabolite sets show significant chemical similarity to at least one MDDR drug set, establishing a reasonable first pass estimate for the percentage of current enzyme targets accessible to the approach presented here.

2.5 Conclusion

Using the SEA method, we have shown that comparison between ligand sets representing MDDR drug classes and ligand sets representing the substrates and products of metabolic reactions yields links between known drugs and enzyme targets at high statistical significance. Because the method is based on chemical similarity and requires only information from these molecule sets rather than the sequence, structure or physiochemistry of the targets, this ligand-based approach is independent from, and complementary to, protein structure and sequence based methods. The results also suggest the potential of this method for predicting previously unknown interactions between drug classes and metabolic targets, recovering routes of metabolism and toxicity in humans, and identifying potential drug targets in emerging pathogens. Thus, by mapping the chemical diversity of drugs to small molecule metabolism using ligand topology, this work establishes a computational framework for ligand-based prediction of a drug class's action, metabolism, and toxicity.

2.6 Methods

2.6.1 Compound sets. All compounds, both drugs and metabolites, are represented using Daylight SMILES strings [61]. Sets comprised of isomers with unique compound names were

retained, even though stereochemistry was later removed as part of the molecule fingerprinting process.

2.6.2 Ligand sets. Reaction sets were extracted from the 8.15.2007 release of MetaCyc based upon the substrates and products annotated to each reaction. Two filters were applied. First, the ten most common metabolites based on the number of occurrences in the MetaCyc metabolic network were removed: water, ATP, ADP, NAD, pyrophosphate, NADH, carbon dioxide, AMP, glutamate, and pyruvate. Second, each reaction set was required to include at least two unique compounds, as indicated by a MetaCyc or a MDDR unique compound id.

2.6.3 Drug sets. Drug sets were extracted from the MDDR, a compilation of about 169,000 drug-like ligands in 688 activity classes, each targeting a specific enzyme (designated by the Enzyme Commission (E.C.) number). The subset of this database for which mappings between enzymes and the MDDR drug classes were available was used. These mappings were based on a previous study that maps E.C. numbers, GPCRs, ion channels and nuclear receptors to MDDR activity classes [32]. Only sets containing five or more ligands were used. Salts and fragments were removed, ligand protonation was normalized and duplicate molecules were removed. Of the 688 targets in the MDDR, 97 were excluded as having too few ligands (<5), and another 345 targets were excluded because their definitions did not describe a molecular target, e.g., drugs associated only with an annotation such as "Anticancer" were not used. The remaining 246 enzyme targets were together associated with a total of 65,241 unique ligands, with a median and mean of 124 and 289 drug ligands per target. For further details, see Keiser et al. [6].

2.6.4 Set comparisons. All pairs of ligands between any two sets were compared using a pair-wise similarity metric, which consists of a descriptor and a similarity criterion. For the similarity descriptor, standard two-dimensional topological fingerprints were computed using the Scitegic ECFP4 fingerprint [62]. The similarity criterion was the widely used Tanimoto coefficient (Tc) [63]. For set comparisons, all pair-wise Tcs between elements across sets were calculated, and those scoring above a threshold were summed, giving a raw score relating the two sets. The Tanimoto coefficient threshold of 0.32 was determined according to a previously published method based upon fit to an extreme value distribution [6]. A model for random similarity similar to that used by BLAST [64] was used to generate expectation (E) values which are used to describe the strengths of relationships discovered using this protocol [6]. All scores reported here are based upon the background distribution and cutoff scores generated using the drug sets extracted from the MDDR collection. For further details, see Keiser et al. [6]. Network visualization was performed in Cytoscape 2.6.2 [65] using the γ -files hierarchical layout algorithm.

2.6.5 MRSA essentiality and synthetic lethal analysis.

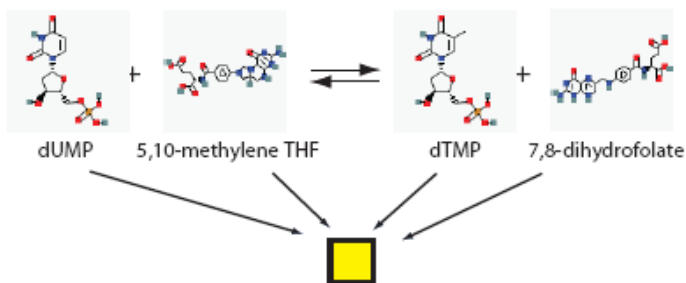
Essentiality and synthetic lethal data generated as described earlier [49]. Briefly, the metabolic network was reconstructed from the genome to include all reactions that have an active flux. The essentiality of a given enzyme was then assessed by the effect of the removal of that enzyme on biomass production. Similarly, synthetic lethal pairs can be identified by systematic pairwise deletion of enzymes and recalculation of biomass production in an ideally rich medium.

2.7 Acknowledgments

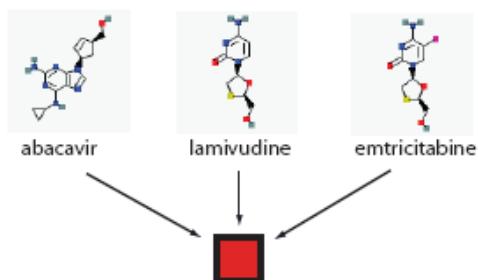
This work was supported by National Institutes of Health (NIH) GM71896 to BK Shoichet, NIH GM60595 to PCB, NIH U01-AI070499 to OW, NIH GA1070499-01 to AL Barabasi. MJK was supported by Training Grant GM67547 and a National Science Foundation graduate fellowship. We also thank Elsevier MDL for the MDDR and Scitegic for PipelinePilot.

A) Compose Sets

Example reaction set: *Thymidylate synthase*



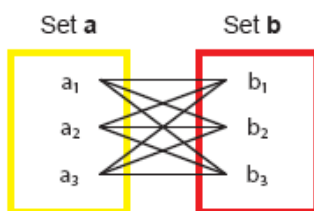
Example drug set: *Nucleoside reverse transcriptase inhibitors*



B) Generate ligand fingerprints



C) Calculate set similarity raw score



D) Compare raw score to background distribution to determine expectation value

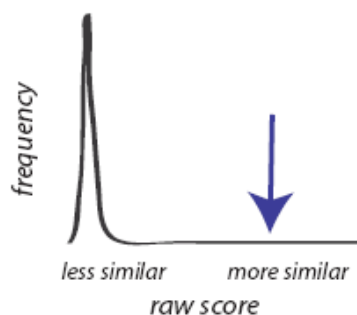


Figure 2.1| Similarity Ensemble Approach (SEA)

Figure 2.2 | Effect-space map showing chemical similarity between drugs and metabolites in human folate and pyrimidine biosynthesis. Each node represents one reaction set – the substrates and products of a single human metabolic reaction. Edges connect the reactions in the canonical pathway as annotated in HumanCyc [35]. As given in the color key, each reaction is colored according to the expectation value indicating the strength of similarity between that target reaction set and the respective MDDR drug sets represented in panels A-C. Diamond shaped nodes indicate reactions catalyzed by enzymes annotated as known drug targets in the MDDR; circles indicate reactions catalyzed by enzymes not annotated as targets. **Reaction key:** 1. Deoxyuridine kinase 2. Thymidine kinase 3. Thymidylate kinase 4. Deoxythymidine diphosphate kinase 5. Thymidylate synthase (TS) 6. Methylene tetrahydrofolate reductase 7. Dihydrofolate reductase (DHFR) 8. Deoxyuridine diphosphate kinase 9. Deoxyuridine triphosphate diphosphatase

Figure 2.2.A| Effect-space map – Nucleoside reverse transcriptase inhibitors (NRTIs)

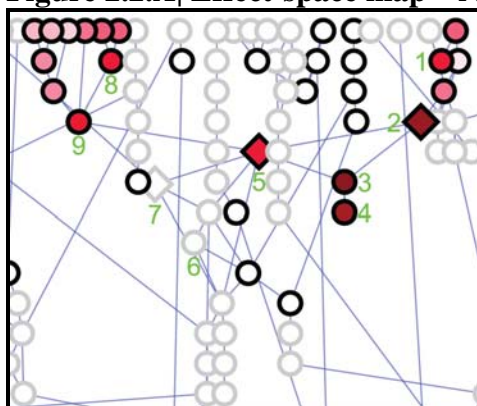


Figure 2.2.B| Effect-space map – Dihydrofolate reductase (DHFR) inhibitors

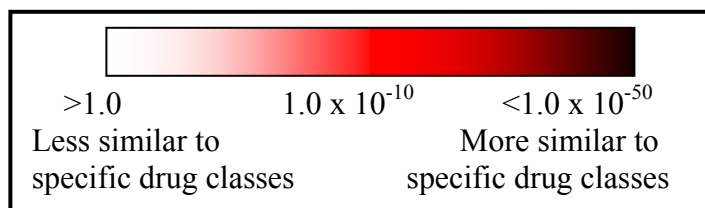
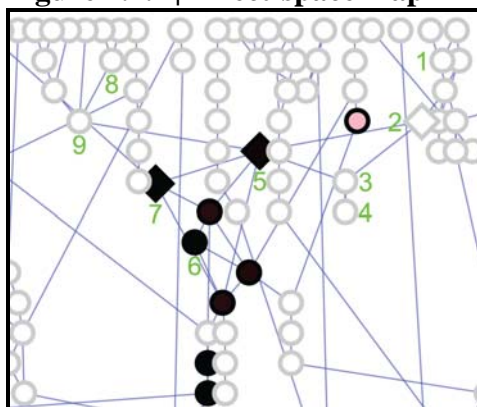
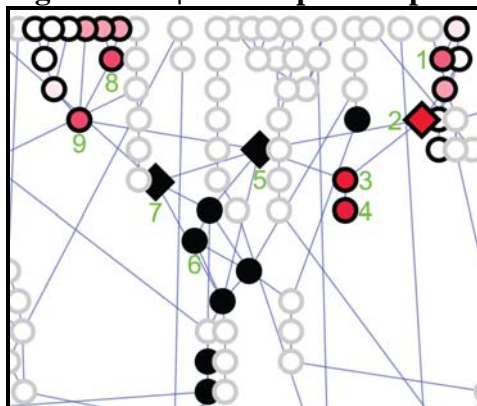


Figure 2.2.C| Effect-space map – Thymidylate synthase (TS) inhibitors



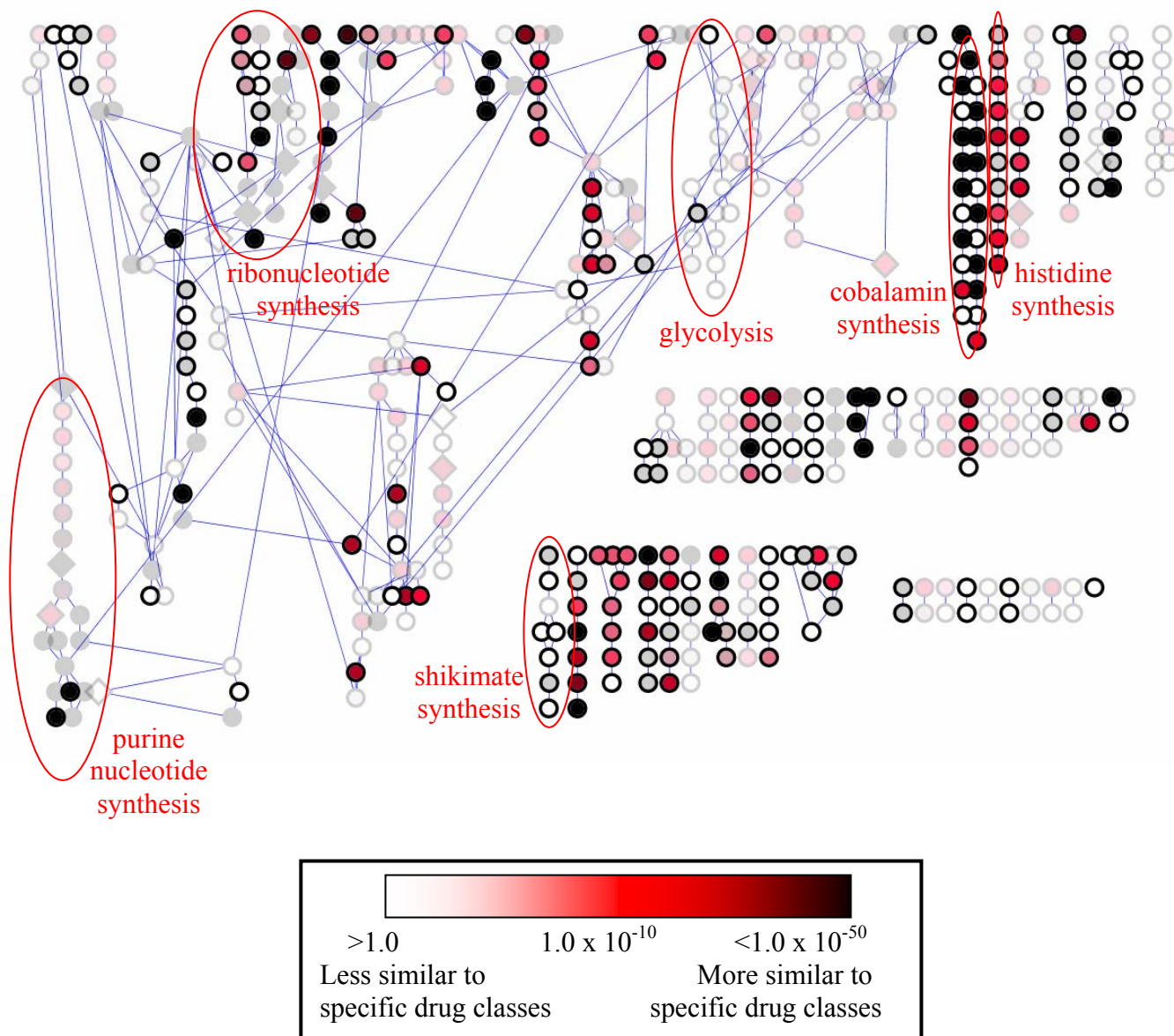
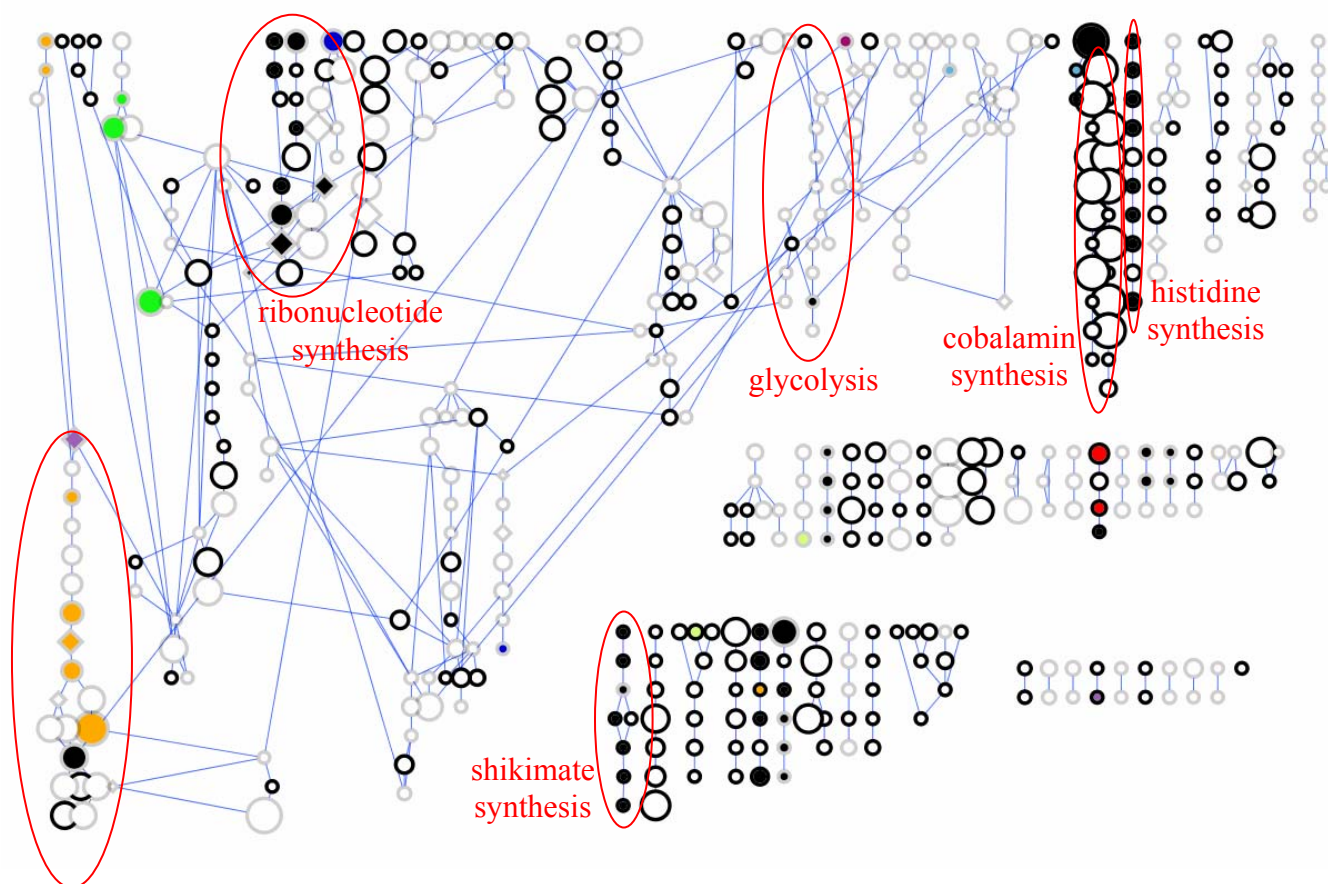


Figure 2.3| Effect-space map showing chemical similarity between drugs and metabolites in MRSA. Canonical pathway representation of methicillin-resistant *Staphylococcus aureus* (MRSA) [12] small molecule metabolism colored by expectation value of the best hit against MDDR. Reactions that are also present in humans have been faded. Layout based upon the Cytoscape 2.5 y-files hierarchical layout. Edge lengths are not significant. For ease of viewing, reactions are not labeled but can be identified in the interactive versions of the maps available at the online resource.



Key:
 Black = Essential reaction
 Other colors = Synthetic lethal reaction pairs
 Node size = similarity to top MDDR hit (bigger is more drug-like)
 Diamond shape = MDDR drug target
 Faded border = human reaction

Figure 2.4| Essential/synthetic lethal map of MRSA metabolism. Canonical pathway representation of methicillin-resistant *Staphylococcus aureus* (MRSA) small molecule metabolism colored by essentiality and synthetic lethality of reactions.

Table 2.1| Metabolic enzyme targets and their best links to MDDR.

| Enzyme Target ^a | EC# | Best Hit MDDR Drug Set | Best Hit E-value |
|---|-------------------|---|------------------|
| Adenosine kinase | 2.7.1.20 | S-Adenosyl-L-Homocysteine Hydrolase Inhibitor | 4.38E-288 |
| <i>Adenosylmethionine decarboxylase</i> | <i>4.1.1.50</i> | <i>S-Adenosyl-L-Homocysteine Hydrolase Inhibitor</i> | <i>2.71E-216</i> |
| Thromboxane-A synthase | 5.3.99.5 | Prostaglandin | 1.66E-204 |
| Adenosylhomocysteinase | 3.3.1.1 | S-Adenosyl-L-Homocysteine Hydrolase Inhibitor | 4.73E-203 |
| <i>Adenosine deaminase</i> | <i>3.5.4.4</i> | <i>Adenosine (A1) Agonist</i> | <i>7.69E-159</i> |
| Thymidine kinase | 2.7.1.21 | Thymidine Kinase Inhibitor | 3.19E-151 |
| <i>Dihydrofolate reductase</i> | <i>1.5.1.3</i> | <i>Glycinamide Ribonucleotide Formyltransferase Inhibitor</i> | <i>1.02E-134</i> |
| <i>Catechol O-methyltransferase</i> | <i>2.1.1.6</i> | <i>S-Adenosyl-L-Homocysteine Hydrolase Inhibitor</i> | <i>4.67E-127</i> |
| Prostaglandin-endoperoxide synthase | 1.14.99.1 | Prostaglandin | 8.57E-110 |
| <i>Purine-nucleoside phosphorylase</i> | <i>2.4.2.1</i> | <i>Adenosine (A1) Agonist</i> | <i>8.35E-105</i> |
| <i>Ribose-phosphate pyrophosphokinase</i> | <i>2.7.6.1</i> | <i>S-Adenosyl-L-Homocysteine Hydrolase Inhibitor</i> | <i>4.33E-91</i> |
| Phosphoribosylglycinamide formyltransferase | 2.1.2.2 | Glycinamide Ribonucleotide Formyltransferase Inhibitor | 1.55E-82 |
| Phosphoribosylaminoimidazolecarboxamide formyltransferase | 2.1.2.3 | Glycinamide Ribonucleotide Formyltransferase Inhibitor | 9.12E-80 |
| <i>3',5'-cyclic-nucleotide phosphodiesterase</i> | <i>3.1.4.17</i> | <i>S-Adenosyl-L-Homocysteine Hydrolase Inhibitor</i> | <i>1.23E-77</i> |
| Thymidylate synthase | 2.1.1.45 | Thymidylate Synthetase Inhibitor | 2.54E-75 |
| Steryl-sulfatase | 3.1.6.2 | Aromatase Inhibitor | 4.90E-62 |
| <i>Guanylate cyclase</i> | <i>4.6.1.2</i> | <i>Purine Nucleoside Phosphorylase Inhibitor</i> | <i>2.68E-60</i> |
| Cholestenone 5-alpha-reductase | 1.3.1.22 | Steroid (Salpha) Reductase Inhibitor | 3.63E-60 |
| Steroid 17-alpha-monooxygenase | 1.14.99.9 | Steroid (5alpha) Reductase Inhibitor | 1.37E-58 |
| <i>RNA-directed DNA polymerase</i> | <i>2.7.7.49</i> | <i>S-Adenosyl-L-Homocysteine Hydrolase Inhibitor</i> | <i>1.06E-52</i> |
| Alpha-glucosidase | 3.2.1.20 | Glucosidase (alpha) Inhibitor | 1.00E-51 |
| Farnesyl-diphosphate farnesyltransferase | 2.5.1.21 | Squalene Synthase Inhibitor | 2.12E-46 |
| Beta-galactosidase | 3.2.1.23 | Glucosidase (alpha) Inhibitor | 4.04E-46 |
| <i>Sterol esterase</i> | <i>3.1.1.13</i> | <i>Phospholipase A2 Inhibitor</i> | <i>3.18E-44</i> |
| Leukotriene-A4 hydrolase | 3.3.2.6 | Prostaglandin | 5.16E-40 |
| Squalene monooxygenase | 1.14.99.7 | Squalene Synthase Inhibitor | 7.59E-40 |
| <i>Ribonucleoside-diphosphate reductase</i> | <i>1.17.4.1</i> | <i>S-Adenosyl-L-Homocysteine Hydrolase Inhibitor</i> | <i>2.47E-38</i> |
| 3-hydroxyanthranilate 3,4-dioxygenase | 1.13.11.6 | 3-Hydroxyanthranilate Oxygenase Inhibitor | 1.14E-33 |
| Dihydroorotase | 3.5.2.3 | Dihydroorotase Inhibitor | 2.25E-32 |
| Nitric-oxide synthase | 1.14.13.39 | Nitric Oxide Synthase Inhibitor | 8.86E-28 |
| Phospholipase A2 | 3.1.1.4 | Phospholipase A2 Inhibitor | 9.82E-26 |
| <i>Diaminopimelate epimerase</i> | <i>5.1.1.7</i> | <i>Nitric Oxide Synthase Inhibitor</i> | <i>2.43E-24</i> |
| <i>Membrane dipeptidase</i> | <i>3.4.13.19</i> | <i>Nitric Oxide Synthase Inhibitor</i> | <i>2.81E-23</i> |
| 3-alpha(or 20-beta)-hydroxysteroid dehydrogenase | 1.1.1.53 | Aromatase Inhibitor | 1.51E-22 |
| <i>Sterol O-acyltransferase</i> | <i>2.3.1.26</i> | <i>Adenosine (A2) Agonist</i> | <i>4.95E-22</i> |
| <i>Hydroxymethylglutaryl-CoA reductase (NADPH)</i> | <i>1.1.1.34</i> | <i>Adenosine (A2) Agonist</i> | <i>4.95E-22</i> |
| <i>IMP dehydrogenase</i> | <i>1.1.1.205</i> | <i>Adenosine (A1) Agonist</i> | <i>8.98E-17</i> |
| <i>ATP-citrate (pro-S)-lyase</i> | <i>4.1.3.8</i> | <i>Adenosine (A2) Agonist</i> | <i>1.83E-15</i> |
| <i>Glutamate--cysteine ligase</i> | <i>6.3.2.2</i> | <i>Nitric Oxide Synthase Inhibitor</i> | <i>2.71E-11</i> |
| <i>Dopamine-beta-monooxygenase</i> | <i>1.14.17.1</i> | <i>Adrenergic (beta1) Agonist</i> | <i>3.81E-11</i> |
| Lanosterol synthase | 5.4.99.7 | Squalene Synthase Inhibitor | 1.38E-10 |
| <i>Nucleoside-diphosphate kinase</i> | <i>2.7.4.6</i> | <i>P2T Purinoreceptor Antagonist</i> | <i>2.76E-10</i> |

^aExact matches (the enzyme is the canonical target of the best MDDR hit) are shown in **bold type**, pathway matches (the enzyme shares the same pathway as the canonical target of the best MDDR hit) are shown in normal type, and enzymes not in the same pathway as the canonical target are shown in *italic type*.

Table 2.2| Selected best hits between MetaCyc reaction sets and MDDR drug sets

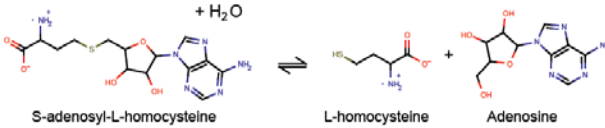

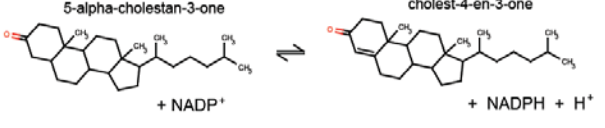
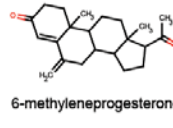
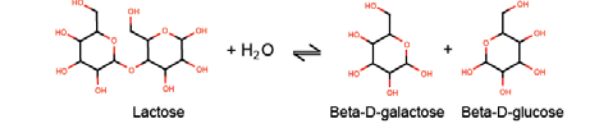

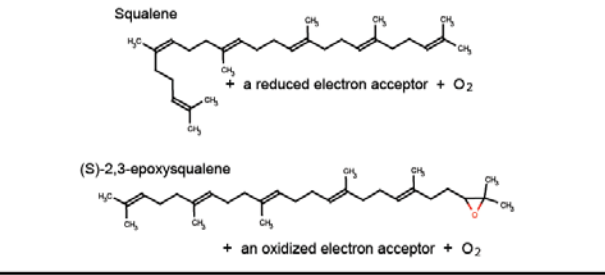
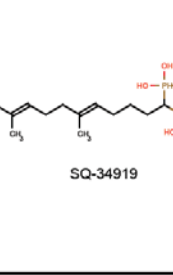
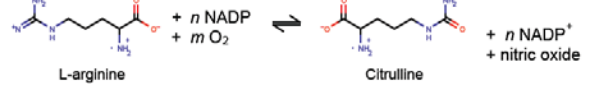
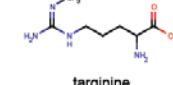
| Enzyme Target | Substrates and products | E-value of best hit | Representative Inhibitor |
|-------------------------------------|--|--|--|
| S-Adenosyl-L-Homocysteine Hydrolase |  <p>S-adenosyl-L-homocysteine + H₂O ⇌ L-homocysteine + Adenosine</p> | 4.73E-203 S-Adenosyl-L-Homocysteine Hydrolase Inhibitor |  <p>6-C-methyleneplanocin A (RMNPA)</p> |
| Cholestenone-5-alpha-reductase |  <p>5-alpha-cholestan-3-one + NADP⁺ ⇌ cholest-4-en-3-one + NADPH + H⁺</p> | 3.63E-60 Steroid-5-alpha reductase inhibitors |  <p>6-methyleneprogesterone</p> |
| Lactase |  <p>Lactose + H₂O ⇌ Beta-D-galactose + Beta-D-glucose</p> | 4.04E-46 Glucosidase-alpha inhibitor |  <p>Camiglibose</p> |
| Squalene monooxygenase |  <p>Squalene + a reduced electron acceptor + O₂ ⇌ (S)-2,3-epoxysqualene + an oxidized electron acceptor</p> | 7.59E-40 Squalene synthase inhibitor |  <p>SQ-34919</p> |
| Nitric oxide synthase (NOS) |  <p>L-arginine + n NADP + m O₂ ⇌ Citrulline + n NADP⁺ + nitric oxide</p> | 8.86E-28 Nitric oxide synthase inhibitors |  <p>targinine</p> |

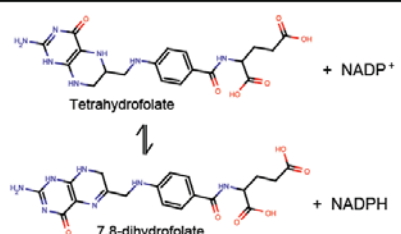
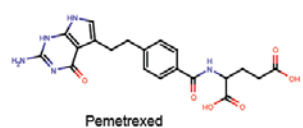
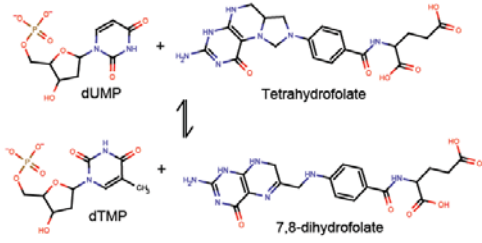
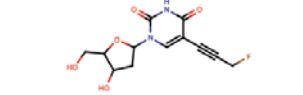
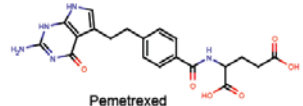
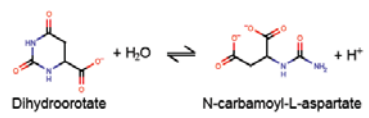
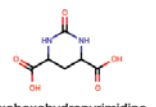
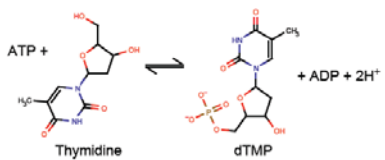
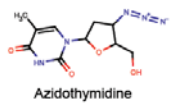
Table 2.3| Selected links between human metabolic reactions and current drugs.

| A. Select drug classes link to human metabolic reactions^a | | |
|---|--|----------------|
| Rank | Thymidylate Synthetase (TS) Inhibitor | E-value |
| 1 | Dihydrofolate reductase (DHFR) | 1.96E-123 |
| 2 | Methyltetrahydrofolate-corrinoid-iron-sulfur protein methyltransferase | 3.58E-102 |
| 3 | Methionyl-tRNA formyltransferase | 1.97E-99 |
| 4 | Methylenetetrahydrofolate reductase | 2.67E-86 |
| 5 | Thymidylate synthase (TS) | 2.54E-75 |
| 6 | Formate-tetrahydrofolate ligase | 1.44E-74 |
| 7 | Dihydrofolate synthetase | 1.35E-70 |
| 8 | Aminomethyltransferase | 7.13E-63 |
| 9 | 5-methyltetrahydrofolate-homocysteine S-methyltransferase | 2.80E-62 |
| 10 | Phosphoribosylaminoimidazolecarboxamide (AICAR) formyltransferase | 1.50E-60 |
| 11 | Phosphoribosylglycinamide formyltransferase (GART) | 1.50E-60 |
| Rank | Dihydrofolate Reductase (DHFR) Inhibitor | E-value |
| 1 | Dihydrofolate reductase (DHFR) | 1.46E-82 |
| 2 | Methyltetrahydrofolate-corrinoid-iron-sulfur protein methyltransferase | 2.84E-75 |
| 3 | Methylenetetrahydrofolate reductase | 6.01E-73 |
| 4 | Methionyl-tRNA formyltransferase | 7.00E-66 |
| 5 | Aminomethyltransferase | 6.90E-55 |
| 6 | Formate-tetrahydrofolate ligase | 6.15E-49 |
| 7 | Thymidylate synthase (TS) | 1.91E-48 |
| 8 | 5-methyltetrahydrofolate-homocysteine S-methyltransferase | 2.60E-45 |
| 9 | 3-methyl-2-oxobutanoate hydroxymethyltransferase | 2.68E-44 |
| 10 | Glycine decarboxylase | 2.68E-44 |
| 11 | Glycine hydroxymethyltransferase (SHMT) | 2.68E-44 |
| 12 | Dihydrofolate synthetase | 9.65E-42 |
| 13 | Phosphoribosylaminoimidazolecarboxamide (AICAR) formyltransferase | 2.21E-39 |
| 14 | Phosphoribosylglycinamide formyltransferase (GART) | 2.21E-39 |
| Rank | Nucleoside Reverse Transcriptase Inhibitor (NRTI) | E-value |
| 1 | Thymidylate kinase | 7.48E-28 |
| 2 | Thymidine kinase | 3.48E-26 |
| 3 | Deoxythymidine diphosphate kinase | 1.54E-24 |
| 4 | Ribonucleoside-triphosphate reductase | 2.88E-14 |
| 5 | Deoxyuridine triphosphate pyrophosphatase | 5.60E-12 |
| 6 | Deoxyuridine kinase | 1.14E-11 |
| 7 | Deoxyuridine diphosphate kinase | 1.45E-11 |
| 8 | Thymidylate synthase (TS) | 5.68E-11 |
| B. Select metabolic reactions link to current drug classes^b | | |
| Rank | Thymidylate Synthetase (TS) Reaction | E-value |
| 1 | Thymidylate synthase inhibitor (TS) | 2.54E-75 |
| 2 | Glycinamide ribonucleotide formyltransferase inhibitor (GART) | 4.76E-73 |
| 3 | Thymidine kinase inhibitor (TK) | 1.18E-62 |
| 4 | Dihydrofolate reductase inhibitor (DHFR) | 1.91E-48 |
| 5 | Folypolyglutamate synthetase inhibitor | 2.27E-31 |
| 6 | Nucleoside reverse transcriptase inhibitor (NRTI) | 5.68E-11 |
| Rank | Dihydrofolate Reductase (DHFR) Reaction | E-value |
| 1 | Glycinamide Ribonucleotide Formyltransferase Inhibitor | 1.02E-134 |
| 2 | Thymidylate Synthetase Inhibitor | 1.96E-123 |
| 3 | Dihydrofolate Reductase Inhibitor | 1.46E-82 |
| 4 | Folypolyglutamate Synthetase Inhibitor | 3.15E-62 |

^a Top ranked links to human metabolic reaction sets using select MDDR drug classes as query sets

^b Top ranked links to MDDR drug classes using selected human metabolic reactions as query sets

Table 2.4| Selected links between MDDR drug classes and human folate and pyrimidine metabolism

| Enzyme Target | Substrates and products | E-value of best hit | Representative Inhibitor |
|--------------------------------|--|---|---|
| Dihydrofolate reductase (DHFR) |  <p>Tetrahydrofolate + NADP⁺ → 7,8-dihydrofolate + NADPH</p> | 1.46E-82 DHFR inhibitors, best hit |  <p>Pemetrexed</p> |
| Thymidylate synthase (TS) |  <p>dUMP + Tetrahydrofolate → dTMP + 7,8-dihydrofolate</p> | 8.86E-75 TS inhibitors, best hit |  <p>Fluoropropynyl deoxyuridine</p>  <p>Pemetrexed</p> |
| Dihydroorotate |  <p>Dihydroorotate + H₂O → N-carbamoyl-L-aspartate + H⁺</p> | 2.25E-32 Dihydroorotate inhibitors, best hit |  <p>2-oxohexahydropyrimidine-4,6-dicarboxylic acid</p> |
| Thymidine kinase (TK) |  <p>ATP + Thymidine → dTMP + ADP + 2H⁺</p> | 3.48E-26 Reverse transcriptase inhibitors |  <p>Azidothymidine</p> |

2.8 References

1. Johnson M, Lajiness M, Maggiora G (1989) Molecular similarity: a basis for designing drug screening programs. *Prog Clin Biol Res* 291: 167-171.
2. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL (2007) Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov* 6: 29-40.
3. Kramer JA, Sagartz JE, Morris DL (2007) The application of discovery toxicology and pathology towards the design of safer pharmaceutical lead candidates. *Nat Rev Drug Discov* 6: 636-649.
4. Drews J (2006) Case histories, magic bullets and the state of drug discovery. *Nat Rev Drug Discov* 5: 635-640.
5. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. *Nat Biotechnol* 24: 805-815.
6. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, et al. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25: 197-206.
7. Cleves AE, Jain AN (2006) Robust ligand-based modeling of the biological targets of known drugs. *J Med Chem* 49: 2921-2938.
8. Watkins SM, German JB (2002) Metabolomics and biochemical profiling in drug discovery and development. *Curr Opin Mol Ther* 4: 224-228.
9. Shyur LF, Yang NS (2008) Metabolomics for phytomedicine research and drug development. *Curr Opin Chem Biol* 12: 66-71.
10. Rochfort S (2005) Metabolomics reviewed: a new "omics" platform technology for systems biology and implications for natural products research. *J Nat Prod* 68: 1813-1820.
11. Kell DB (2006) Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov Today* 11: 1085-1092.
12. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 36: D623-631.
13. Dobson CM (2004) Chemical space and biology. *Nature* 432: 824-828.
14. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M (2007) Drug-target network. *Nat Biotechnol* 25: 1119-1126.

15. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24: i232-240.
16. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, et al. (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 25: 71-75.
17. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. *Proc Natl Acad Sci U S A* 104: 8685-8690.
18. Hajduk PJ, Huth JR, Tse C (2005) Predicting protein druggability. *Drug Discov Today* 10: 1675-1682.
19. Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* 1: 727-730.
20. Imming P, Sinning C, Meyer A (2006) Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* 5: 821-834.
21. Meisner NC, Hintersteiner M, Uhl V, Weidemann T, Schmied M, et al. (2004) The chemical hunt for the identification of drugable targets. *Curr Opin Chem Biol* 8: 424-431.
22. Russ AP, Lampel S (2005) The druggable genome: an update. *Drug Discov Today* 10: 1607-1610.
23. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, et al. (2008) The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A* 105: 9880-9885.
24. Navarro MB, Huttner B, Harbarth S (2008) Methicillin-resistant *Staphylococcus aureus* control in the 21st century: beyond the acute care hospital. *Curr Opin Infect Dis* 21: 372-379.
25. Powell JP, Wenzel RP (2008) Antibiotic options for treating community-acquired MRSA. *Expert Rev Anti Infect Ther* 6: 299-307.
26. Clements A, Halton K, Graves N, Pettitt A, Morton A, et al. (2008) Overcrowding and understaffing in modern health-care systems: key determinants in methicillin-resistant *Staphylococcus aureus* transmission. *Lancet Infect Dis* 8: 427-434.
27. Avdic E, Cosgrove SE (2008) Management and control strategies for community-associated methicillin-resistant *Staphylococcus aureus*. *Expert Opin Pharmacother* 9: 1463-1479.

28. Nicasio AM, Kuti JL, Nicolau DP (2008) The current state of multidrug-resistant gram-negative bacilli in North America. *Pharmacotherapy* 28: 235-249.
29. Klevens RM, Morrison MA, Nadle J, Petit S, Gershman K, et al. (2007) Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States. *JAMA* 298: 1763-1771.
30. MDL Information Systems I (2006) MDL Drug Data Report. San Leandro, CA: MDL Information Systems, Inc.
31. Tipton KF (1992) *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB)*. New York: NC-IUBMB.
32. Schuffenhauer A, Zimmermann J, Stoop R, van der Vyver JJ, Lecchini S, et al. (2002) An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J Chem Inf Comput Sci* 42: 947-955.
33. Ciruela F, Saura C, Canela EI, Mallol J, Lluís C, et al. (1996) Adenosine deaminase affects ligand-induced signalling by interacting with cell surface adenosine receptors. *FEBS Lett* 380: 219-223.
34. Mackay K, Starr JR, Lawn RM, Ellsworth JL (1997) Phosphatidylcholine hydrolysis is required for pancreatic cholesterol esterase- and phospholipase A₂-facilitated cholesterol uptake into intestinal Caco-2 cells. *J Biol Chem* 272: 13380-13389.
35. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, et al. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* 6: R2.
36. Martin R, Rose D, Yu K, Barros S (2006) Toxicogenomics strategies for predicting drug toxicity. *Pharmacogenomics* 7: 1003-1016.
37. Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, et al. (2005) Computational prediction of human drug metabolism. *Expert Opin Drug Metab Toxicol* 1: 303-324.
38. Lewis W (2004) Cardiomyopathy, nucleoside reverse transcriptase inhibitors and mitochondria are linked through AIDS and its therapy. *Mitochondrion* 4: 141-152.
39. Petit F, Fromenty B, Owen A, Estaquier J (2005) Mitochondria are sensors for HIV drugs. *Trends Pharmacol Sci* 26: 258-264.
40. Lewis W, Kohler JJ, Hosseini SH, Haase CP, Copeland WC, et al. (2006) Antiretroviral nucleosides, deoxynucleotide carrier and mitochondrial DNA: evidence supporting the DNA pol gamma hypothesis. *Aids* 20: 675-684.

41. Kisliuk RL (2000) Synergistic interactions among antifolates. *Pharmacol Ther* 85: 183-190.
42. Faessel HM, Slocum HK, Rustum YM, Greco WR (1999) Folic acid-enhanced synergy for the combination of trimetrexate plus the glycinamide ribonucleotide formyltransferase inhibitor 4-[2-(2-amino-4-oxo-4,6,7,8-tetrahydro-3H-pyrimidino[5,4,6][1,4]thiazin -6-yl)-(S)-ethyl]-2,5-thienoylamino-L-glutamic acid (AG2034): comparison across sensitive and resistant human tumor cell lines. *Biochem Pharmacol* 57: 567-577.
43. Chan DC, Anderson AC (2006) Towards species-specific antifolates. *Curr Med Chem* 13: 377-398.
44. Costi MP, Ferrari S, Venturelli A, Calo S, Tondi D, et al. (2005) Thymidylate synthase structure, function and implication in drug discovery. *Curr Med Chem* 12: 2241-2258.
45. Gmeiner WH (2005) Novel chemical strategies for thymidylate synthase inhibition. *Curr Med Chem* 12: 191-202.
46. McGuire JJ (2003) Anticancer antifolates: current status and future directions. *Curr Pharm Des* 9: 2593-2613.
47. Chu E, Callender MA, Farrell MP, Schmitz JC (2003) Thymidylate synthase inhibitors as anticancer agents: from bench to bedside. *Cancer Chemother Pharmacol* 52 Suppl 1: S80-89.
48. Motter AE, Gulbahce N, Almaas E, Barabasi AL (2008) Predicting synthetic rescues in metabolic networks. *Mol Syst Biol* 4: 168.
49. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2: 886-897.
50. Lee DS, Burd H, Liu J, Almaas E, Wiest O, et al. (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel anti-microbial drug targets. *J Bacteriol*: accepted.
51. Kerbarh O, Bulloch EM, Payne RJ, Sahr T, Rebeille F, et al. (2005) Mechanistic and inhibition studies of chorismate-utilizing enzymes. *Biochem Soc Trans* 33: 763-766.
52. Chiang PK (1998) Biological effects of inhibitors of S-adenosylhomocysteine hydrolase. *Pharmacol Ther* 77: 115-134.

53. Pugh CS, Borchardt RT (1982) Effects of S-adenosylhomocysteine analogues on vaccinia viral messenger ribonucleic acid synthesis and methylation. *Biochemistry* 21: 1535-1541.
54. Ahmad S, Jensen RA (1987) The prephenate dehydrogenase component of the bifunctional T-protein in enteric bacteria can utilize L-arogenate. *FEBS Lett* 216: 133-139.
55. Rando RR, Relyea N, Cheng L (1976) Mechanism of the irreversible inhibition of aspartate aminotransferase by the bacterial toxin L-2-amino-4-methoxy-trans-3-butenoic acid. *J Biol Chem* 251: 3306-3312.
56. Liu D, Pozharski E, Lepore BW, Fu M, Silverman RB, et al. (2007) Inactivation of *Escherichia coli* L-aspartate aminotransferase by (S)-4-amino-4,5-dihydro-2-thiophenecarboxylic acid reveals "a tale of two mechanisms". *Biochemistry* 46: 10517-10527.
57. Cornell NW, Zuurendonk PF, Kerich MJ, Straight CB (1984) Selective inhibition of alanine aminotransferase and aspartate aminotransferase in rat hepatocytes. *Biochem J* 220: 707-716.
58. Rubin JL, Gaines CG, Jensen RA (1982) Enzymological Basis for Herbicidal Action of Glyphosate. *Plant Physiol* 70: 833-839.
59. Rubin JL, Gaines CG, Jensen RA (1984) Glyphosate Inhibition of 5-Enolpyruvylshikimate 3-Phosphate Synthase from Suspension-Cultured Cells of *Nicotiana glauca*. *Plant Physiol* 75: 839-845.
60. Zhang C, Crasta O, Cammer S, Will R, Kenyon R, et al. (2008) An emerging cyberinfrastructure for biodefense pathogen and pathogen-host data. *Nucleic Acids Res* 36: D884-891.
61. James C, Weininger D, Delaney J (1992-2005) *Daylight Theory Manual*. Mission Viejo, CA: Daylight Chemical Information Systems Inc.
62. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, et al. (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org Biomol Chem* 2: 3256-3266.
63. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* 11: 1046-1053.
64. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Jour Mol Biol* 215: 403-410.

65. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498-2504.

Chapter 3. Global Predictors for the Evolutionary Rates of Enzymes

3.1 Abstract

Numerous variables have been invoked as predictors for the evolutionary rates of proteins. While expression level is the most prominent, the independent contribution of each variable is inherently difficult to discern due to shared mechanisms and overlapping statistical effects. For enzymes, we propose that three key selectable properties represent important, and separable, axes of evolutionary constraint. First, we investigate gene expression, which requires proper folding and thermodynamic stability to produce functional proteins. Second, homologs with similar structure but dissimilar molecular function represent a protein's inherent structural constraints or evolvability. Finally, the output of metabolic systems, the chemical conversion of a substrate through multiple reactions into a biologically useful product, constitutes a key aspect of organismal fitness. In *Saccharomyces* small molecule metabolism, expression level is the single strongest predictor of evolutionary rate as measured by mRNA level ($r_{dn} = -0.49$, $r_{\omega} = -0.25$, $n = 1745$ genes) and codon adaptation index ($r_{dn} = -0.74$, $r_{\omega} = -0.46$, $n = 145$ genes). Coexpression, in contrast to expression level, does not correlate strongly with evolutionary rate ($r_{dn} = 0.09$, $r_{\omega} = 0.07$, $n = 98$ genes). We demonstrate a correlation in evolutionary rate (d_n) and selective pressure (ω) between structural superfamily members ($r_{dn} = 0.25$, $r_{\omega} = 0.24$, $n = 264$ genes) and metabolic network neighbors ($r_{dn} = 0.29$, $r_{\omega} =$

0.29, $n = 143$ genes), respectively. Taken together, metabolic network neighbors and superfamily members correlate significantly better in measures of evolutionary rate and selective pressure than either constraint alone ($r_{dn} = 0.56$, $r_w = 0.53$, $n = 100$ genes). These predictors are measurable in genomic data, can be relatively weighed, and contribute information independent from expression level. Together with expression level and dispensability, these predictors explain the majority of variance in d_n across our dataset ($r_{dn}^2 = 0.65$, $n = 64$).

3.2 Background

3.2.1 Introduction

Many hypotheses have been offered regarding the correlation between expression level and evolutionary rate, but the precise mechanisms remain elusive [1-3]. It has been suggested that highly expressed proteins are involved in more critical biological functions, and therefore are subject to stronger purifying selection. Another view recognizes that proper folding and thermodynamic stability are pre-requisites for proper protein function. The severity of deleterious mutations and the burden of unfolded, misfolded, or unstable proteins increases with expression level. Therefore, biological systems should be less tolerant of change in highly expressed proteins. A third hypothesis, translational selection, supposes that differences between amino acids in codon usage bias, based upon the speed, accuracy, and metabolic cost of translation, could constrain evolutionary rates. By ‘constrain’ we mean that some amino acid changes improve protein function and are selected for, others are neutral and subject to drift, while

many disrupt function and are selected against. Finally, the effects on evolutionary rate could also be caused, or at least enhanced, by secondary correlations with other variables.

Among these additional properties that have been correlated with evolutionary rate are dispensability, functional class, expression breadth, developmental timing, and degree of connectivity in protein-protein interaction networks [2-8]. The requirements of selection at various levels of function in biological systems differ tremendously from site-specific protein structural constraints in proteins or protein complexes [9-12], to translational efficiency [1, 2, 13], to substrate flow in metabolic pathways [14-16]. Overlapping effects of these properties complicate the analysis of evolutionary rates. For instance, genome regions with elevated mutation rates also show increased recombination rates and higher expression levels, while high recombination rates and expression correlate with reduced fixation rates for harmful mutations [2].

For enzymes, we propose that in addition to gene expression two key selectable properties of metabolic systems – structural similarity in homologs of dissimilar function and the output of metabolic pathways – represent important axes of constraint that uniquely distinguish metabolic pathways from other biological systems. First, investigation of structural homologs of dissimilar function allows dissection of the structural constraints on enzyme evolution from overlapping constraints on their functional roles in metabolic pathways. Second, the chemical conversion of a substrate, via multiple reactions within a metabolic pathway, into a biologically useful product provides the primary readout of metabolic fitness. To provide an estimate of the overall

importance of structural homologs and metabolic pathway context as predictors of evolutionary rates, we have compared the contribution of these two properties to those of both expression level and coexpressed groups of genes. Together these constitute a third axis of gene expression. Despite the importance of all these axes of constraint for understanding the evolutionary rates of enzymes, no analyses have yet addressed the relative extent to which they make independent contributions to evolutionary rate and selective pressure in the evolution of metabolic enzymes. Dissecting the individual contribution of each of these determinants will provide additional insight into the detailed mechanisms by which biological systems evolve.

To achieve this goal, we have three primary aims. First, to better understand the sources of variation in evolutionary rates, we aim to determine to what extent expression level accounts for variation in the evolutionary rates of enzymes, and whether the addition of orthogonal information improves these correlations. We also investigate correlations in evolutionary rate between coexpressed genes in the same transcriptional module. Given the challenges of obtaining protein and mRNA expression profiles for each new model organism studied, we are particularly interested in predictors of a gene's evolutionary rate encoded directly in the genome. We therefore included codon adaptation index in our analyses, along with mRNA level and protein abundance, as a proxy measure of expression level.

Second, we wish to determine whether homologs with similar structure but dissimilar function correlate in evolutionary rate. Therefore, we investigate correlations

in evolutionary rate among superfamily members as defined by the Structural Classification of Proteins (SCOP) [17]. SCOP superfamilies by definition descend from a common ancestor and share similar protein structures, but often encompass highly divergent sets of enzymes that can catalyze distinct and even widely different chemical reactions. These diverse superfamilies may be related through a similar aspect of function, such as binding of the same substrate or substrate sub-structure, or performing the same partial chemical reaction, both of which can be associated with conserved structural features [11]. Superfamilies frequently span sequences with less than 30% identity. Given strong structural similarity coupled with such great sequence divergence, we ask whether superfamily members correlate in evolutionary rate. By focusing upon evolutionarily related structures in the same superfamily, but annotated to distinct metabolic reactions, we aim to capture constraints on evolvability inherent in the structure and not based upon a single enzymatic function.

Third, we investigate whether metabolic network context is correlated with evolutionary rate. Evolutionary rate correlations have frequently been reported for protein-protein interaction networks [18, 19]. However, small molecule metabolic networks are distinct in that the majority of enzymes catalyzing the reactions of central metabolism are not known to directly bind or complex with each other [20]. Therefore, unlike protein-protein interaction networks, any correlations in evolutionary rate are likely due to causes other than compensating mutations at binding interfaces. Rather, these correlations may relate to an organism's need to maintain an appropriate relationship in timing and concentration of small molecule substrate flow through

metabolic pathways. In other work, metabolic network context has been addressed in terms of flux analysis, which also presumes the primary importance to fitness of an appropriate level of substrate flow through metabolic pathways [8].

3.2.2 A Pairs-based Approach

In this work, we focus on evolutionary rate correlations among three groups of small molecule metabolic enzymes – coexpressed enzymes in the same transcriptional module, enzymes in the same structural superfamily but with different overall reactions, and enzymes adjacent in the metabolic network. We conducted our study using the yeast *Saccharomyces cerevisiae* and its three close relatives *S. paradoxus*, *S. mikatae*, and *S. bayanus*. Together, these genomes provide excellent resolution of molecular evolution across a phylogeny of closely related eukaryotic organisms. Advantages of using the *Saccharomyces* yeast model system include detailed expression profiles [21, 22], well-characterized biochemical pathways from the *Saccharomyces* Genome Database (SGD [23]), and widely available protein structure data (SCOP).

To dissect the role of protein structure and metabolic network context from expression level in small molecule metabolic enzyme evolution, we compare pairs of enzymes with respect to the three axes of constraint investigated in this paper, i.e. pairs of structurally related enzymes, pairs of enzymes adjacent in the metabolic network, and pairs of enzymes expressed together. Metabolic network context constantly varies as the metabolic network neighbors differ for every enzyme. The same is true for superfamily membership and transcriptional modules. Pairs-based analyses account for this constantly

changing perspective. Metabolic pathways, superfamilies, and transcriptional modules also vary widely in size. Reliable ortholog sets and evolutionary rate information are not available for every enzyme. Our pairs-based approach was designed to manage these variations.

Using *S. cerevisiae* as our model system, we first generate three sets of pairs representing constraints of interest: all possible enzyme pairs in the same structural superfamily excluding identical pairs, all enzyme pairs that catalyze adjacent reactions in the metabolic network, and all possible pairs in the same transcriptional module excluding identical pairs (see **Supplemental Data**). We then map evolutionary rates to these *S. cerevisiae* enzyme pairs. We use previously published rates from orthologs that can be clearly identified across the four closely related *Saccharomyces* yeast species: evolutionary rate (d_n), synonymous mutation rate (d_s), and selective pressure ($d_n/d_s = \omega$). These data have previously been corrected for the well-known codon usage bias in *S. cerevisiae* [24, 25]. When evolutionary rate information for an enzyme is not available across all four yeast species, all pairs that contain that enzyme are excluded from analysis. The correlation between pairs can then be calculated across the three sets of pairs. Significance estimates are obtained by comparison to a label-permuted null model. For further details, see relevant sections below and **Methods**.

3.3 Results

3.3.1 Expression Level Accounts for Largest Percentage of Evolutionary Rate Variance

We begin by asking how much of the variation in evolutionary rate across our dataset is accounted for by expression level, and how much may be due to other factors. Following the method of Drummond et al., we applied principal component analysis (PCA) to our small molecule metabolic enzyme dataset. Using the Drummond et al. variables as a starting point, we performed PCA with the following seven variables: codon adaptation index (CAI), mRNA level, protein abundance, dispensability, superfamily average d_n , metabolic network neighbor average d_n , and (in analogy to protein-protein interaction networks) degree of connectivity in the metabolic network ($n = 64$). A strength of PCA is that it does not assume independence among variables. Therefore, it is frequently applied to reduce the dimensionality of datasets and find the subset of variables that account for the greatest variance.

Two strongly significant components emerge from PCA of the evolutionary rates (components 1 and 2 from **Fig. 1**). Component 1 totals 33% of the overall variance and is dominated by measures related to expression level including codon adaptation index, mRNA level, and protein level. Component 2 is the next largest comprising 30% of the total variance. This component contains measures related to network context (degree of connectivity in the metabolic network and average d_n of enzymes adjacent in the network), homologous structures (average d_n of an enzyme's SCOP superfamily members excluding the enzyme itself), and dispensability (a measure of gene essentiality

from knockout studies). The emergence of dispensability and network connectivity in consistent with previous reports ([8]

As shown in **Table 1**, expression related measures explained the greatest percentage of the overall variance in d_n , 29%. Superfamily members, enzymes adjacent in the metabolic network, and degree of connectivity in the metabolic network together explain an additional 24% of the overall variance in d_n . The opposite was true for selective pressure (ω) with superfamily members, adjacent enzymes, and connectivity collectively explaining 34% of the total variance, and expression level just 19%. Dispensability was the largest single variable contributing to variance in ω at 18%. All seven variables used in the PCA together accounted for 65% of the total variance in d_n (multiple $r_{dn}^2 = 0.6497$, adjusted $r_{dn}^2 = 0.6059$, $P = 9.073e-11$) and 70% of the total variance in ω (multiple $r_{dn}^2 = 0.704$, adjusted $r_{dn}^2 = 0.667$, $P = 9.862e-13$). These results confirm previously reported correlations between d_n and expression level [2, 3, 5, 26], as well as d_n and dispensability, but still leave 35% of the overall variance in evolutionary rate unaccounted for. We now focus on the contributions of protein structure and metabolic network context, which have not previously been addressed in an integrated analysis, to variance in evolutionary rate.

3.3.2 Evolutionary Rates Correlate among SCOP Superfamily Members

We examined whether structurally related enzymes, i.e. members of the same SCOP superfamily, significantly share measures of evolutionary rate and selective pressure and contribute to the overall variance in evolutionary rate. As defined by the Structural Classification of Proteins Database, superfamilies contain proteins with significant sequence divergence, but whose structures suggest that a common evolutionary origin is probable [17]. Despite clear structural similarity, sequence identity between superfamily members generally falls below 30%, and these homologs often perform distinct chemical reactions.

Our dataset contains 96 SCOP superfamilies for which evolutionary rate data are available for at least two members. Our results show that SCOP superfamily members correlate significantly in evolutionary rate and selective pressure ($r_{dn} = 0.25$, $P < 0.0001$; $r_{\omega} = 0.24$, $P < 0.0001$; $n = 264$ genes, 897 pairs, **Fig. 2**). Because isozymes could be a confounding factor in our analysis, we investigated their impact on our results. Enzyme annotated to both the same SCOP superfamily and to the same reaction comprised just 6% of the pairs in this analysis. Removing these pairs yields similar results ($r_{dn} = 0.20$, $P < 0.0001$; $r_{\omega} = 0.21$, $P < 0.0001$; $n = 247$ genes, 839 pairs).

3.3.3 Evolutionary Rates Correlate in Adjacent Metabolic Network Enzymes

We also examined whether evolutionary metrics for neighboring genes in the yeast small molecule metabolic network correlate with measures of evolutionary rate. To perform this analysis, the metabolic network was reconstructed based upon the 142

individual biochemical pathways that are currently represented in the SGD. This sample set contains all gene pairs with unambiguously identified orthologs in all four *Saccharomyces* yeast species, 42% and 25% of the annotated genes and adjacent gene pairs, respectively, of the SGD biochemical pathways. For this study, each gene pair in the network is unique and counted only once, no matter how many SGD pathways are annotated with the pair. The results show that in adjacent network pairs both evolutionary rate and selective pressure correlate with strong statistical significance ($r_{dn} = 0.29$, $P = 0.0008$; $r_{\omega} = 0.29$, $P = .0010$; $n = 145$ genes, 130 pairs; **Fig. 2**). As expected, and consistent with previous reports,[27] synonymous mutation rate shows poor correlation and no statistical significance ($r_{ds} = 0.06$, $P = 0.2192$).

Given that metabolic network context appears as a significant constraint in our results, the question arises whether these constraints on pathway pairs also constrain pathway order. To determine whether pathway order correlates with evolutionary rate and selective pressure, we repeated the analysis using only the set of individual SGD biochemical pathways, not the reconstructed network. We employed two different null models: one permuting globally across all pathways and a second permuting only locally within pathways. While pathway pairs correlate with high significance when permuting globally ($r_{dn} = 0.28$, $P = 0.0037$; $r_{\omega} = 0.29$, $P = .0040$; $n = 136$ genes, 125 pairs), they do not when permuting locally ($r_{dn} = 0.28$, $P = 0.0744$; $r_{\omega} = 0.29$, $P = .0893$), indicating that all pathway members tend to evolve together without regard to pathway order.

To account for direction of substrate flow, previously suggested to correlate with

selective pressure and evolutionary rate [28, 29], we also looked at all unbranched, unidirectional pathways. On average, upstream adjacent genes had lower measures of evolutionary rate ($d_n = 0.082$ vs. 0.097) and selective pressure ($\omega = 0.038$ vs. 0.047). We note, however, that the sample size is small ($n = 30$ pairs) and the result is not statistically significant (paired student's t-test: $t_{dn} = -1.37$, d.f. = 29 , $P = 0.27$). A more definitive answer to the question cannot be resolved without a larger dataset than was available for this work.

We also confirmed that these results would not be biased by protein-protein interaction. Of the few adjacent pathway pairs in the reconstructed metabolic network known to bind each other directly, none were sufficiently conserved across the four yeast species to also appear in our results [20]. Like protein-protein interaction networks, there is conflicting evidence whether degree of connectivity in the metabolic network correlates with either evolutionary rate or selective pressure. While we did not find such a correlation in a direct pairwise analysis ($r_{dn} = -0.07$, $P = 3.3e10$; $r_{\omega} = -0.07$, $P = 3.2e-1$; $n = 165$ genes), correlations have been reported in the literature [8]. Our pairwise correlation result contrasts with the PCA where degree of connectivity accounted for a small portion of variance (6.4%) in d_n . We note that our sample sets comprise different numbers of genes, and employ different approaches, pairwise correlation versus the global maximization approach of PCA. Taken together, these observations suggest that direct molecular interaction between enzymes, at least in this small dataset, cannot be a dominant mechanism of coevolution in metabolic networks. By coevolution we mean that proteins evolve at similar rates through compensating mutations. Thus, to the extent that

direct binding between protein pairs is a mechanism of coevolution in protein-protein interaction networks, metabolic pathways are distinct in that connectivity is carried instead by the flow of small molecule substrates.

3.3.4 Expression Levels Correlate in Adjacent Metabolic Network Pairs – But Not Transcriptional Module Pairs

Many reports have indicated that expression level correlates strongly with evolutionary rate [4, 30-32]. The same is true in our dataset (CAI $r_{dn} = -0.74$, $P = 2e-32$; $r_{\omega} = -0.46$, $P = 1e-10$; $n = 145$ genes). However, in the present work we are explicitly concerned with predictive correlations between different genes, not just different properties within the same gene. Therefore, we also investigated the correlations between transcription level [33], CAI [25], and protein abundance [34] in adjacent network enzymes (**Table 2**). As suggested by the correlations in evolutionary rate between network neighbors coupled with the correlation between CAI and evolutionary rate, we find that CAI, mRNA levels, and protein levels all significantly correlate in adjacent network pairs.

To examine the effect of transcriptional coregulation on evolutionary rate and selective pressure in the context of this study, we tested transcriptional modules as defined by Ihmels et al. [22] Transcriptional modules are self-consistent regulatory units comprising a set of coregulated genes and the experimental conditions that induce their coregulation. We found statistically significant but negligible correlations in d_n and ω

between transcriptional module pairs ($r_{dn} = 0.09$, $P = 3.0e-3$; $r_{ds} = 0.01$, $P = 6.7e-1$; $r_{\omega} = 0.07$, $P = 2.4e-2$; $n = 98$ genes; 1,137 pairs; **Fig. 2**), but these results cannot be considered reliable over varying confidence values (see Methods). Moreover, transcriptional module membership did not improve predictions of evolutionary rate or selective pressure when combined with either metabolic network context or SCOP superfamily. These results were unexpected given that 83% and 70% of the transcriptional module genes also appear in the network neighbor and SCOP superfamily gene sets, respectively. We rationalize these results by noting that although many enzymes grouped within pathways are coregulated, only 2% of the 1,137 transcriptional module pairs analyzed are adjacent in the metabolic network. Consistent with this interpretation, repeating the analysis requiring that coregulated enzyme pairs be within five metabolic network steps yields results similar to those using only the SGD adjacent network pairs (d_n : $r = 0.27$, $P = 0.0337$; ω : $r = 0.37$, $P = .0091$; $n = 76$ genes, 213 pairs).

3.3.5 The Independence of Network Context and SCOP Superfamily

To evaluate the combined contributions of network context and structural similarity to enzyme evolution, we looked at all genes ($n = 100$) that had both neighboring enzymes in the metabolic network and additional members in the same SCOP superfamily that are not adjacent in the metabolic network. Given the sequence metrics d_n , d_s , and ω , for each gene, we defined three quantities: s is the average of the SCOP superfamily members excluding the gene in question, p is the average of all enzymes adjacent in the network, and c is the average of s and p . We find that the combined metric outperforms all others, suggesting that the information in p and s is

additive. (**Fig 3, Table 3A**) In support of these results, multiple regression analysis on the 100 genes with both additional superfamily members s and adjacent pathway genes p yields a covariance of 30% for evolutionary rate and 27% for selective pressure across this set of metabolic genes ($r_{dn}^2 = .2998$, $r_{ds}^2 = .0148$, $r_{\omega}^2 = .2676$, $n = 100$ genes).

As a further test for independence, we performed partial correlation to help determine whether the correlations in evolutionary rate among superfamily members and among network neighbors were due to an underlying correlation with expression level. Partial correlation analysis can be used to test for correlations between two variables while controlling for a third. Here, $r_{AB|C}$ denotes the partial correlation coefficient between any two variables, A and B, while controlling for a third, C. Due to limited mRNA expression and protein abundance data, we used the commonly applied CAI as the best available proxy measure of expression level [4, 6, 13, 25]. Partial correlation analysis indicates that pathway neighbors and SCOP superfamily represent largely independent predictors equivalent in magnitude (**Table 3B**). In particular, p and s do not correlate significantly with each other ($d_n = 0.17$, $P = 0.0850$; $\omega = 0.16$, $P = 0.1063$) when controlling for variation in evolutionary rate. Importantly, a gene's evolutionary rate correlates neither with the expression level of its metabolic network neighbors ($d_n = 0.06$, $P = 0.51$; $\omega = 0.04$, $P = 0.66$, **Table 3C**), nor the expression level of other superfamily members ($d_n = 0.14$, $P = 0.17$; $\omega = -0.13$, $P = 0.17$, **Table 3D**), controlling for variation in evolutionary rate across network neighbors and SCOP superfamily members, respectively.

3.4 Discussion

This study accounts for variance in the evolutionary rates of *Saccharomyces* small molecule metabolic enzymes unexplained by expression level, and identifies additional potential evolutionary constraints that may be encoded directly in the genome. We demonstrate modest but significant correlations in evolutionary rate among SCOP superfamily members and among enzymes adjacent in the metabolic network. The SCOP superfamily members studied here are distant homologs that retain similar protein structure, yet catalyze distinct chemical reactions. Adjacent enzymes in small molecule metabolic pathways are joined by their role in the conversion of a substrate into a biologically useful product. Our results suggest that these constraints are independent from both each other and from expression level.

These results mirror those in a recent review that identified four factors with purifying selection and influence on evolutionary rate – expression level, structure and stability (represented in part by our correlations among SCOP superfamily members), pleiotropy (reflected in our metabolic network context analyses), and dispensability [2].

As a tool for understanding protein evolutionary rates, expression level fails to explain a significant portion of the constraint on evolutionary rate and selective pressure. For our dataset, and as a recent study suggested for a different and larger sample set, expression level explains less than half of the observed variance in evolutionary rate [26].

Of the total variance in d_n across our dataset expression level accounts for just 29% and dispensability 11%. Importantly, superfamily and metabolic network context account for another 25% (**Fig 1, Tables 1 and 2**). The remaining 35% variance in evolutionary rate may be addressed by some of the many factors noted in the literature, but remains an open question in this study.

In addition to the inability of expression level to explain the majority of variance in our dataset, we note that expression level is a problematic metric for both practical and conceptual reasons. As a practical matter, obtaining expression profiles and protein abundances for all organisms whose genome has been completely sequenced presents an enormous challenge. Useful proxy measures for expression level such as CAI are based upon codon usage bias. Yet CAI cannot be consistently applied across all genomes since codon usage bias varies widely between organisms, and sometimes does not appear at all, notably in humans. Conceptually, the mechanisms underlying the intriguing correlation between expression level and evolutionary rate are uncertain and the biological interpretation unclear. For example, expression profiles may diverge rapidly with gene duplication or speciation while the expressed genes themselves remain relatively unchanged over long periods of time [36]. It is unknown how this difference in evolutionary window between changes in expressional level and changes in the expressed genes impacts the observed correlation. For these reasons, studies are needed to shed additional light on the relative importance and specific mechanisms of constraints on protein evolution.

3.4.1 The Independence of SCOP Superfamily and Network Context from Expression Level

We have presented evidence based upon both principal component and partial correlation analyses that our correlations in evolutionary rate among SCOP superfamily members and among metabolic network neighbors are largely independent from expression level. The principal component analysis reveals three significant components. Measures related to expression level dominate the largest component, but the SCOP superfamily and metabolic network context dominate the second largest component and explain considerable variance (24%) across the dataset. This is the first systematic study to demonstrate an independent contribution to evolutionary rate for these two variables.

Partial correlation analysis contributes further evidence that the SCOP and network neighbor correlations are not only independent from expression, but also independent from each other. Metabolic enzyme evolutionary rates have a strong negative correlation with CAI, a proxy measure of expression level. Conversely, a gene's evolutionary rate does not correlate with the CAI of superfamily members. Surprisingly, a gene's evolutionary rate also does not correlate with the CAI of metabolic network neighbors. We note that noise is known to be a confounding factor in partial correlation analysis. In particular, due to noise, two variables could falsely appear correlated because of underlying and independent correlations with a third variable. Although our results do not definitively prove the independence of these variables, given the absence of any significant partial correlation in CAI between superfamily members or between network

neighbors, we conclude that these correlations are indeed independent from expression level.

3.4.2 The Independence of SCOP Superfamily from Network Context

We have presented evidence that network context and structural similarity are independent in their correlations with evolutionary rate. Regarding functional independence, only 7 of the SCOP superfamily pairs appear in the metabolic network pairs dataset. The low incidence of SCOP superfamily members appearing in metabolic pathways is consistent with other reports. For example, working in the *E. coli* model system, Teichmann et al. found that only 8 of 106 metabolic pathways contained significant numbers of homologs [35]. Teichmann et al. also reported that homologs were twice as likely to be distributed between pathways as within pathways. Regarding correlations in evolutionary rate, we note that SCOP superfamily and metabolic network neighbors together correlate better with a gene's evolutionary rate than either measure alone (**Fig. 3**). Partial correlation analyses as well supports the hypothesis of independence. When controlling for variation in a gene's evolutionary rate, there is no significant correlation between SCOP superfamily and adjacent pathway members (**Table 3**). Together, these lines of evidence point to independence between SCOP superfamily and metabolic network context with respect to evolutionary rate. Finally, we note that whether or not our work establishes SCOP superfamily and network context as independent from each other, together they account for a quarter of the variation in evolutionary of the metabolic enzymes, independent of expression level.

3.4.3 SCOP Superfamily

Our results suggest that, from a global perspective, some enzyme scaffolds may indeed be more constrained in evolutionary rate than others. This is consistent with a broad and deep body of literature showing that some protein scaffolds can tolerate more mutations than others and still function [11, 36-41]. For example, roughly 10% of all enzymes contain an $(\alpha/\beta)_8$ barrels domain [42-44]. The impressive functional versatility and sequence diversity of this fold suggest high inherent evolvability. Indeed, single-site mutants of the barrel domains in both the L-Ala-D/L-Glu epimerase from *Escherichia coli* (AEE) and the muconate lactonizing enzyme II from *Pseudomonas* sp. P51 (MLE II) catalyze the *o*-succinylbenzoate synthase (OSBS) reaction while maintaining competence for the wild-type reaction [45].

Such tolerance for mutations may be attributed to at least two factors. First, abundant evidence has been gathered from *in vitro* studies linking evolvability, or ‘designability’, with thermodynamic stability. However, *in vivo* reports remain anecdotal. Thus, the stability hypothesis must be tempered by the fact that although minimum stability is a precondition for proper protein function, stability does not directly correlate with functional competence.

Second, functional requirements of proteins are known to constrain evolutionary rates at specific residues [37, 46-48]. This fundamental rule underlies numerous comparative analyses of proteins and nucleic acids [11, 38, 39, 49-52]. For example, genomic sequencing efforts have been remarkably successful at generating functional

hypotheses for novel genes. The process typically begins with an inference of homology based on sequence or structure, followed by functional annotation transfer from previously characterized proteins. Functional annotation transfer presumes that homologous proteins will perform similar functions. Many widely used genomic and proteomic databases, such as PFAM [53] and SCOP [17], are explicitly structured around this principle. This approach presumes site-specific evolutionary constraints arise primarily from individual protein structural and chemical requirements. Such site-specific constraints contribute to the mechanisms by which some protein scaffolds are inherently more or less evolvable than others [36, 54, 55].

Because it has been so broadly observed that functional requirements in homologous enzymes constrain evolutionary rate at specific residues, constraints represented by SCOP superfamilies might be expected to dominate small molecule metabolic enzyme evolution. Surprisingly, while SCOP superfamily members correlate in evolutionary rate and selective pressure, SCOP superfamily represents only a modest fraction of the total variance (12%). This result is consistent with observations that the structures of individual enzymes contributing to primary and secondary metabolism can vary appreciably [56]. For example, in different species multiple enzymes that are evolutionarily unrelated have been shown to catalyze identical reactions in both thiamine biosynthesis [57] and S-adenosyl methionine modification [58]. Thus, it appears that the need for metabolic competence can conscript different proteins scaffolds that are evolutionarily unrelated to evolve new functions.

3.4.4 Metabolic Network Context

While structural similarity among homologous enzymes accounts for a modest percentage of variance in evolutionary rates, this study also demonstrates that enzymes adjacent in the metabolic network evolve at similar rates. This effect is separable from expression level, codon adaptation index, protein abundance, and inherent constraints in the structural superfamilies represented in the pathway. There are many reasons selection may occur in a metabolic network. For example, metabolic pathway reactions usually involve highly similar substrates whose chemistry may impose similar constraints across the pathway [14].

Kinetic efficiency across metabolic pathways has been shown to be important for pathway stability and could therefore be subject to selection [14]. Thus, selection for optimal kinetics in a pathway, rather than in individual enzymes, suggests another potential mechanism by which selection could act on metabolic pathways. A classic example of this is triose phosphate isomerase (TIM). Extensive studies have shown that TIM has achieved diffusion-limited catalytic efficiency [9]. This feature is biologically less meaningful, however, if the substrate is not provided in concentrations sufficient to benefit from this high level of efficiency, or if the product of catalysis is not then required by the cell at equally high rates.

Finally, we note that conserved pathways must generate a biologically useful product, or degrade a harmful substrate. Selection therefore conserves pathway output, but not necessarily the individual enzymes comprising the pathway. As illustrated by the

case of S-adenosyl methionine modification discussed above, multiple unrelated structures may perform the same reaction, allowing conservation of pathway output without conservation of the specific components within pathways. Furthermore, enzymes once thought to catalyze a single specific reaction are increasingly recognized as promiscuous, allowing a single structure to contribute different functions [59]. For example, tetrachlorohydroquinone (TCHQ) dehalogenase from *Sphingobium chlorophenolicum* catalyzes the replacement of two chlorine substituents on TCHQ, allowing the soil bacterium to degrade the anthropogenic pesticide pentachlorophenol. Yet in the same active site, the enzyme also has maleylacetoacetate isomerase activity [60]. Such pathway mutability and *ad hoc* assembly is beginning to feature prominently in how we describe the evolutionary dynamics of, and ultimately engineer, new metabolic pathways [61, 62]. The fitness benefit conferred by these pathways, along with the frequency and intensity of selection, may thus drive similar evolutionary rates among the constituent enzymes.

3.5 Conclusion

Our results demonstrate that both network context and protein structure provide constraints on evolutionary rates among small molecule metabolic enzymes. These predictors are measurable in genomic data, can be relatively weighed, and contribute information independent from expression level. Altogether, measures related to expression level, dispensability, structural superfamily, and metabolic network context explain 64% of the variance in evolutionary rate across our dataset.

Although we cannot explain all the variance in evolutionary rate, studies from disparate disciplines have added to a long list of factors implicated as determinants of evolutionary rate that may address the additional variance unaccounted for here. As advocated elsewhere [2], integrated analyses such as provided in this work that both demonstrate significant correlations with evolutionary rate and distinguish the proportion of variance a given factor may explain, constitute the next step in further elucidating the principles of protein evolution. Such syntheses will impact fundamental questions such as how we search for positive selection in our own genome, how we predict the function of genes from recently sequenced organisms, and how we engineer novel ones.

3.6 Methods

3.6.1 Evolutionary Metrics. We use the evolutionary rate data published by Hirsh et al. The data derive from the comparative analysis of orthologous coding sequences identified across four yeast species (*Saccharomyces cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*) by Kellis et al.[24] Together, these genomes provide excellent resolution of molecular evolution across a phylogeny of closely related eukaryotic organisms (80% genic sequence identity) with minimal lateral gene transfer. Using a correction for the well-known codon usage bias in yeast [30], d_n , d_s , and d_n/d_s (also known as ω) have been calculated by Hirsh et al based upon the Kellis ortholog assignments. Coding mutation rates (d_n) provide an estimate of evolutionary rate with respect to orthologs in other species, essentially sequence distance since time of divergence. Non-coding mutation

rates (d_s), presumed to be neutral to selection, serve as an internal control by approximating time since divergence. Selective pressure on a particular gene can be estimated by calculating ratios of coding versus non-coding mutation rates (ω) with respect to orthologs in other species. Ratios of 1, >1 , and <1 indicate neutral evolution, positive selection, and negative selection, respectively. The combined metric c is the average of s and p , where s is the average of the other members of the same superfamily and p is the average of all adjacent genes in the metabolic network.

3.6.2 Biochemical Network. For our metabolic network model, we use the Yeast Biochemical Pathways, formerly known as YeastCyc, hosted by the Saccharomyces Genome Database (SGD) [23]. The current set (generated September 19, 2005) of 138 pathways contains 781 polypeptides corresponding to 925 enzymatic reactions on 675 compounds. Since the functions of many of the yeast genes are not yet known, pathways and reaction assignments may be incomplete. Nevertheless, this metabolic network model represents the best-curated dataset available uniting yeast metabolic pathways and genomic data.

All enzymes in the reconstructed network are unique, but may be annotated to multiple pathways. For example, fructose 1,6-bisphosphate aldolase is annotated to several overlapping pathways including glycolysis, gluconeogenesis, glucose fermentation, mannitol degradation, and sorbitol degradation. When analyzing network pairs, each enzyme pair including fructose 1,6-bisphosphate aldolase enzyme is counted only once. However, when we permute within pathways to test pathway order, enzyme pairs are not

unique because they may catalyze the same adjacent reactions in multiple pathways. For instance, fructose 1,6-bisphosphate aldolase is annotated with identical adjacent enzymes to both the glucose fermentation and glycolysis pathways.

SCOP. The Structural Classification of Proteins (SCOP) [17] database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships. Whole genome assignments based upon the 1.67 SCOP release are hosted on the Superfamily Database [63]. The Superfamily Database search method uses a library (covering all proteins of known structure) consisting of 1447 SCOP superfamilies, each of which is represented by a group of hidden Markov models. In this report, we use the best match for each domain with an e-value cutoff of e^{-10} . Correlations and significance are similar over a wide range of e-value cutoffs, but begin to decrease with more permissive e-values. Proteins may be assigned to more than one superfamily, but in practice this is rare, given that most yeast small molecule metabolic enzymes annotated in SGD comprise single domains. All SCOP superfamily members included in our analyses are annotated to a small molecule metabolic reaction. Additionally, each superfamily included in our analyses has at least two members annotated to different reactions.

3.6.3 Transcriptional Modules. Ihmels et al. [21] established that genes associated with similar metabolic functions are likely to exhibit a similar expression pattern. They characterized the regulation of genes associated with adjacent metabolic reactions and provide a database describing the coregulated genes and the regulatory conditions

associated with most metabolic pathways. Confidence values are given for each gene assignment to a particular metabolic transcriptional module. Here we report results calculated on this data set using a confidence value cutoff of 80% for inclusion of a gene in a transcriptional module.

3.6.4 Statistical Methods. We began by generating three sets of pairs representing constraints of interest: all possible enzyme pairs in the same structural superfamily, all enzyme pairs that catalyze adjacent reactions in the metabolic network, and all possible pairs in the same transcriptional module (see Supplemental Data). Pairs are excluded from analysis when evolutionary rate data is not available for both genes. To avoid undue bias towards the largest large superfamilies and transcriptional modules, we used a cutoff of 400 gene pairs (i.e. 20 genes) per superfamily or transcriptional module. Correlation coefficients are then calculated for the respective sets of pairs. For the sake of clarity and readability, we generally report only parametric (Pearson) results. In cases where parametric and non-parametric (Spearman rank order) disagree considerably, we have listed both. Calculated P-values are all given in scientific notation. Bootstrap P-values from the null model are given exactly in full notation.

We use a label-permuted null model to bootstrap p-values for the network context, SCOP superfamily, and the combined analyses (Figs. 2 and 3). We bootstrap these P-values due to non-normal distribution of evolutionary rates and the presence of the same gene in multiple pairs. The bootstrap P-values tend to be more conservative than calculated P-values by approximately 10^1 . These results are report exactly to four significant digits, all

other results are given in scientific notation. The null model was generated by randomizing data values across a given sample set of genes for 10,000 iterations. Randomizing in this way preserves network connectivity. For example, enzymes catalyzing reactions that feed into many biochemical pathways are similarly represented in the null model, but with a randomly assigned evolutionary rate. The randomization of data labels is performed only across the genes in a given sample set, not across the entire metabolic network.

We used R (Ihaka and Gentleman 1996) with the `.pls` package to perform the multiple regression and principal component analyses, as previously described [26]. We log transformed codon adaptation index and mRNA expression. We decided whether or not to log transform a variable based on whether log transformation led to a higher variance (r^2). We scaled the predictor variables to zero mean and unit variance before carrying out the principal component analysis. For further details, see [26].

3.7 Acknowledgments

We wish to thank Naama Barkai (Weizmann Institute) and the curators at the Saccharomyces Genome Database for providing access to their datasets. We also wish to thank Jeff Chuang, Annie Tsong, and Mike Kim (UCSF) for helpful discussion and careful critique of this manuscript. However, any faults are ours alone. This work was supported by GM60595 (P.C.B.) and the UCSF Mentorship & Research Assistantship Program (J.C.A.).

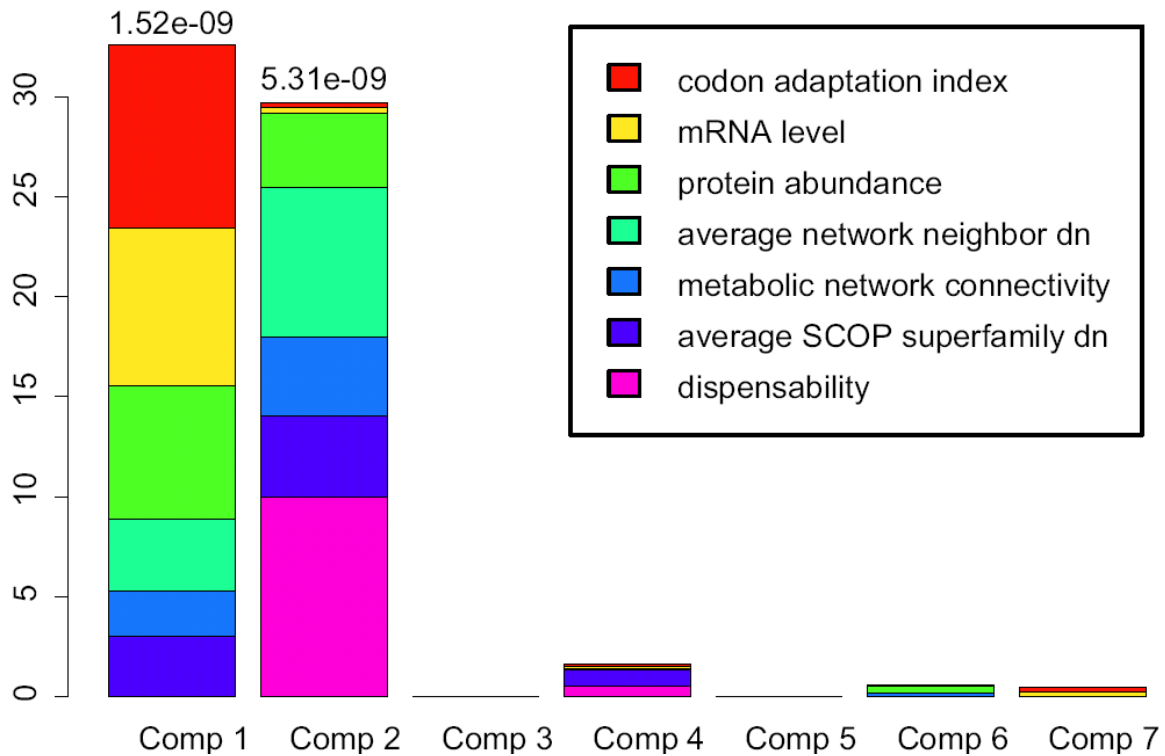


Figure 3. 1| Principal component analysis (PCA) of evolutionary rate (d_n) in small molecule metabolic enzymes yields two highly significant components. PCA of the evolutionary rate of small molecule metabolic enzymes ($n = 64$) using seven input variables (see box above) yields two highly significant components (P-values appear above the significant components).

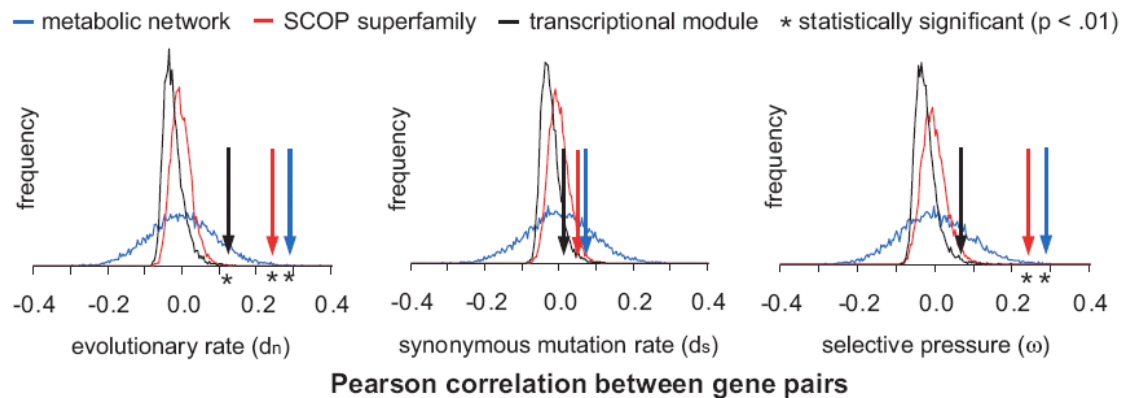


Figure 3. 2| Pearson correlation between small molecule metabolic enzyme pairs according to constraint: metabolic network, SCOP superfamily, and transcriptional module. Histograms display null distributions with drop arrows indicating observed values from the four *Saccharomyces* yeast species. The analysis includes 143 genes in 120 adjacent metabolic network pairs, 264 genes in 897 SCOP superfamily pairs, and 98 genes in 1,137 transcriptional module pairs. Null distributions were generated from 10,000 label-permuted networks that preserved overall topology. See Additional File for plots of the raw data used in this analysis.

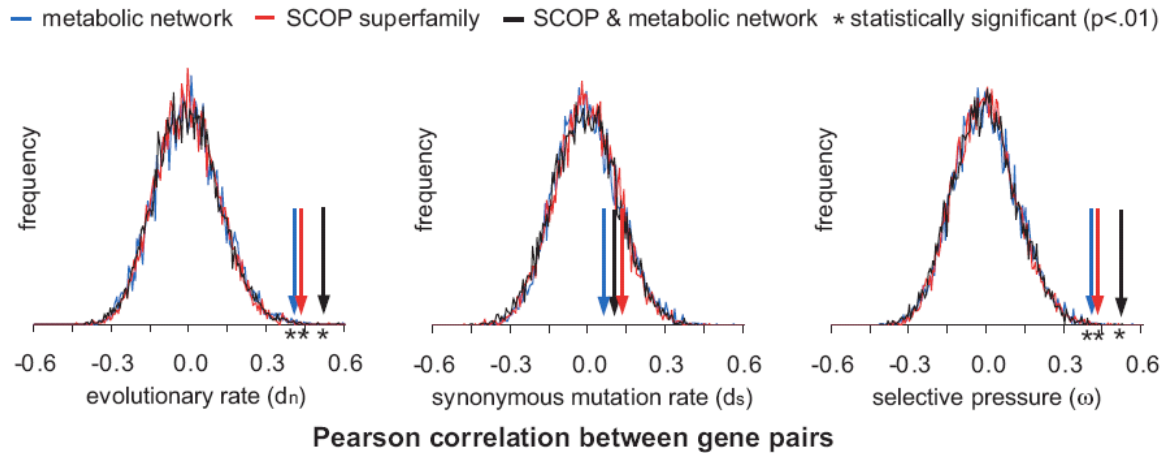


Figure 3. 3| Pearson correlation between small molecule metabolic enzyme pairs with orthogonal constraints: metabolic network only, SCOP superfamily only, metabolic network and SCOP superfamily combined. Histograms display null distributions with drop arrows indicating observed values from the four *Saccharomyces* yeast species. The figure includes only the 100 enzymes with both adjacent metabolic network pairs and additional SCOP superfamily members conserved across all four yeast species. Correlations are between the query gene and the average of the adjacent network pairs (p), the average of the additional SCOP superfamily members (s), or the average of both (c , or $(p+s)/2$). Null distributions were generated from 10,000 label-permuted networks that preserved overall topology. See Additional File for plots of the raw data used in this analysis.

| Variable | % variance d_n explained | % variance ω explained |
|-----------------------------------|--|---|
| network neighbor d_n | 11.189 | 14.918 |
| protein level | 10.704 | 14.945 |
| dispensability | 10.631 | 17.954 |
| codon adaptation index | 9.781 | 0.984 |
| mRNA level | 8.487 | 2.644 |
| superfamily d_n | 7.809 | 8.085 |
| connectivity in metabolic network | 6.371 | 10.869 |
| total | 64.97 | 70.40 |

Table 3.1| Principal component analysis yields multiple variables that together explain variation in metabolic enzyme evolutionary rates.

| | Pearson r_{dn} | P-value | Spearman r_{dn} | P-value | # genes | # pairs |
|------------------|------------------|----------|-------------------|----------|---------|---------|
| codon adaptation | 0.39 | 1.40E-17 | 0.28 | 2.80E-09 | 333 | 433 |
| mRNA level | 0.43 | 4.30E-19 | 0.19 | 1.70E-04 | 311 | 266 |
| protein level | 0.13 | 3.20E-02 | 0.22 | 3.80E-04 | 235 | 266 |

Table 3.2| Expression level measures correlate in adjacent metabolic network pairs.

| A. Combining Pathway Context and Structural Superfamily | | | | | | |
|---|-----------------------|---------|-----------------------|---------|-----------------------|---------|
| | r_s | P-value | r_p | P-value | r_c | P-value |
| d_n | 0.45 | 0.0005 | 0.47 | 0.0009 | 0.56 | <0.0001 |
| ω | 0.43 | 0.0005 | 0.44 | 0.0008 | 0.53 | <0.0001 |
| d_s | 0.09 | 0.22 | 0.13 | 0.1309 | 0.13 | 0.1288 |
| B. Partial Correlation Analysis - Pathway Context and SCOP | | | | | | |
| | $r_{\text{gene},s p}$ | P-value | $r_{\text{gene},p s}$ | P-value | $r_{p,s \text{gene}}$ | P-value |
| d_n | 0.35 | 0.0027 | 0.35 | 0.004 | 0.17 | 0.093 |
| ω | 0.33 | 0.0049 | 0.34 | 0.0038 | 0.16 | 0.1092 |
| d_s | 0.1 | 0.1737 | 0.05 | 0.3176 | 0.17 | 0.0949 |
| C. Partial Correlation Analysis - Pathway Context and Expression Level | | | | | | |
| | $r_{\text{gene},p x}$ | P-value | $r_{p,x \text{gene}}$ | P-value | $r_{\text{gene},x p}$ | P-value |
| d_n | 0.4 | 0.0017 | -0.55 | <0.0001 | 0.06 | 0.268 |
| ω | 0.37 | 0.0023 | -0.544 | <0.0001 | 0.04 | 0.3094 |
| d_s | -0.08 | 0.2148 | -0.13 | 0.155 | 0.06 | 0.287 |
| D. Partial Correlation Analysis - SCOP Superfamily and Expression | | | | | | |
| | $r_{\text{gene},s x}$ | P-value | $r_{s,x \text{gene}}$ | P-value | $r_{\text{gene},x s}$ | P-value |
| d_n | 0.28 | 0.0101 | -0.57 | <0.0001 | -0.14 | 0.0802 |
| ω | 0.26 | 0.0146 | -0.64 | <0.0001 | -0.13 | 0.085 |
| d_s | 0.12 | 0.1365 | 0.07 | 0.3377 | -0.14 | 0.0864 |

Table 3.3| Pathway context and structural superfamily independently correlate with metabolic enzyme evolutionary rates. Boxes are highlighted in pale blue to indicate significant results (P-value < 0.05) and blue to indicate highly significant results (P-value < 0.01). Correlations are shown for evolutionary rate (d_n), selective pressure (ω), and synonymous mutation rate (d_s). The combined metric c is the average of s and p , where s is the average of the other members of the same structural (SCOP) superfamily and p is the average of all adjacent genes in the metabolic network. x is codon adaptation index (CAI) of metabolic network neighbors (C) or SCOP superfamily members (D). r_{ABC} denotes the partial correlation coefficient between any two variables A and B, while controlling for a third C.

3.8 References

1. Rocha EP: **The quest for the universals of protein evolution.** *Trends Genet* 2006, **22**(8):412-416.
2. Pal C, Papp B, Lercher MJ: **An integrated view of protein evolution.** *Nat Rev Genet* 2006, **7**(5):337-348.
3. Koonin EV: **Systemic determinants of gene evolution and function.** *Mol Syst Biol* 2005, **10**(1038):msb4100029.
4. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH: **Why highly expressed proteins evolve slowly.** *Proc Natl Acad Sci U S A* 2005, **102**(40):14338-14343.
5. Fraser HB, Hirsh AE: **Evolutionary rate depends on number of protein-protein interactions independently of gene expression level.** *BMC Evol Biol* 2004, **4**:13.
6. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW: **Functional genomic analysis of the rates of protein evolution.** *Proc Natl Acad Sci U S A* 2005, **102**(15):5483-5488.
7. Chuang JH, Li H: **Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome.** *PLoS Biol* 2004, **2**(2):E29.
8. Vitkup D, Kharchenko P, Wagner A: **Influence of metabolic network structure and function on enzyme evolution.** *Genome Biol* 2006, **7**(5):R39.
9. Albery WJ, Knowles JR: **Evolution of enzyme function and the development of catalytic efficiency.** *Biochemistry* 1976, **15**(25):5631-5640.
10. Benkovic SJ, Hammes-Schiffer S: **A perspective on enzyme catalysis.** *Science* 2003, **301**(5637):1196-1202.
11. Gerlt JA, Babbitt PC: **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies.** *Annu Rev Biochem* 2001, **70**:209-246.
12. Pain RH: **The Evolution of Enzyme Activity.** *Nature* 1982, **299**(07 October 1982):486-487.
13. Sharp PM, Li WH: **The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.** *Nucleic Acids Res* 1987, **15**(3):1281-1295.

14. Alves R, Chaleil RA, Sternberg MJ: **Evolution of enzymes in metabolism: a network perspective.** *J Mol Biol* 2002, **320**(4):751-770.
15. Rison SC, Thornton JM: **Pathway evolution, structurally speaking.** *Curr Opin Struct Biol* 2002, **12**(3):374-382.
16. Schmidt S, Sunyaev S, Bork P, Dandekar T: **Metabolites: a helping hand for pathway evolution?** *Trends Biochem Sci* 2003, **28**(6):336-341.
17. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247**(4):536-540.
18. Valencia A, Pazos F: **Computational methods for the prediction of protein interactions.** *Curr Opin Struct Biol* 2002, **12**(3):368-373.
19. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE: **Co-evolution of proteins with their interaction partners.** *J Mol Biol* 2000, **299**(2):283-293.
20. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the database of interacting proteins.** *Nucleic Acids Res* 2000, **28**(1):289-291.
21. Ihmels J, Levy R, Barkai N: **Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*.** *Nat Biotechnol* 2004, **22**(1):86-92.
22. Ihmels J, Bergmann S, Barkai N: **Defining transcription modules using large-scale gene expression data.** *Bioinformatics* 2004, **20**(13):1993-2003.
23. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, Engel SR, Feierbach B, Fisk DG, Hirschman JE *et al*: **Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms.** *Nucleic Acids Res* 2004, **32**(Database issue):D311-314.
24. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*.** *Nature* 2004, **428**(6983):617-624.
25. Hirsh AE, Fraser HB, Wall DP: **Adjusting for selection on synonymous sites in estimates of evolutionary distance.** *Mol Biol Evol* 2005, **22**(1):174-177.
26. Drummond DA, Raval A, Wilke CO: **A Single Determinant Dominates the Rate of Yeast Protein Evolution.** *Mol Biol Evol* 2005.

27. Chin CS, Chuang JH, Li H: **Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence.** *Genome Res* 2005, **15**(2):205-213.
28. Rausher MD, Miller RE, Tiffin P: **Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway.** *Mol Biol Evol* 1999, **16**(2):266-274.
29. Cork JM, Purugganan MD: **The evolution of molecular genetic pathways and networks.** *Bioessays* 2004, **26**(5):479-484.
30. Kliman RM, Irving N, Santiago M: **Selection conflicts, gene expression, and codon usage trends in yeast.** *J Mol Evol* 2003, **57**(1):98-109.
31. Gu X, Zhang Z, Huang W: **Rapid evolution of expression and regulatory divergences after yeast gene duplication.** *Proc Natl Acad Sci U S A* 2005, **102**(3):707-712.
32. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV: **Conservation and coevolution in the scale-free human gene coexpression network.** *Mol Biol Evol* 2004, **21**(11):2058-2070.
33. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**(5):717-728.
34. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast.** *Nature* 2003, **425**(6959):737-741.
35. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli.** *J Mol Biol* 2001, **311**(4):693-708.
36. Babbitt PC, Gerlt JA: **New functions from old scaffolds: how nature reengineers enzymes for new functions.** *Adv Protein Chem* 2000, **55**:1-28.
37. Dickerson RE, Geis I: **The structure and action of proteins.** Menlo Park, Calif.: Benjamin/Cummings Pub. Co.; 1969.
38. Golding GB, Dean AM: **The structural basis of molecular adaptation.** *Mol Biol Evol* 1998, **15**(4):355-369.
39. Knudsen B, Miyamoto MM, Laipis PJ, Silverman DN: **Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases.** *Genetics* 2003, **164**(4):1261-1269.

40. Smock RG, Gierasch LM: **Finding the fittest fold: using the evolutionary record to design new proteins.** *Cell* 2005, **122**(6):832-834.
41. Tian W, Skolnick J: **How well is enzyme function conserved as a function of pairwise sequence identity?** *J Mol Biol* 2003, **333**(4):863-882.
42. Reardon D, Farber GK: **The structure and evolution of alpha/beta barrel proteins.** *Faseb J* 1995, **9**(7):497-503.
43. Hegyi H, Lin J, Greenbaum D, Gerstein M: **Structural genomics analysis: characteristics of atypical, common, and horizontally transferred folds.** *Proteins* 2002, **47**(2):126-141.
44. Farber GK, Petsko GA: **The evolution of alpha/beta barrel enzymes.** *Trends Biochem Sci* 1990, **15**(6):228-234.
45. Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, Babbitt PC, Minshull J, Gerlt JA: **Evolutionary potential of (beta/alpha)₈-barrels: functional promiscuity produced by single substitutions in the enolase superfamily.** *Biochemistry* 2003, **42**(28):8387-8393.
46. Li W-H: **Molecular evolution.** Sunderland, Mass.: Sinauer Associates; 1997.
47. Kimura M: **The neutral theory of molecular evolution.** Cambridge [Cambridgeshire] ; New York: Cambridge University Press; 1983.
48. Nei M, Kumar S: **Molecular evolution and phylogenetics.** Oxford ; New York: Oxford University Press; 2000.
49. Gaucher EA, Gu X, Miyamoto MM, Benner SA: **Predicting functional divergence in protein evolution by site-specific rate shifts.** *Trends Biochem Sci* 2002, **27**(6):315-321.
50. Thorne JL: **Models of protein sequence evolution and their applications.** *Curr Opin Genet Dev* 2000, **10**(6):602-605.
51. Babbitt PC: **Definitions of enzyme function for the structural genomics era.** *Curr Opin Chem Biol* 2003, **7**(2):230-237.
52. Jones S, Thornton JM: **Searching for functional sites in protein structures.** *Curr Opin Chem Biol* 2004, **8**(1):3-7.
53. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32**(Database issue):D138-141.

54. Meyerguz L, Grasso C, Kleinberg J, Elber R: **Computational analysis of sequence selection mechanisms.** *Structure (Camb)* 2004, **12**(4):547-557.
55. Dokholyan NV, Shakhnovich EI: **Understanding hierarchical protein evolution from first principles.** *J Mol Biol* 2001, **312**(1):289-307.
56. Copley RR, Letunic I, Bork P: **Genome and protein evolution in eukaryotes.** *Curr Opin Chem Biol* 2002, **6**(1):39-45.
57. Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P: **Systematic discovery of analogous enzymes in thiamin biosynthesis.** *Nat Biotechnol* 2003, **21**(7):790-795.
58. Schubert HL, Blumenthal RM, Cheng X: **Many paths to methyltransfer: a chronicle of convergence.** *Trends Biochem Sci* 2003, **28**(6):329-335.
59. O'Brien PJ, Herschlag D: **Catalytic promiscuity and the evolution of new enzymatic activities.** *Chem Biol* 1999, **6**(4):R91-R105.
60. Copley SD: **Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach.** *Trends Biochem Sci* 2000, **25**(6):261-265.
61. Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS: **The 'evolvability' of promiscuous protein functions.** *Nat Genet* 2005, **37**(1):73-76.
62. Bornscheuer UT, Kazlauskas RJ: **Catalytic promiscuity in biocatalysis: using old enzymes to form new bonds and follow new pathways.** *Angew Chem Int Ed Engl* 2004, **43**(45):6032-6040.
63. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313**(4):903-919.

3.8 Epilogue

While this manuscript was in review, Plotkin and Fraser published an article titled “Assessing the Determinants of Evolutionary Rates in the Presence of Noise” [1]. The article purports to demonstrate that differences in noise when measuring biological predictor variables, combined with the susceptibility of principal component analysis to variability in noise, invalidates principal component analysis as a method to tease apart the drivers of the evolutionary rate of proteins. The article by Plotkin and Fraser was a direct response to the work of Drummond and colleagues [2] which itself claimed to invalidate the reliability of another technique, multiple regression analysis, widely used to study evolutionary rate. Multiple regression analysis is not robust to outlier data, and is similarly vulnerable to noisy predictor variables. Drummond and colleagues proposed that principal component analysis could overcome these issues, and based upon their analysis, claimed that expression level was the single determinant for the evolutionary rate of proteins. The mathematical basis of their claim was flatly rejected by Plotkin and Fraser, whose findings have recently been extended to population level studies. Kryazhimskiy and Plotkin have demonstrated not only that the ratio of synonymous to non-synonymous mutation rate (d_n/d_s) is relatively insensitive to selective pressure at the population level, but that the hallmark signature of selection, $d_n/d_s > 1$, is frequently violated at the population level [3].

Disputes over methodology continue unabated in the field of molecular evolution. More sophisticated mathematical techniques and, more importantly, biologically sophisticated experimental models will be required to tease apart the contributions of

various constraints on evolutionary rate. While we are convinced that our work remains important, robust, and correct – particularly with respect to selection at the pathway level – serious methodological issues must first be resolved in this area of evolutionary biology. In light of these facts, and understanding that both mathematical modeling and the design of biochemical systems to dissect mechanisms of selection are not our primary research focus, we made the difficult decision to withdraw the manuscript from review. Others investigators continue to publish relying on the same methods discussed here (for a recent example see [4]), but the value of those contributions and the potentially spurious results will remain in doubt until these fundamental methodological issues have been resolved.

3.9 Epilogue References

1. Plotkin JB, Fraser HB (2007) Assessing the determinants of evolutionary rates in the presence of noise. *Mol Biol Evol* 24: 1113-1121.
2. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23: 327-337.
3. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS Genet* 4: e1000304.
4. Jovelin R, Phillips PC (2009) Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol* 10: R35.

Appendix A. The Chemical Diversity of Drugs and Metabolites

A.1 Chemical similarity among drugs and among metabolites

Although drugs and metabolites typically differ in their physiochemical properties, significant and specific similarity links nonetheless emerged. Using SEA at an expectation value cutoff of $E = 1.0 \times 10^{-10}$, 54% (132 of 246) of drug sets link to an average of 43.7 (median = 10) or 0.9% of metabolic reactions (**Figure A.1A**). None of the remaining 46% (114 of 246) of drug sets link to any metabolic reaction sets. Similarly, 67% (1,790 of 5,371) of reaction sets do not link to any MDDR drug set at the expectation value cutoff of $E = 1.0 \times 10^{-10}$, but those that do hit an average of 2.8 (median = 2) or 1.1% of drug sets (**Figure A.1B**).

Our analysis of the intersection between drugs and metabolism reveals substantial regions of unexplored chemical space. To investigate why so many regions remain open, we compared the patterns of chemical similarity among drugs to the patterns of chemical similarity among metabolites. Most set comparisons in an exhaustive all-by-all analysis yield little or no significant similarity. However, through sequential linkage, archipelagos of strong similarity emerge that allow us to connect almost all sets together into two single networks – a drug network and a metabolic network. Of the 246 MDDR drug sets, each set linked to an average of 6.6 (median = 4) or 2.3% of all other drug sets with an

expectation value of 1.0×10^{-10} or better (**Figure A.1C**). Metabolic reactions were more promiscuous with each of the 5,056 MetaCyc reactions linking to an average of 473.0 (median = 370) or 9.4% of all other reactions (**Figure A.1D**).

To extract the most chemically relevant links, we applied the Floyd-Warshall algorithm which guarantees the best path (weighted by expectation value) between any two nodes. Biologically related nodes cluster together by the chemical similarity of their ligands, even though no explicit biological information was used to link them (**Figure A.2**). In the drug network, the serotonin (5HT) receptor agonists/antagonists group together, as do adenosine receptor agonists/antagonists and phospholipase inhibitors. Some classes additionally segregate into their appropriate subtypes, such as the α -adrenergic and β -adrenergic receptor agonists/antagonists. In the metabolic network, clusters often reflect neighboring reactions within a pathway or similar reactions shared across different pathways. For instance, the first five reactions in the purine degradation pathway all cluster together, as do the last five reactions of flavin synthesis. Viewing metabolic networks by chemical similarity allows links among ligand sets independent of their canonical pathway organization. Three reactions involving precursors and metabolites of chorismate (chorismate synthase, chorismate mutase, and isochorismate mutase) group together despite annotation to three different pathways (chorismate, tyrosine, and menaquinone biosynthesis, respectively). Similarly, the antifolate drug targets DHFR, GART, and thymidylate synthase catalyze reactions in multiple pathways including tetrahydrofolate, pyrimidine, and purine biosynthesis. Yet all three enzymes

recognize antifolates, demonstrate synergy as targets, and cluster together in the MDDR drug network.

A.2 Topological differences between drug and metabolic space

Although the same technique was used to prune and visualize these two networks, a prominent difference in connectivity (k) emerged. In the Floyd-Warshall graphs, MetaCyc reaction sets connect on average to 8.6 (median $k = 2$) other metabolic nodes (**Figure A.2A**) while MDDR drug sets connect on average to just 2.0 (median $k = 2$) other drug sets (**Figure A.2B**).

Several highly connected metabolic network hubs reveal the striking difference in connectivity. Amino acid racemase ($k = 420$), UDP sugar hydrolase ($k = 354$), and S-adenosyl-L-homocysteine hydrolase ($k = 333$) were the three most highly connected hubs in the metabolic network. Strongly interconnected clusters where each node connects to each other node also emerged in the metabolic network (clustering coefficient $C_n = 0.322$), but not the drug network ($C_n = 0.015$). Methyl-transferase reactions ($n = 70$, $k = 200$) all utilizing the conversion between S-adenosyl-L-methionine and S-adenosyl-L-homocysteine dominate the largest MetaCyc cluster. The difference in connectivity cannot be attributed solely to the difference in network size between MDDR (nodes = 220, edges = 222) and MetaCyc (nodes = 5,039, edges = 21,604). We applied the same algorithm to the smaller metabolic network of MRSA (nodes = 554, edges = 803) revealing a pattern of connectivity ($k = 2.9$, mean = 2) and clustering ($C_n = 0.194$) similar to MetaCyc (**Figure A.2C**).

A second connectivity difference between the two networks is the presence of cycles. In contrast to our earlier work (Keiser, 2007), the network construction technique allows cycles when connections among highly similar sets would otherwise be lost. Surprisingly, the resulting drug network remains highly acyclic, with only seven cycles in total, all among four beta-lactam related drug classes (**Figure A.2B**). The beta-lactams break the otherwise treelike nature of the drug network by forming cycles. By contrast, the metabolic reaction networks are not at all treelike and have many cycles, which appeared widespread across the entire network (**Figures A.2A and A.2C**). Due to high cycle count, we could not fully count the cycles even in the smaller MRSA network. Enumerating cycles in a graph is an NP-complete problem.

Drug and metabolic chemical similarity networks follow distinctly different connectivity patterns, reflecting the ways each has been explored over time. Drug exploration, by medicinal chemistry and drug series growth, starts at multiple points and follows distinct branches of discrete, localized similarity. Even within a given biological effect category, it is rare for drug nodes to connect through more than one path, with four adenosine analog nodes as the only exception (**Figure A.2**). In stark contrast, a small number of highly connected central hubs emerge in metabolic networks, interconnected by redundant similarity links. One such central hub is the adenylate cyclase reaction that converts ATP to ADP, which reflects ATP's central role as a common metabolic currency. Explicit removal of metabolic common carriers decreases the size of central hubs, but does not remove them. Such a pattern is biologically sensible for at least two

reasons. First, metabolites must be recyclable. The breakdown product of one pathway becomes essential substrate for another. Reversible reactions and changing metabolic demand lead to an ebb and flow of substrates through catabolic and anabolic pathways. This vastly narrows the metabolic chemical space available to any single organism. Second, while metabolic chemical scaffolds are limited, the enzymes and proteins that bind them are prone to mutation and selection. Therefore, compared to drugs, nature can sample the relatively smaller space of metabolism more densely. We can quantify these differences in the random chemical background fits. Random metabolic backgrounds for single organisms are on average 10 to 20 times more internally similar than drug ones, indicating smaller breadth. However, the standard deviations are lower than drug backgrounds, indicating that metabolic space is more densely sampled (**Figure A.3**). We note that the chemical diversity for all of MetaCyc, a compendium of metabolic reactions from over 900 species, approaches within 6-fold the chemical diversity of drug space more closely than the metabolic complement of any single organism. This suggests that nature can access broader chemical diversity through metabolism, but that evolution selects for a much narrower slice of chemical space.

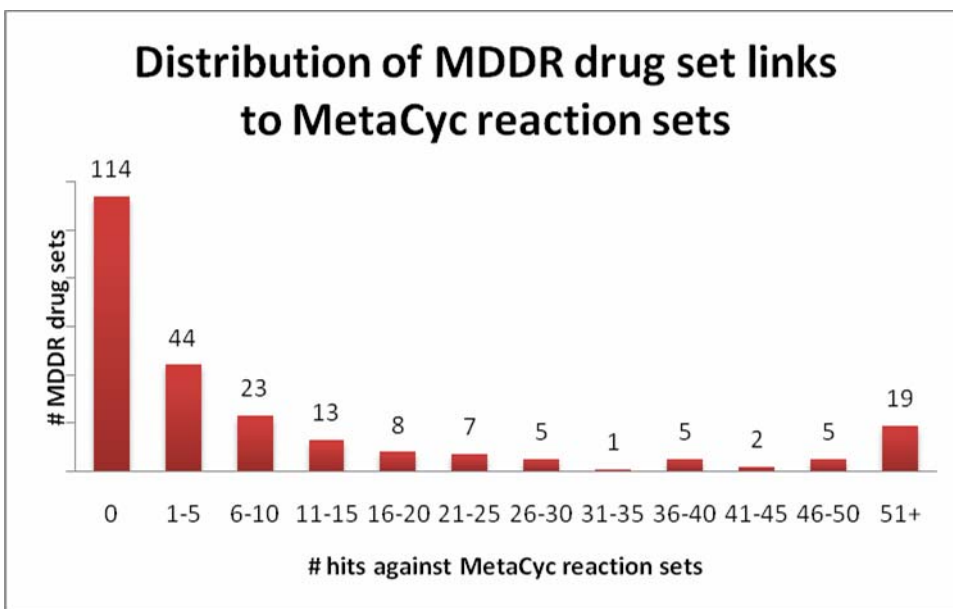


Figure A.1.A| Distribution of MDDR drug set links to MetaCyc reaction sets

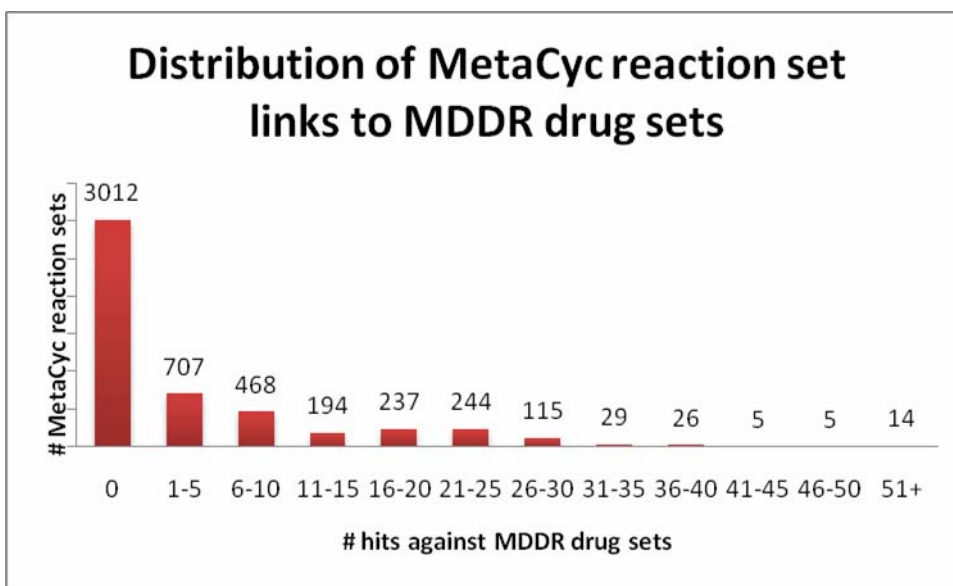


Figure A.1.B| Distribution of MetaCyc reaction set links to MDDR drug sets

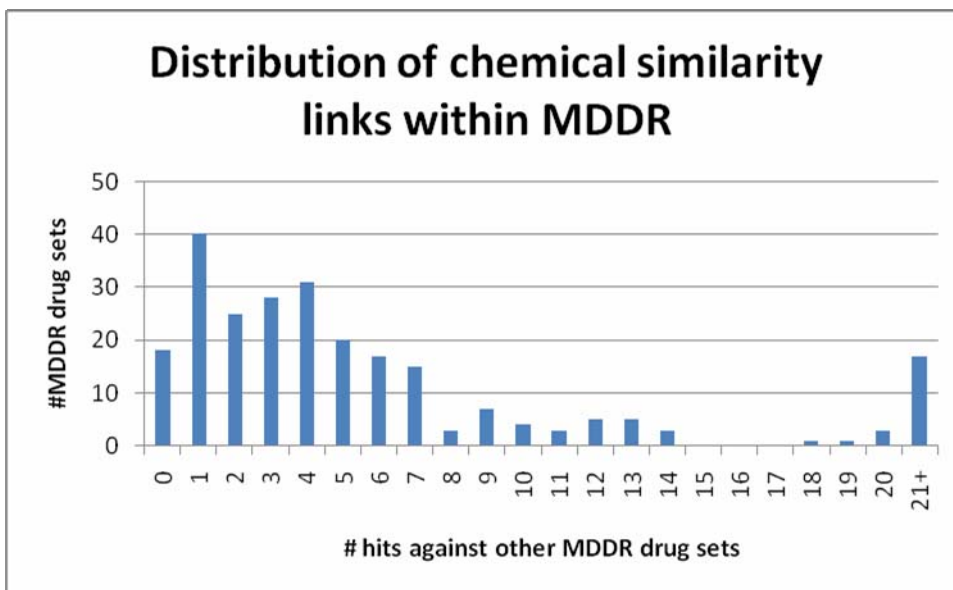


Figure A.1.C| Distribution of chemical similarity links within MDDR

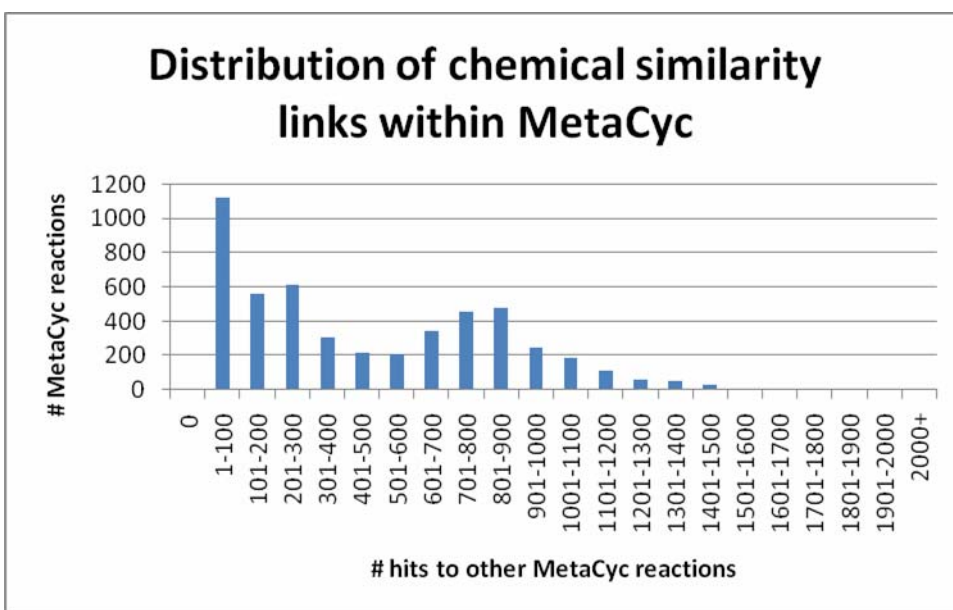


Figure A.1.D| Distribution of chemical similarity links within MetaCyc

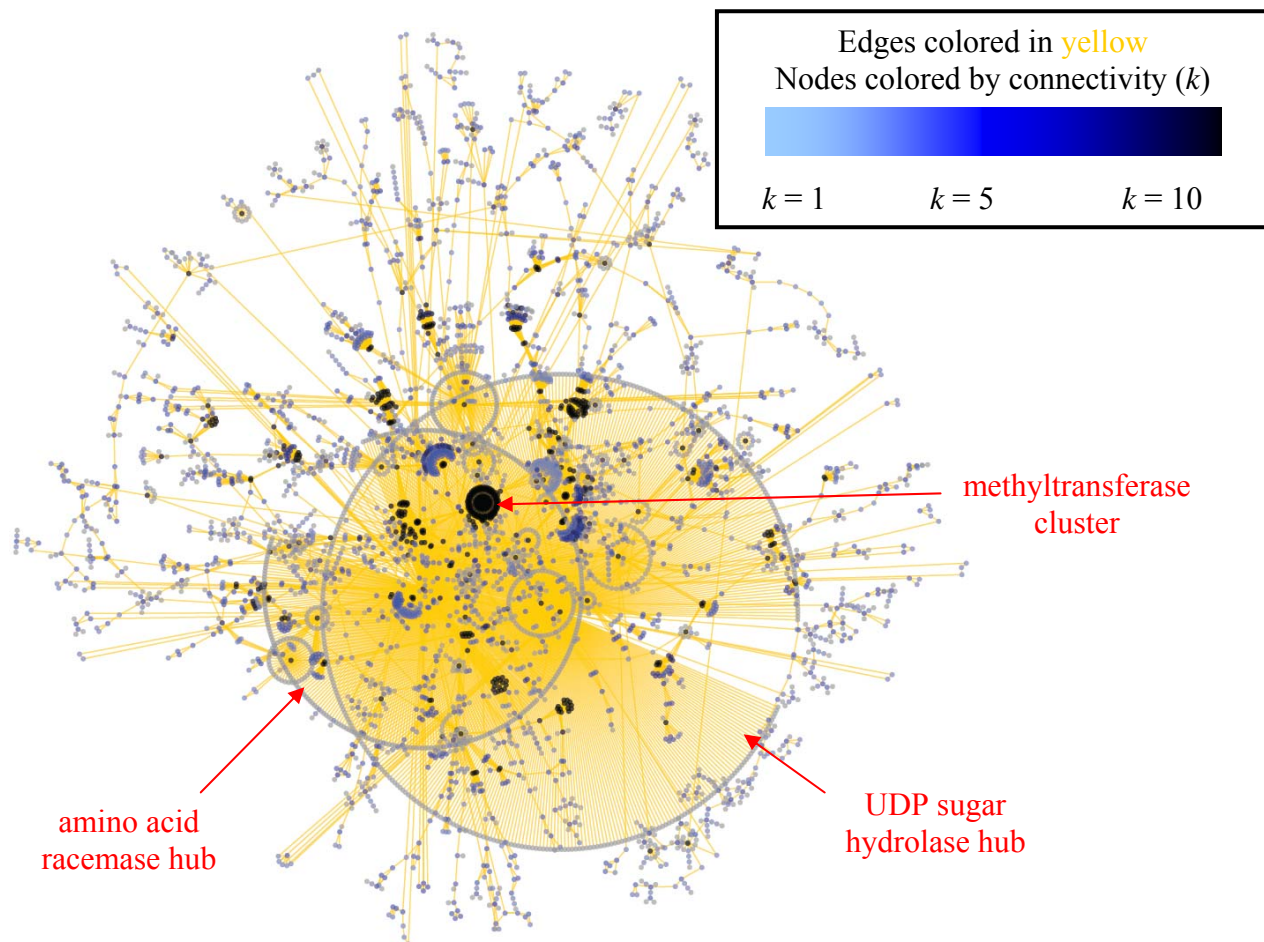


Figure A.2.A| MetaCyc metabolic network

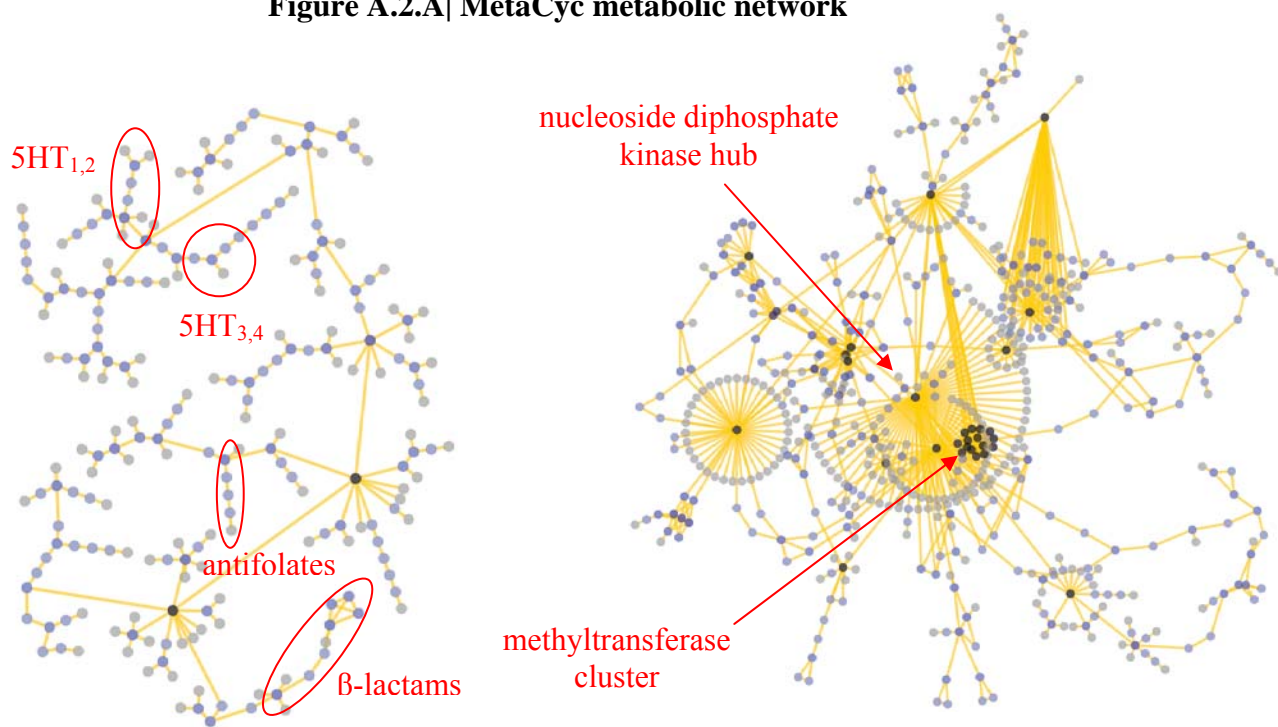


Figure A.2.B| MRSA metabolic

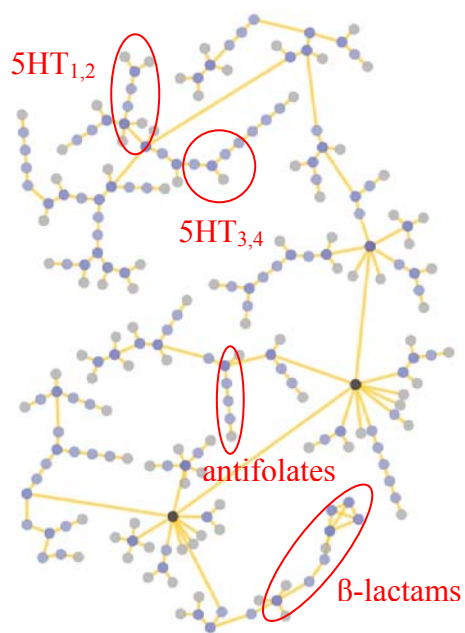


Figure A.2.C| MDDR drug network

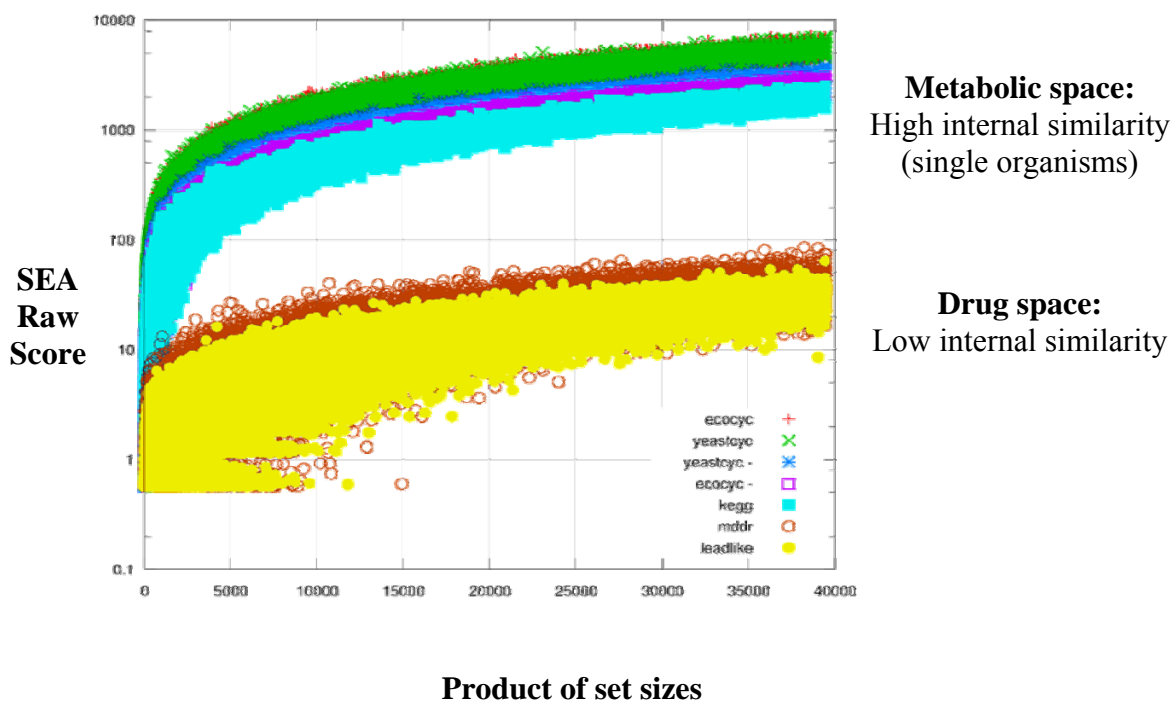


Figure A.3| Quantifying chemical diversity

Appendix B. MRSA Growth Assays

B.1 Rationale

Due to the dearth of essential MRSA reactions with strong links to MDDR, we also investigated synthetic lethal reactions that could be targeted in combination. Again using predictions from flux balance analysis, we mapped 19 synthetic lethal enzyme pairs to MRSA. All but one of the synthetic lethal pairs, aroenate and prephenate dehydrogenase in tyrosine precursor biosynthesis, had at least one orthologous enzyme also present in humans. Of all the potential targets without human orthologs, whether essential or synthetic lethal, aroenate dehydrogenase found the strongest links to MDDR drug sets with an expectation value of 4.80×10^{-28} . Interestingly, both aroenate and prephenate dehydrogenase are catalyzed by the *tyrA* enzyme in the same active site, making the *tyrA* essential for MRSA survival. Furthermore, aroenate dehydrogenase was also predicted to be synthetic lethal with aspartate aminotransferase (AAT). Based upon these SEA results, combined with the MRSA synthetic lethal analysis, we predicted that the synthetic lethal enzyme pair of *tyrA* and AAT would be the targets most accessible to current drug chemistry.

B.2 Experimental Design

To demonstrate proof of concept, we tested known inhibitors (m-fluorotyrosine for tyrA and aminooxyacetate for AAT) alone and in combination against the COL laboratory strain of MRSA. We used a 1:2 serial dilution growth assay in a microtiter plate format. AOAA alone inhibited MRSA growth at an IC_{50} of 340 μmol while MFT alone failed to inhibit at concentrations up to 10,000 μmol (**Figure B.1**). Surprisingly, while combining AOAA with elevated levels MFT at ratios up to 1:16 did not significantly alter MRSA growth rates, combining AOAA with less MFT at a ratio of 1:2 lowered the IC_{50} of AOAA to 206 μmol . These results confirm interaction between the AOAA, MFT, and MRSA metabolism.

B.3 Discussion

This work presented in Chapter 2 lays the computational foundation for ligand-based prediction of interactions between drug compounds and metabolic enzymes. However, prediction of the dynamic response of a pathogen or patient to a given therapeutic intervention lies beyond the scope of our method. The MRSA growth inhibition results reported here demonstrate the often counter-intuitive effects of perturbing metabolic networks. While AOAA combined with low levels of MFT effectively inhibited MRSA growth, combining AOAA with higher ratios of MFT did not significantly alter MRSA growth rates. These curious results may be rationalized by the fact that some tyrosine analogs function as feedback inhibitors. In some bacterial strains,

MFT down-regulates *de novo* synthesis of aromatic amino acids while also decreasing amino acid turnover, leading to an overall increase in the levels of aromatic and other amino acids. Thus, elevated levels of MFT may paradoxically lead to greater MRSA survival. Complimentary methods embraced by the emerging field of systems biology directly address this challenge of modeling the dynamic response of biological systems to chemical perturbations.

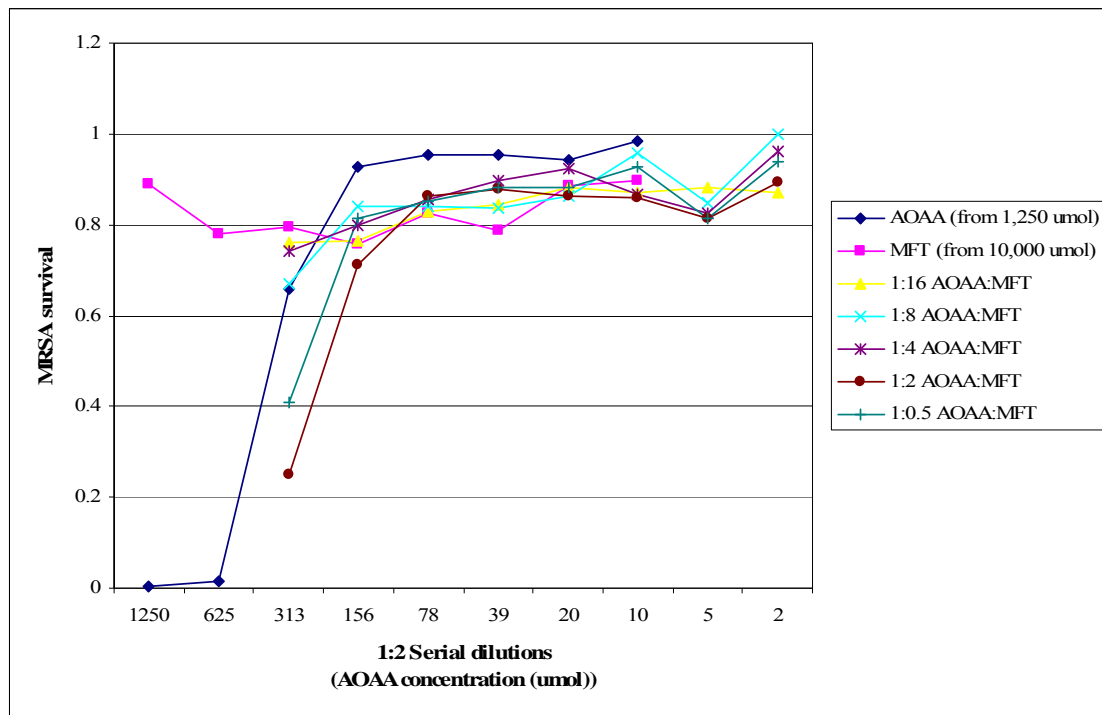


Figure B. 1| MRSA growth inhibition by AOAA and MFT. The COL strain of methicillin-resistant *Staphylococcus aureus* was incubated for 24hrs in the presence of aminooxyacetate (AOAA), m-fluorotyrosine (MFT), and AOAA/MFT combinations at varying ratios using a 1:2 serial dilution format on microtiter plates.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

James Corey Adams 6/11/09

Author Signature Date

(This page must be signed and dated by the author and include the correct pagination – the last page number of your document.)