USING EVOLUTIONARY STATISTICS TO UNDERSTAND CELLULAR SYSTEMS

APPROVED BY SUPERVISORY COMMITTEE

Kimberly Reynolds, Ph.D. (Advisor)

Milo Lin, Ph.D. (Chair)

Michael Reese, Ph.D.

Benjamin Tu, Ph.D.

DEDICATION

To my parents and my significant other Cindy Xu

USING EVOLUTIONARY STATISTICS TO UNDERSTAND CELLULAR SYSTEMS

by

ANDREW FRANK SCHOBER

DISSERTATION / THESIS

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

December, 2019

USING EVOLUTIONARY STATISTICS TO UNDERSTAND CELLULAR SYSTEMS

Publication No. 2

Andrew Frank Schober, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, Graduation Year

Supervising Professor: Kimberly Reynolds Ph.D.

Metabolic enzyme function is dependent on the larger context of a biochemical pathway. Despite detailed characterization of the requisite molecular "parts," it remains difficult to predict the adaptive response to a simple perturbation. That is: if the activity or expression of a single enzyme is changed, what other proteins (if any) require compensatory mutation? Comparative genomics and experimental evolution provide two powerful approaches to begin addressing these questions. In my thesis work, I examined adaptive interactions with the essential enzyme dihydrofolate reductase (DHFR). Analyses of gene synteny and co-occurrence across 1445 bacterial genomes indicated that DHFR coevolves with thymidylate synthase (TYMS), but is relatively decoupled from the rest of the folate metabolic pathway

(and genome). Through directed evolution of *E. coli*, I demonstrated that these two enzymes adapt cooperatively in response to antibiotic stress. An allele replacement experiment confirmed that a pair of mutations to DHFR and TYMS were sufficient to reconstitute the entire trimethoprim resistance phenotype, establishing that the two enzymes are capable of independently driving adaptation. In the final component of my thesis, I drew on the 'mirror-tree' method to define a new measure of residue-residue coevolution which corrects for the phylogenetic relationship among species. In summary, my results verify that small groups of genes within larger metabolic pathways can form adaptive modules that evolve as a unit in response to environmental or mutational stress. Moreover, my mirror-tree inspired analysis provides a path forward for understanding how coupled adaptation between genes manifests at the resolution of site specific constraints on the protein sequence.

Table of Contents

PRIOR PUBLICATIONS

Schober, A.F., et al., *A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism.* Cell Rep, 2019. **27**(11): p. 3359-3370.e7.

List of Figures

List of Tables

CHAPTER ONE

Introduction

1.1 Mapping adaptive interactions in central metabolism

Central metabolism results from the collective action of many enzymes. For model organisms like *E. coli*, prior work has enumerated many of the biochemical reactions catalyzed and how they are assembled to produce functioning biochemical pathways [2]. Large-scale collection of genome sequences has now provided metabolic "parts lists" for many other organisms [3], enabling genome-scale metabolic network reconstructions for a diversity of species [4-6]. Despite these efforts, it remains difficult to predict the fitness consequences of single point mutations or even gene knockouts [7, 8]. It is further complicated to understand how these systems adapt in response to environmental change or stress. For example: if the activity or expression of a single enzyme is changed, what other proteins (if any) require compensatory mutation? Such adaptive interactions are derived from the functional dependence of individual enzymes on their greater genetic context. Our ability to predict the effect of perturbations [9, 10], quantify the relationship between mutation and disease [11, 12], or rationally engineer new metabolic systems [13-15] is limited by an unknown pattern of functional coupling. Thus, an ability to generally map the adaptive interactions between enzymes would aid in focusing mechanistic work, suggest new strategies for the engineering of novel cellular systems, and provide a path toward the predictive modeling of cellular phenotypes.

1.2 Predicting adaptive interactions through coevolutionary inference

Comparative genomic analyses provide a general strategy for inferring the adaptive couplings between proteins that have shaped the evolution of cellular systems. The basic premise is that nature has already conducted a long-term experiment of perturbation and adaptation, which is recorded in the genomes of extant species. In these models, conservation is taken as an indicator of functional importance, while correlation is regarded as a signal of coevolution. Because coevolutionary analyses only report on the outcomes of selection, they can reveal interactions across diverse scales and mechanisms [16-20]. Statistical coevolution has been modeled in the context of both single protein families (residue-residue interaction) and entire genomes (gene-gene interaction) [20]. We propose that coevolving proteins, as determined by statistical analyses, represent core, conserved adaptive interactions. As such, these methods provide computational hypotheses which motivate direct testing in individual organisms through targeted perturbation and directed evolution experiments. The central goal of this thesis is to produce a computational and experimental roadmap for one might quantitatively evaluate our capacity to predict adaptive interactions through comparative genomics. Hereafter, I will provide a brief review of comparative genomic methods and their associated biological ramifications.

Analyses of coevolution can be roughly divided up into three categories, in order of decreasing granularity:

1. Co-occurrence: the correlated loss and gain of genes across species
2. Synteny: the conservation of physical proximity on the chromosome

3. Sequence coevolution: correlated changes in amino acid identity between positions of the protein sequence

Each approach has been shown to successfully associate gene products with known interaction [18, 19, 21]. However, it is non-obvious that the three different approaches would capture the exact same set of interactions when optimized. How effectively different mechanisms of interaction are imprinted on these evolutionary reporters is not well understood. As such, there is value in rigorous development and examination of each method in parallel.

Gene co-occurrence, sometimes called phylogenetic profiling, represents the coarsest measure of coevolution [21]. The biological reasoning that motivates this approach is that once a key component of a pathway or physical complex is lost, selection would dictate that the cells also dispense of the remaining no-longer-functional genes. Conversely, horizontal transfer represents a mechanism by which collections of functionally related genes can be passed from one organism to the next [22]. Acquiring a new phenotype can be dependent on the joint function of two or more genes. Therefore, organisms receiving only part of an enzymatic pathway would likely be outcompeted by those with the complete set of necessary genes. These events are tracked in an indirect manner through an analysis of gene co-occurrence across extant species. Each family of orthologous genes is represented by its ‘phylogenetic profile,’ a one-dimensional binary vector indicating whether that gene is present across a collection of species. Gene families can then be clustered based on their phylogenetic profile by a variety of statistical metrics [21, 23, 24].

A related strategy for the inference of evolutionary coupling is the statistical analysis of gene synteny, which refers to the conservation of genetic context. The arrangement of genes on a bacterial chromosome is a highly conserved genetic feature [25, 26]. The selective underpinnings of this result come from the informational processes of DNA replication and gene expression. Replication asymmetry drives the positioning of highly expressed genes near the origin, facilitating increased copy number during periods of fast growth [27]. Essential genes demonstrate a bias for the leading strand as to avoid head-on collisions between DNA and RNA polymerases [28]. The selective focus of synteny is gene co-expression. One-dimensional distance between genes has been shown to be the strongest determinant of co-expression in several bacteria [18]. This property is thought to have driven the formation of operons and supraoperons containing functionally related genes [18, 22, 29, 30].

The most stringent notion of synteny is called gene order conservation (GOC). As its name suggests, the method is concerned with genes that are directly adjacent on the chromosome. The size of intergenic regions is neglected, which means it is only practically applicable to prokaryotes. GOC is defined as the relative frequency that two contiguous genes have their respective orthologs contiguous in another species. After 500 million years of evolution, about 50% of genes that are initially contiguous will remain so for the average bacterium [26]. This indicates strong purifying selection against deleterious rearrangements. Analysis of GOC show that the data is well-described by an empirical model in which gene pairs fall into two categories: fast rearranging and slow rearranging [26]. The pattern of GOC

decay over evolutionary time is thus consistent with the case where selection on the proximity of specific families of orthologs is conserved across many species.

A more general definition of synteny considers the conservation of genes within kilobase (kb) sized segments of the chromosome. Recent work examined synteny as the number of times two genes occurred within a normalized distance threshold on the chromosome. [24]. The frequency of this event was then compared against a null expectation based on a model where genes are randomly and uniformly shuffled across species. The biological rationale of this approach stems from the observation that *E. coli* and *B. subtilis* genes within 10kb of one another uniformly co-express [18]. The vast majority of gene pairs in *E. coli* are well described by the null model, despite its simplicity. Specific pairs of genes show statistically significant coevolution according to gene synteny. These have been shown to be enriched for proteins that are known to physically interact as well as enzymes with a shared metabolic intermediate.

Lastly, perhaps the most challenging domain of comparative genomics is the analysis of amino acid sequence coevolution. In this class of methods, protein families are represented by their multiple sequence alignment (MSA). The dimensionality of this data is much greater than that of a simple binary indicator or gene-gene distance. The problem is further complicated by the fact that the strength of selection varies widely across the length of the protein [17]. Due to the factors of increased dimensionality and variable selection, this data is more susceptible to noise and the presence of a confounding phylogenetic signal [31]. Despite these challenges, various methods have been developed for the analysis of individual

protein families [16, 20, 32]. The output of such analyses is a map of coevolution between pairs of positions in a given multiple sequence alignment.

Statistical coupling analysis (SCA) is a framework that identifies coevolving networks of amino acids in the three dimensional protein structure [33]. This type of network, termed a 'sector,' has been implicated in the mediation of allostery, substrate specificity, and other facets of protein function [32, 34-36]. However, application of SCA to protein-protein coevolution is nontrivial. Work in one superfamily of multi-domain proteins suggests that SCA may be informative beyond the context of single domains, but the generality of this finding is yet unknown [37]. Critically, SCA lacks a natural definition for an 'interaction score' between a pair of protein families. On the other side of the coin is direct coupling analysis (DCA), which uses coevolution to infer the physical contact map of a protein [16]. In contrast to the interconnected networks described by SCA, DCA identifies a sparse pattern of physically contacting residue-residue pairs smattered across the three dimensional structure. These DCA contacts have proved powerful in three-dimensional structure prediction [16]. DCA has also been shown to detect contacts between proteins, namely the physical interfaces within a macromolecular complex [38, 39]. However, DCA is limited in its ability to predict protein-protein interactions *a priori*. This is due to a combination of computational expense and limited sensitivity[40]. As a result, physical interface prediction using DCA is either restricted to small datasets or aided by a prior screening via synteny or co-occurrence [39]. Because DCA identifies interaction based on the coevolutionary signature at a physical interface, the method would not be able to capture the full range of adaptive dependencies present in central metabolism.

A contrasting model of protein sequence coevolution, named ‘mirror-tree,’ is expressly purposed for the prediction of adaptive interaction between proteins. Instead of analyzing correlations in amino acid identity between specific positions of a multiple sequence alignment, mirror-tree tracks the change in total sequence similarity across species [19]. This feature allows the analysis to explicitly account for the phylogenetic relationship of each species when determining protein-protein coevolution [41, 42]. Similar to synteny and co-occurrence, this method culminates with a coarse grained ‘interaction-score’ consisting of a single number indicating the strength of coevolution between two proteins. While it was initially conceived as a method for identifying physically-interacting proteins, some evidence suggests that the application of mirror-tree could be more general [19, 43]. The tradeoff is that unlike SCA or DCA, the existing mirror-tree analyses provides no insight into which positions in the protein structure drive coevolution. So at present, successfully predicting protein-protein interaction using sequence information comes at the sacrifice of positional resolution. Developing a framework unifying the analysis of residue-residue coevolution with the prediction of evolutionary interactions is an open problem, which I explore later in my thesis.

Overall the tools of comparative genomics, consisting of co-occurrence, synteny, and amino acid sequence coevolution, provide a general strategy for the inference of adaptive interaction between proteins. While previous work has largely focused on using these tools to annotate protein function and predict physical interaction [1, 39, 40], I propose that the *absence* of statistical coevolution between most proteins can be taken as a prediction of adaptive independence. Intriguingly, genome wide analyses of synteny and co-occurrence

across bacteria have identified quasi-modular groups of genes which are evolutionarily coupled to one another but less so to the remainder of their cellular system and genome [18, 23]. These results suggest the possibility that cellular systems might be decomposed into smaller adaptive units, consisting of a few genes that evolve together in response to environmental stress or mutation. In my thesis work, I investigated this possibility using *E. coli* folate metabolism as a model system. My results provide a proof-of-concept that coevolutionary analysis might be used to identify small adaptive units embedded in larger cellular systems.

1.3 Folate metabolism as a model system

To experimentally test the idea that evolutionary couplings can be used to identify adaptive interactions in a given organism, I focused on the metabolic enzyme Dihydrofolate Reductase (DHFR, encoded by the *E. coli* gene *folA*). DHFR is essential for the synthesis of purine nucleotides, thymidine, and several amino acids [44]. As a consequence, it is a common target for antibiotics, antimalarials, and chemotherapeutics [45, 46]. DHFR has become a prominent model system for studying evolution due to its metabolic centrality, and our capacity to perturb its function through small molecule inhibitors. Recent work has utilized DHFR as a vehicle for understanding the evolution of drug resistance [47-51], protein conformational dynamics [52-54], and the evolutionary constraints on horizontal gene transfer [55, 56]. However, it remains unclear how the relationship between DHFR and cellular fitness is dependent on its greater metabolic context. Understanding how mutations to other enzymes might be able to compensate for a reduction in DHFR function is important

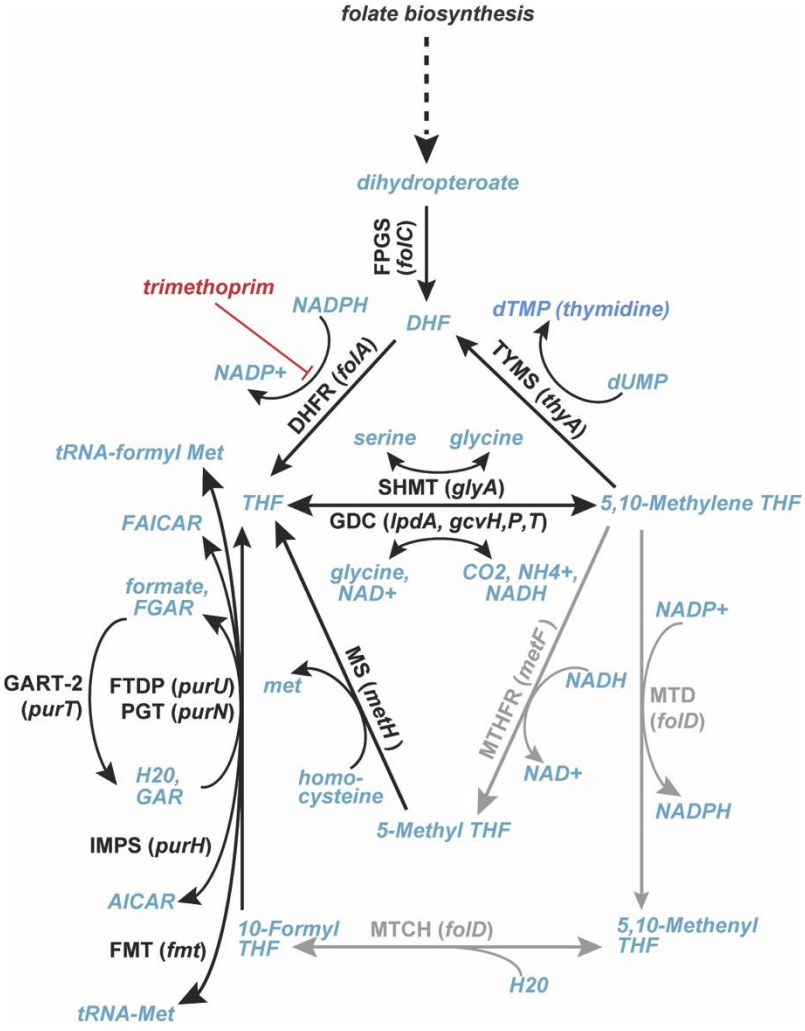

Figure 1.1 Biochemical map of folate metabolism. Abbreviated enzyme names are displayed in black or grey text. Black text and lines correspond to enzymes that were annotated as highest confidence interactions with DHFR (*folA*) in STRINGdb v10.5 [1]. Metabolites are indicated in blue. See abbreviations for complete names of each enzyme and compound.

to questions of antibiotic resistance and more generally the evolution of folate metabolism. As such, DHFR provided a well-studied system to test the capacity for comparative genomics to predict adaptive interactions.

To map the network of enzymes that have the potential to adaptively interact with DHFR, I used STRINGdb. STRING is a database that integrates many types of biological

data, including gene synteny, co-occurrence, co-expression, and high-throughput experiments (e.g. yeast two-hybrid). Its stated goal is to uncover "both direct and indirect functional interactions" between proteins. The amalgamated STRING score is benchmarked against its ability to recapitulate KEGG pathways. The STRING database (v10.5, [1]) and KEGG pathway maps [3] indicate 16 core enzymes that complete the one-carbon cycle and are biochemically coupled to DHFR (e.g. they are linked by a product or substrate, Figure 1.1). The complete folate metabolic pathway, as identified by STRING, interconverts folic acid derivatives through a series of one-carbon group transfers. Methionine, serine, thymidine, and purine nucleotides are produced in the process [44]. The input of the pathway is 7,8-dihydrofolate (DHF), which is produced through the addition of L-glutamate to dihydropteroate by the bifunctional enzyme dihydrofolate synthase/folylpolyglutamate synthase (FPGS). Dihydrofolate is not active in one carbon metabolism so it must first be reduced by DHFR. DHFR catalyzes the stereospecific conversion of DHF to 5,6,7,8-tetrahydrofolate (THF) in an NAPDH-dependent manner. THF is then modified at both the N-5 and N-10 positions by a diversity of one-carbon groups. Various THF species serve as a one-carbon donor for the synthesis of amino acids and purine nucleotides (Figure 1.1). Oxidation of reduced folate (5,10-Methylene THF) back to DHF is exclusively carried out by thymidylate synthase (TYMS), which concurrently convers uridine monophosphate (dUMP) to thymidine monophosphate (dTMP).

Like many microorganisms, *E. coli* lack the ability to transport folate into the cell. As a result, the folate metabolic pathway is the target of several well-known antibiotics. These include sulfamethoxazole (SMX) and trimethoprim (TMP), which are commonly prescribed

in combination [57]. Sulfamethoxazole impedes the production of DHF through the inhibition of FPGS activity, while trimethoprim competitively binds DHFR. Together, these compounds limit the production of reduced folates, resulting in adverse changes to metabolite abundances and reaction velocities. The final outcome of trimethoprim inhibition on metabolism and bacterial growth is dependent, in part, on environmental conditions.

In a nutrient poor context (e.g. media lacking amino acids), trimethoprim addition causes cell stasis. Mass spectrometry profiling of intracellular metabolite concentrations in *E. coli* has detailed the underlying cascade of metabolic events [58, 59]. The most immediate consequence of DHFR inhibition is the accumulation of its substrate DHF. At high concentrations, DHF exhibits product inhibition of upstream FPGS [59]. In other organisms, this metabolite has also been shown to inhibit the activities of MTHFR and TYMS [60-62]. Reduced flux through the pathway leads to the hierarchical depletion of folate-dependent metabolites [58]. Intracellular glycine depreciates on the fastest timescale, accounting for an almost immediate halt in growth. This is followed by the simultaneous depletion of thymidine (dTTP) and Methionine, followed by adenosine triphosphate (ATP) sometime later. Related cofactors of the folate pathway, such as AICAR and dUMP, accumulate concurrently. The fastest starvation event, the depletion of glycine, stops growth through activation of the stringent response [63]. Long term stasis is stabilized by the purine deficiency which develops sometime later.

In a nutrient rich environment containing exogenous amino acids, trimethoprim inhibition primarily results in cell death. Since glycine starvation and thus the stringent response are abated, the depletion of thymidine constitutes the fastest and most dominant

effect in the cascade. Thymine starvation has been shown to induce the premature initiation of DNA replication [27]. While the ensuing chain of cause and effect are not completely understood, this has lethal consequences due to double-stranded breaks, single strand gaps, and recombination intermediates at the origin [64]. The localized DNA breakage eventually results in degradation of the origin of replication (*oriC*). This phenomenon is referred to as *thymineless death* [65].

Given that trimethoprim has pleiotropic metabolic ramifications spanning the entire folate metabolic pathway, it presents a non-trivial yet well-defined opportunity to study how complex cellular systems adapt to targeted perturbation. That is, which enzymes adapt in response to inhibition of DHFR with trimethoprim? As a first step towards this question, we examined the extent to which the biochemical connections between folate metabolic enzymes lead to co-evolution across species, using comparative genomics.

1.4 Coevolution in the folate pathway

We analyzed gene synteny and co-occurrence across 1445 completely sequenced bacterial genomes to study the pattern of evolutionary coupling between the 16 core enzymes of folate metabolism [24]. As a null model, the amalgamated STRING scores an associated pathway map of folate metabolism suggested a dense pattern of biochemical interactions in which most enzymes of the pathway are coupled to one another. In contrast, synteny and co-occurrence indicate that evolutionary coupling in the pathway is both sparse and modular (Figure 1.2). The maps of coevolution produced by synteny and co-occurrence are qualitatively consistent with one another. Most pairs of enzymes do not show any statistical

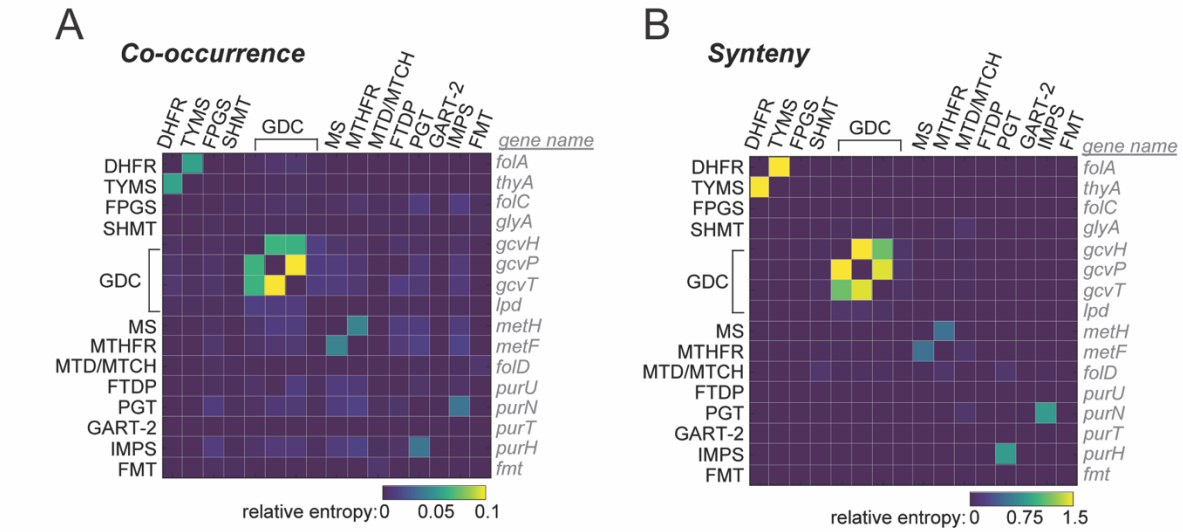

Figure 1.2 Coevolutionary maps of folate metabolism. A-B Statistical coevolution according to analyses of gene co-occurrence and synteny as computed across 1445 complete bacterial genomes. Coupling between gene pairs in folate metabolism is indicated as a relative entropy D_{ij}^{intra}, shown by pixel intensity. Enzyme names are labeled on the top and left of each heatmap, with corresponding gene names in grey italics along the right. In *E. coli*, a single gene (*folD*) encodes a bifunctional enzyme which catalyzes both the methylene tetrahydrofolate dehydrogenase (MTD) and methenyltetrahydrofolate cyclohydrolase (MTCH) reactions as shown by the biochemical pathway in Figure (1.1).

coevolution, despite their apparent proximity in biochemical space. We observed several small groups of enzymes which demonstrated internal coevolution but remained statistically independent from the rest of the pathway. As one might expect, one such evolutionary unit consisted of the glycine cleavage system proteins H, P, and T (*gcvH*, *gcvP*, and *gcvT* in *E. coli*). These gene products comprise a single macromolecular complex, which facilitates the conversion of glycine to serine [66]. Modular coevolution is also observed between three pairs of biochemically related enzymes: 1) DHFR and TYMS 2) methionine synthase (MS) with methionine tetrahydrofolate reductase (MTHFR) and 3) the purine biosynthesis proteins PGT and IMPS. These interactions are likely mediated by a biochemical mechanism, since the proteins are not known to physically bind.

There are a few technical caveats relevant to the hypothesis of adaptive independence that is motivated by our statistical results. First, some false negatives are expected due to limited statistical power. These would constitute enzyme pairs that do in fact coevolve, but that signal is not recovered by our present analysis. We find only six evolutionary couplings out of 120 enzyme pairs, despite the fact that 83 enzyme pairs are coupled biochemically through a shared product or substrate. Therefore, a high false negative rate would necessary to explain the sparsity observed in our statistical maps of coevolution. More generally, the pattern of adaptive interactions in a pathway need not strictly resemble its biochemical map given the non-linear relationship between enzyme activities, metabolite concentrations, and fitness. Thus, the observed sparsity may reflect a modular organization of adaptive constraints that have been conserved throughout evolution. Since our results are the product of statistical inference across thousands of genomes, they are not expected to capture idiosyncratic interactions that are specific to a particular organism or environmental condition. We hypothesize that coevolutionary maps represent a prediction of the core, conserved adaptive couplings and their degree of independence from the rest of the genome. If true, these would provide for identifying the evolutionary building blocks of metabolic pathways, consisting of smaller, multi-gene adaptive units, which do not depend strongly on the choice of model organism or environment.

References

1. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
2. Keseler, I.M., et al., *EcoCyc: a comprehensive view of Escherichia coli biology.* Nucleic Acids Res, 2009. **37**(Database issue): p. D464-70.
3. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
4. Ma, H. and A.P. Zeng, *Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms.* Bioinformatics, 2003. **19**(2): p. 270-7.
5. Baart, G.J. and D.E. Martens, *Genome-scale metabolic models: reconstruction and analysis.* Methods Mol Biol, 2012. **799**: p. 107-26.
6. Fondi, M. and P. Lio, *Genome-scale metabolic network reconstruction.* Methods Mol Biol, 2015. **1231**: p. 233-56.
7. Tang, H. and P.D. Thomas, *Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation.* Genetics, 2016. **203**(2): p. 635-47.
8. Peleg, T., et al., *Network-Free Inference of Knockout Effects in Yeast.* PLOS Computational Biology, 2010. **6**(1): p. e1000635.
9. Kim, J., et al., *Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis.* Mol Syst Biol, 2010. **6**: p. 436.
10. Long, C.P., et al., *Dissecting the genetic and metabolic mechanisms of adaptation to the knockout of a major metabolic enzyme in Escherichia coli.* Proc Natl Acad Sci U S A, 2018. **115**(1): p. 222-227.
11. Kondrashov, A.S., S. Sunyaev, and F.A. Kondrashov, *Dobzhansky-Muller incompatibilities in protein evolution.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 14878-83.
12. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1193-8.
13. Kim, J. and S.D. Copley, *Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network.* Proc Natl Acad Sci U S A, 2012. **109**(42): p. E2856-64.
14. Michener, J.K., et al., *Effective use of a horizontally-transferred pathway for dichloromethane catabolism requires post-transfer refinement.* Elife, 2014. **3**.
15. Michener, J.K., et al., *Phylogeny poorly predicts the utility of a challenging horizontally transferred gene in Methylobacterium strains.* J Bacteriol, 2014. **196**(11): p. 2101-7.
16. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293-301.
17. Rivoire, O., K.A. Reynolds, and R. Ranganathan, *Evolution-Based Functional Decomposition of Proteins.* PLoS Comput Biol, 2016. **12**(6): p. e1004817.

18. Junier, I. and O. Rivoire, *Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation.* PLoS One, 2016. **11**(5): p. e0155740.
19. Pazos, F. and A. Valencia, *Similarity of phylogenetic trees as indicator of protein-protein interaction.* Protein Eng, 2001. **14**(9): p. 609-14.
20. de Juan, D., F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution.* Nat Rev Genet, 2013. **14**(4): p. 249-61.
21. Pellegrini, M., et al., *Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4285-8.
22. Pang, T.Y. and M.J. Lercher, *Supra-operonic clusters of functionally related genes (SOCs) are a source of horizontal gene co-transfers.* Scientific reports, 2017. **7**: p. 40294-40294.
23. Rivoire, O., *Elements of coevolution in biological sequences.* Phys Rev Lett, 2013. **110**(17): p. 178102.
24. Schober, A.F., et al., *A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism.* Cell Rep, 2019. **27**(11): p. 3359-3370.e7.
25. Lathe, W.C., 3rd, B. Snel, and P. Bork, *Gene context conservation of a higher order than operons.* Trends Biochem Sci, 2000. **25**(10): p. 474-9.
26. Rocha, E.P., *Inference and analysis of the relative stability of bacterial chromosomes.* Mol Biol Evol, 2006. **23**(3): p. 513-22.
27. Pritchard, R.H. and K.G. Lark, *INDUCTION OF REPLICATION BY THYMINE STARVATION AT THE CHROMOSOME ORIGIN IN ESCHERICHIA COLI.* J Mol Biol, 1964. **9**: p. 288-307.
28. Rocha, E.P. and A. Danchin, *Essentiality, not expressiveness, drives gene-strand bias in bacteria.* Nature genetics, 2003. **34**(4): p. 377.
29. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* Journal of molecular biology, 1961. **3**(3): p. 318-356.
30. Snel, B., P. Bork, and M.A. Huynen, *The identification of functional modules from the genomic association of genes.* Proc Natl Acad Sci U S A, 2002. **99**(9): p. 5890-5.
31. Tesileanu, T., L.J. Colwell, and S. Leibler, *Protein sectors: statistical coupling analysis versus conservation.* PLoS Comput Biol, 2015. **11**(2): p. e1004091.
32. Halabi, N., et al., *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-86.
33. Reynolds, K.A., et al., *Evolution-based design of proteins.* Methods Enzymol, 2013. **523**: p. 213-35.
34. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.
35. Reynolds, K.A., R.N. McLaughlin, and R. Ranganathan, *Hot spots for allosteric regulation on protein surfaces.* Cell, 2011. **147**(7): p. 1564-75.
36. Raman, A.S., K.I. White, and R. Ranganathan, *Origins of Allostery and Evolvability in Proteins: A Case Study.* Cell, 2016. **166**(2): p. 468-480.
37. Smock, R.G., et al., *An interdomain sector mediating allostery in Hsp70 molecular chaperones.* Mol Syst Biol, 2010. **6**: p. 414.

38. Hopf, T.A., et al., *Sequence co-evolution gives 3D contacts and structures of protein complexes.* Elife, 2014. **3**.
39. Ovchinnikov, S., H. Kamisetty, and D. Baker, *Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information.* Elife, 2014. **3**: p. e02030.
40. Croce, G., et al., *A multi-scale coevolutionary approach to predict interactions between protein domains.* bioRxiv, 2019: p. 558379.
41. Sato, T., et al., *The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.* Bioinformatics, 2005. **21**(17): p. 3482-9.
42. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-92.
43. Clark, N.L., E. Alani, and C.F. Aquadro, *Evolutionary rate covariation reveals shared functionality and coexpression of genes.* Genome Res, 2012. **22**(4): p. 714-20.
44. Green, J.M. and R.G. Matthews, *Folate Biosynthesis, Reduction, and Polyglutamylation and the Interconversion of Folate Derivatives.* EcoSal Plus, 2013.
45. Ducker, G.S. and J.D. Rabinowitz, *One-Carbon Metabolism in Health and Disease.* Cell Metab, 2016.
46. Gangjee, A., H.D. Jain, and S. Kurup, *Recent advances in classical and non-classical antifolates as antitumor and antiopportunistic infection agents: part I.* Anticancer Agents Med Chem, 2007. **7**(5): p. 524-42.
47. Costanzo, M.S. and D.L. Hartl, *The evolutionary landscape of antifolate resistance in Plasmodium falciparum.* J Genet, 2011. **90**(2): p. 187-90.
48. Ogbunugafor, C.B., et al., *Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance.* PLoS Comput Biol, 2016. **12**(1): p. e1004710.
49. Palmer, A.C., et al., *Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes.* Nat Commun, 2015. **6**: p. 7385.
50. Rodrigues, J.V., et al., *Biophysical principles predict fitness landscapes of drug resistance.* Proc Natl Acad Sci U S A, 2016. **113**(11): p. E1470-8.
51. Toprak, E., et al., *Evolutionary paths to antibiotic resistance under dynamically sustained drug selection.* Nat Genet, 2012. **44**(1): p. 101-5.
52. Bhabha, G., et al., *Divergent evolution of protein conformational dynamics in dihydrofolate reductase.* Nat Struct Mol Biol, 2013. **20**(11): p. 1243-9.
53. Francis, K., V. Stojkovic, and A. Kohen, *Preservation of protein dynamics in dihydrofolate reductase evolution.* J Biol Chem, 2013. **288**(50): p. 35961-8.
54. Liu, C.T., et al., *Functional significance of evolving protein sequence in dihydrofolate reductase from bacteria to humans.* Proc Natl Acad Sci U S A, 2013. **110**(25): p. 10159-64.
55. Bershtein, S., et al., *Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria.* PLoS Genet, 2015. **11**(10): p. e1005612.
56. Bhattacharyya, S., et al., *Transient protein-protein interactions perturb E. coli metabolome and cause gene dosage toxicity.* Elife, 2016. **5**.

57. Reeves, D.S., et al., *Trimethoprim--sulphamethoxazole: comparative study in urinary infection in hospital.* Br Med J, 1969. **1**(5643): p. 541-4.
58. Kwon, Y.K., M.B. Higgins, and J.D. Rabinowitz, *Antifolate-induced depletion of intracellular glycine and purines inhibits thymineless death in E. coli.* ACS Chem Biol, 2010. **5**(8): p. 787-95.
59. Kwon, Y.K., et al., *A domino effect in antifolate drug action in Escherichia coli.* Nat Chem Biol, 2008. **4**(10): p. 602-8.
60. Dolnick, B.J. and Y.C. Cheng, *Human thymidylate synthetase. II. Derivatives of pteroylmono- and -polyglutamates as substrates and inhibitors.* J Biol Chem, 1978. **253**(10): p. 3563-7.
61. Kisliuk, R.L., Y. Gaumont, and C.M. Baugh, *Polyglutamyl derivatives of folate as substrates and inhibitors of thymidylate synthetase.* J Biol Chem, 1974. **249**(13): p. 4100-3.
62. Matthews, R.G. and C.M. Baugh, *Interactions of pig liver methylenetetrahydrofolate reductase with methylenetetrahydropteroylpolyglutamate substrates and with dihydropteroylpolyglutamate inhibitors.* Biochemistry, 1980. **19**(10): p. 2040-5.
63. Cashel, M., et al., *Escherichia coli and Salmonella: cellular and molecular biology.* 1996.
64. Guzman, E.C. and C.M. Martin, *Thymineless death, at the origin.* Front Microbiol, 2015. **6**: p. 499.
65. Bushby, S.R. and G.H. Hitchings, *Trimethoprim, a sulphonamide potentiator.* Br J Pharmacol Chemother, 1968. **33**(1): p. 72-90.
66. Okamura-Ikeda, K., et al., *Cloning and nucleotide sequence of the gcv operon encoding the Escherichia coli glycine-cleavage system.* Eur J Biochem, 1993. **216**(2): p. 539-48.

CHAPTER TWO

Directed Evolution of Trimethoprim Resistance in *E. coli*

2.1 Background and introduction

2.1.1 Harnessing evolutionary inference for hypothesis generation

Extant species are the product of a convolution of various selective pressures with a largely unknown (and possibly shifting) network of functional constraints between genetic elements. We attempt to learn the constraints that guide evolutionary outcomes through comparative genomic analyses. But how do these statistical results translate into meaningful and testable experimental hypotheses? As described in chapter 1, my colleagues and I applied analysis of two coevolutionary measures to bacterial folate metabolism (Fig. 1.2) [1]. A key finding of this analysis was that small groups of genes demonstrated a strong signature of coevolution while maintaining statistical independence from rest of the pathway. We termed these collections of 2-3 genes evolutionary modules. We hypothesized that genes within an evolutionary module would demonstrate two defining characteristics. Based on the presence of statistical coupling, we propose that the fitness effect of perturbing one gene in the module should be dependent on the functional state of the others. Conversely, we expect that genes within a module should be capable of co-adapting to relevant perturbing or change in environment in a way that is independent from the rest of the pathway; this would help explain our observation of statistical independence from the remainder of the genome.

To test this hypothesis, we chose an evolutionary module comprised of the essential metabolic enzymes dihydrofolate reductase (DHFR) and thymidylate synthase (TYMS) for experimental study. The two proteins are not known to physically interact. Rather, they catalyze sequential steps of the folate metabolic pathway, suggesting a possible biochemical mechanism of coupling. We chose *E. coli* as our model organism, where the function of DHFR is directly tied to growth rate [2]. Despite the fact that genes encoding DHFR and TYMS are in synteny in many bacterial species, they are several megabases apart in *E. coli*.

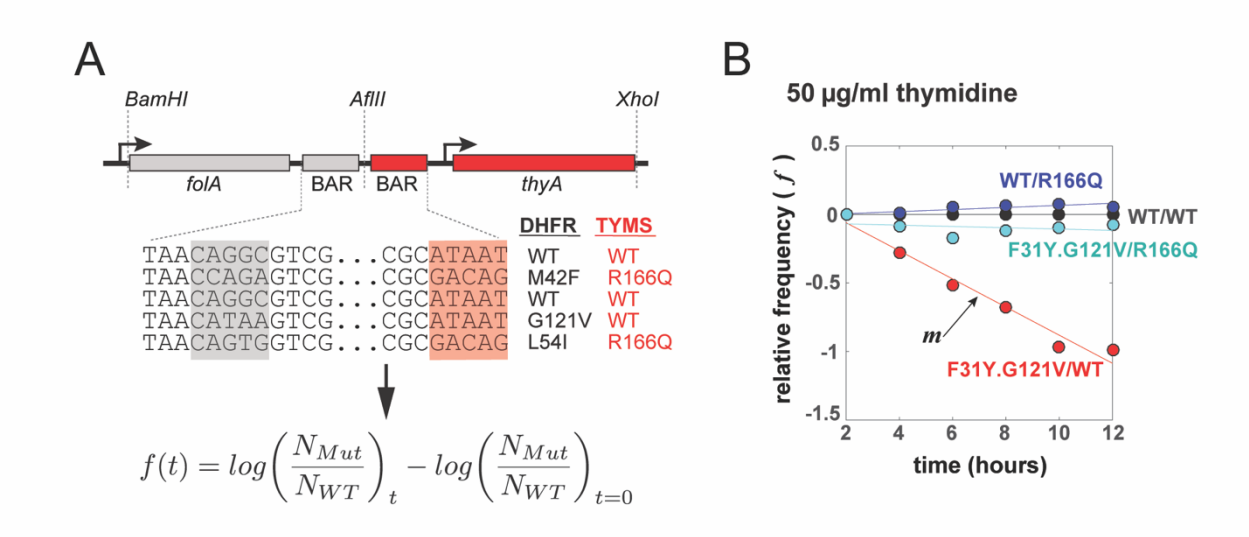

Figure 2.1 Growth competition assay for paired DHFR/TYMS mutants. A, Barcoding strategy for using deep sequencing to count genotype frequencies over time. The schematic represents a plasmid encoding a single copy of the *folA* (DHFR) and *thyA* (TYMS) genes. The library encapsulated 10 variants of *E. coli* DHFR with known catalytic activity, paired with either a wild-type (WT) or loss-of-function (R166Q) TYMS allele (20 genotypes in total). Each allele was labeled with a unique 5 nucleotide identifier denoted 'BAR' in the schematic. The plasmid library was transformed into *E. coli* ER2566 Δ*folA* Δ*thyA* for growth competition. The number of each barcode pair ($\boldsymbol{N_{mut}}$) relative to the reference genotype (WT/WT; counts denoted by $\boldsymbol{N_{WT}}$) is monitored over time through deep sequencing to provide a relative frequency $\boldsymbol{f(t)}$. **B,** Sample plot of log-relative frequency over time for select genotypes. Points represent relative frequencies estimated from next generation sequencing of culture samples collected over time. Data are color coded to match the genotype labels which are in DHFR/TYMS format. Lines represent a least squares fit of linear slope $\boldsymbol{m}$, which is equal to the exponential growth rate of each genotype relative to wild-type. The slope of the WT/WT line is therefore zero by construction. Relative growth rate provides a measure of fitness and indicates whether a genotype will become enriched in the population or deplete over time.

As a result, these experimental tests will also determine whether the functional coupling underlying gene synteny persists even in the absence of proximity on the chromosome. If the prediction that evolutionary modules adapt as a relatively independent unit proves to be true, then this work could provide a strategy for using evolutionary statistics to decompose cellular systems into smaller, adaptive subunits.

2.1.2 Using targeted perturbation to assay functional coupling between enzymes

The statistical coevolution observed between DHFR and TYMS suggests that fitness effect of mutating one of the enzymes should be tied to the functional state of the other. To test this prediction, we examined the fitness (as assessed by growth rate) of *E. coli* for an array of DHFR and TYMS genotype combinations. We utilized a set of 10 previously characterized *E. coli* DHFR mutants spanning a range of catalytic activities (k_{cat}/K_m). Each DHFR was paired with either a wild-type (WT) or catalytically dead (R166Q) TYMS, constituting 20 genotypes in total (Figure 2.1A). DHFR/TYMS pairs were expressed on a plasmid system wherein each genotype was labeled with a genetic barcode. By using next generation sequencing to count barcode frequency over time, we were able to culture all genotypes in a pooled relative growth rate assay (Figure 2.1B). Counts of frequency over time were converted into a relative fitness, which expresses the difference in exponential growth rate between each genotype and the reference (WT/WT). Growth rates were measured in M9 minimal media with 0.4% glucose, 0.1% amicase and either a partial or full rescue of R166Q TYMS through thymidine supplementation (5 or 50 ug/ml thymidine). In both TYMS backgrounds and experimental conditions, we observed that decreasing DHFR activity produces a monotonic decline in relative growth rate (Figure 2.2). In the 5 μg/ml

thymidine condition, introducing the R166Q TYMS mutation resulted in a fitness defect when paired with a wild-type DHFR, as expected (Figure 2.2A). However, the growth defect due to a decrease in DHFR activity was mitigated by the presence of the inactive TYMS mutant. When paired with the slowest DHFRs, the R166Q TYMS allele outperformed its wild-type. The fitness cost associated with an inactive R166Q TYMS was completely rescued by supplementation with 50 µg/ml thymidine. In this condition, R166Q TYMS always outgrew wild-type TYMS as DHFR activity was reduced (Figure 2.2B). These results confirm that selection for DHFR activity is directly coupled to the functional state of TYMS. In particular, we observed a pattern of buffering epistasis in which the cost of reducing

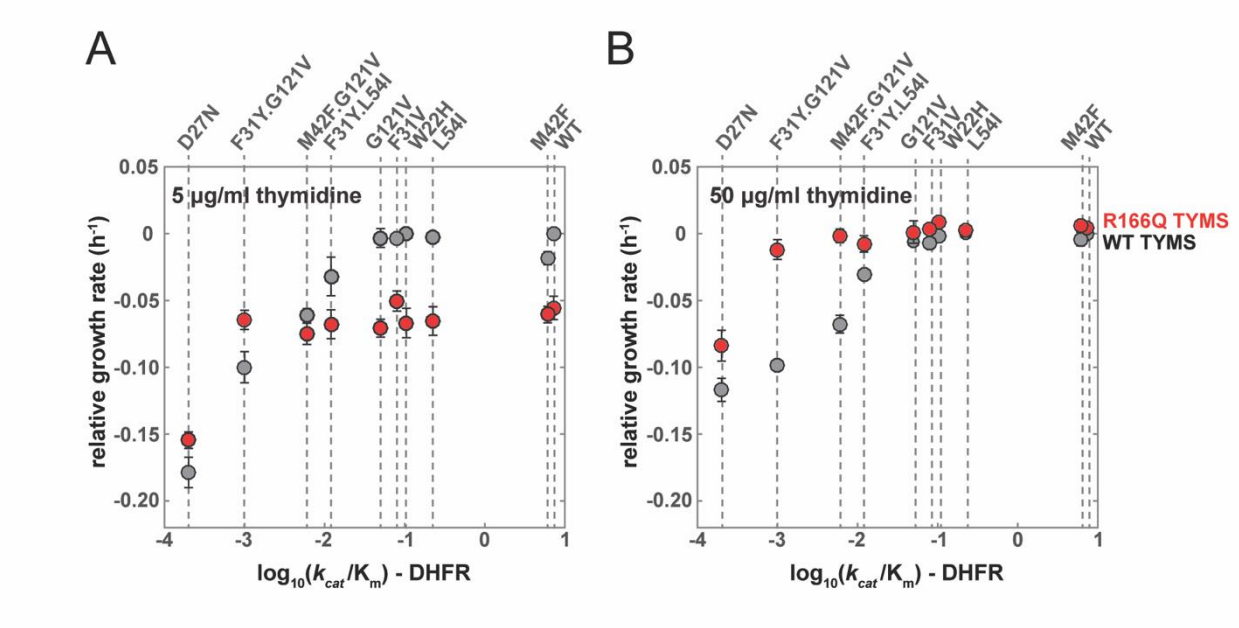

Figure 2.2 The fitness cost of decreased DHFR activity is buffered by a loss-of-function in TYMS. A,B Scatter plots of DHFR mutants spanning an array of catalytic activities (k_{cat}/K_m) when paired with either a wild-type (WT; grey dots) or catalytically dead (R166Q; red dots) TYMS. Relative growth rate (h^{-1}) is normalized against the WT/WT genotype; error bars denote standard error across triplicate measurements. The relative DHFR point mutants are indicated along the top of the plot. The assay was conducted in M9 and 0.1% amicase supplemented with either 5 or 50 µg/ml thymidine. Results from both conditions indicate that the R166Q TYMS mutation buffers the fitness cost of reducing DHFR activity.

DHFR activity can be compensated for with the introduction of a catalytically inactive TYMS. Our findings are consistent with a biochemical constraint wherein the relative activity of TYMS should not greatly exceed that of DHFR.

2.1.3 Metabolomic profiling reveals a constraint on the intermediate dihydrofolate

To better understand the biochemical constraints on DHFR and TYMS function, we characterized the metabolic changes that result from perturbation to these two enzymes. More specifically, we selected 10 DHFR/TYMS pairs from the above relative growth rate measurements for liquid chromatography-mass spectrometry (LC-MS) profiling of folate metabolites. Cells were harvested from log-phase growth in M9 glucose media supplemented with 0.1% amicase and 50 μg/ml thymidine. As shown by the growth competition assay, DHFR mutants in this condition display significant growth defects individually. However, the corresponding DHFR/TYMS double mutants are restored to near wild-type growth. Current mass spectrometry methods allow discernment between the full diversity of folate species, which vary in oxidation, one-carbon modification, and polyglutamylation state [3]. Thus, our approach permits broad study of the metabolic consequences resulting from mutation.

The data show that a reduction in DHFR activity alone causes an accumulation of its substrate, indicated by the increase in intracellular DHF (Figure 2.3A, bottom four rows). This effect is accompanied by a depletion of reduced polyglutamated folates (Glu $\geq$ 3), while a number of mono- and di-glutamated reduced folate species are increased. Prior work found that high concentrations of DHF generate product inhibition of its upstream enzyme FP-γ-GS. In addition to the synthesis of DHF, FP-γ-GS is responsible for the catalysis of reduced

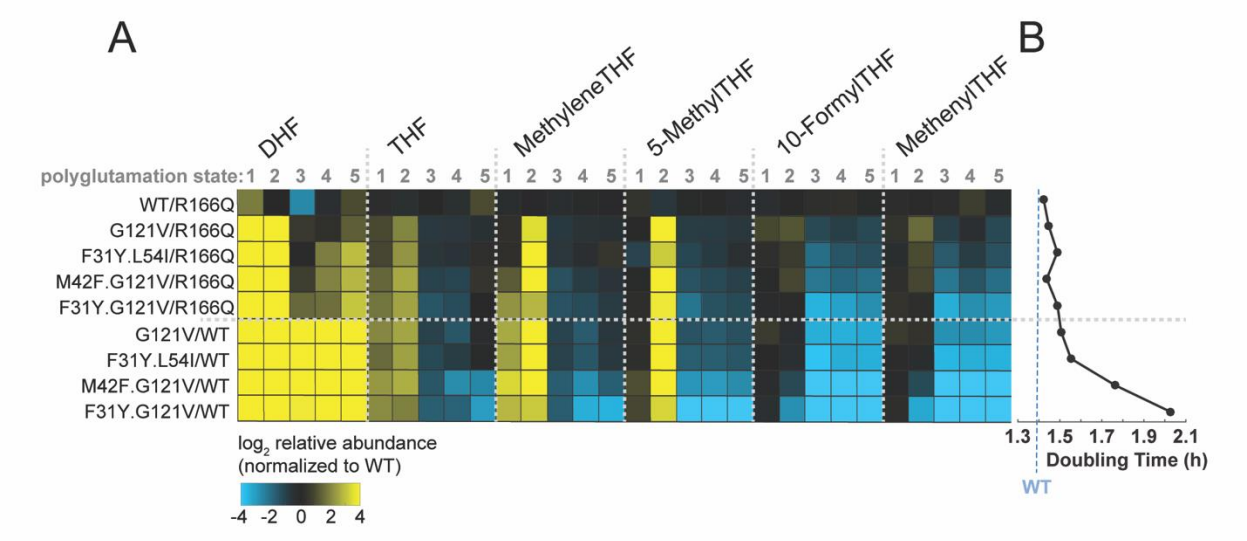

Figure 2.3 Metabolic changes due to reduced DHFR activity can be compensated through a loss-of-function in TYMS. A, Liquid chromatography-mass spectrometry profiling of intracellular folate species for select DHFR/TYMS mutant combinations from the growth competition assay. Cells were cultured in M9 media supplemented with 0.1% amicase and 50 μg/ml thymidine. DHFR/TYMS genotypes are labeled along the right-hand side of the heatmap; folate species and polyglutamylation state are indicated along the top. Data represent the mean of three replicate measurements of $\log_2$abundance relative to wild-type. In the background of a wild-type TYMS, the data show that mutations reducing DHFR activity result in an accumulation of DHF and a depletion of reduced folates (bottom four rows). The effect is partially abrogated by introducing an inactivating mutation to TYMS (rows two through five). **B,** The corresponding doubling time of each mutant pair, as measured in batch culture under the same experimental conditions.

folate polyglutamylation. [4]. The polyglutamylation of reduced folates is important for their retention and use as substrates in several downstream reactions [5]. Our findings are consistent with the inhibition of FP-γ-GS by overabundant DHF. Unsurprisingly, the metabolic signature and fitness defect that we observe due to decreased DHFR activity resembles the effect of its competitive inhibitor trimethoprim [4, 6]. Mutants displaying the most severe accumulation of DHF and depletion of THF grow more slowly (Figure 2.3B). When a loss-of-function TYMS mutant is introduced, the metabolite profiles become much closer to that of wild-type (Figure 2.3A, rows 2-5). The accumulation of DHF is somewhat abrogated and polyglutamated reduced folate levels are increased. Thus, coordinated

decreases of both DHFR and TYMS better maintain the underlying balance of metabolites. Measurements of growth rate in batch culture confirm that this restores growth rate to near wild-type levels (Figure 2.3B). These findings provide a plausible biochemical explanation for the observed statistical association between DHFR and TYMS by synteny and co-occurrence across thousands of bacteria. The bifunctional fused form of DHFR/TYMS found in protists and plants may represent an extreme outcome of this selection which guarantees stochiometric expression [7, 8].

2.1.4 Examining adaptive independence using forward evolution

The above data demonstrate coupling between DHFR and TYMS but does not exclude the possibility that they interact with the other products of the genome. To test for adaptive interactions more broadly, I proposed a second-site suppressor screen. The experiment is simple: apply a perturbation to *E. coli* DHFR, allow cells to adapt, then sequence the genome to identify compensatory mutation. This experiment would either verify the adaptive independence of the DHFR/TYMS pair or unveil other interactions that were unknown *a priori*. The antibiotic trimethoprim (TMP) is a competitive inhibitor of DHFR which allows for a titratable reduction in enzyme activity. Erdal Toprak and colleagues established an experimental system for evolving resistance to trimethoprim, and potentially many other antibiotics, in *E. coli* through sustained drug stress [9]. The workhorse of this study is a continuous culture apparatus called the morbidostat/turbidostat [10]. The morbidostat facilitates the dynamic control of drug concentration in response to growth rate and optical density (Figure 2.4). The basic idea is that drug concentration is increased as long

as the culture maintains a minimum optical density and outgrows the rate of dilution. This prevents the loss of viability due to excessive antibiotic stress, while maintaining constant selection pressure as population resistance increases.

In prior work, Toprak et al used the morbidostat to evolve trimethoprim resistant *E. coli* MG1655 with constant phenotypic adaptation over 20 days. Their experiment featured five replicate populations grown in M9 minimal media. However, the adaptive mutations resulting from this condition where wholly constrained to DHFR. The total set of possible mutations was comprised of two substitutions in the promoter region and nine in the coding

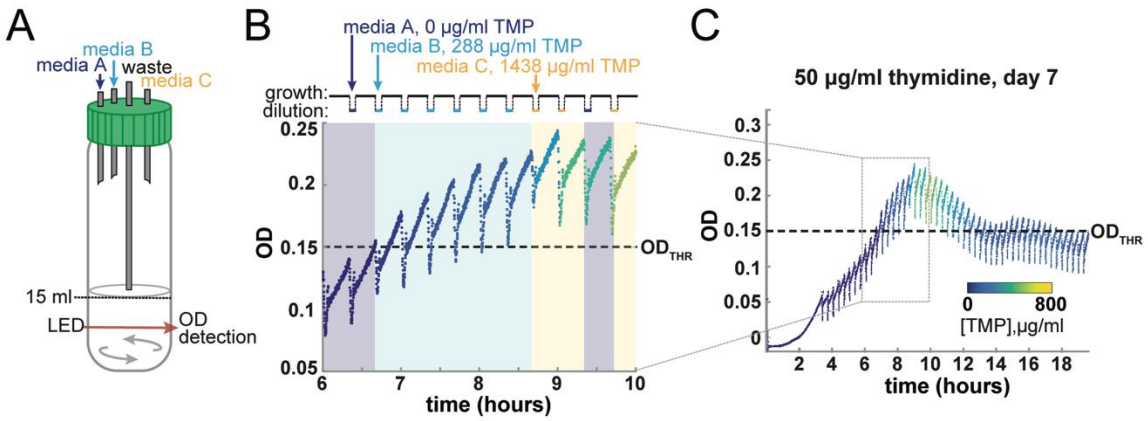

Figure 2.4 Evolution of trimethoprim (TMP) resistance using the morbidostat. A, Schematic of the continuous culture tube. Dilutions with fresh media are made through the series of inlet lines labeled 'A,' 'B,' and 'C.' A constant culture volume of 15 ml is maintained via aspiration through the waste line. **B-C,** Control strategy for the addition of trimethoprim. Pixels indicate the OD_{600} trajectory for a single tube and are colored according to the current trimethoprim concentration. The culture grows unperturbed until it reaches a minimum OD_{600} of 0.05, at which point it is subjected to periodic dilution every 20 minutes. As long as the culture is below an OD_{600} of 0.15, media 'A,' containing no trimethoprim, is used. Once the culture surpasses this second threshold, trimethoprim is introduced through the use of media 'B.' Dilution with media 'B,' continues until the culture growth rate is suppressed below the dilution rate. In the event that the concentration of trimethoprim in the culture reaches 60% of that of media 'B,'the program switches to media 'C' containing 5-fold more trimethoprim. This allows continuous selection even as the population adapts. If media 'C' is used in a given day, the trimethoprim concentration of both 'B' and 'C' are incremented by a factor of 5 following day. This example plot represents an excerpt from the data of my trimethoprim evolution experiment.

sequence. Each of the replicates featured a single promoter mutation and quasi-ordered acquisition of 2-3 coding sequence mutations. While their experiment revealed a number of modifications to the DHFR enzyme that rendered the cell less sensitive to trimethoprim, they did not observe compensatory mutations elsewhere in the pathway or genome. It is possible that the nutrient-scarce media conditions utilized were prohibitive to evolutionary routes beyond the DHFR locus. In contrast to the findings by Toprak et al, TYMS loss-of-function mutations have been observed in trimethoprim resistant clinical isolates from multiple genera of gram negative bacteria [11, 12]. This observation supports the notion that evolving trimethoprim resistance in other environments may yield different evolutionary outcomes. Understanding the evolutionary interactions of DHFR and its capacity for adaptive independence thus requires further experimental study.

2.2 Forward evolution of trimethoprim (TMP) resistance using the morbidostat

I conducted forward evolution of *E. coli* MG1655 in the presence of trimethoprim using the morbidostat. In order to elaborate on past work in this system by Toprak and colleagues, I chose growth media that provided buffering of selection on the folate pathway. In particular, I used M9 glucose media supplemented with 0.2% amicase and several concentrations of thymidine (5, 10, and 50 μg/ml). Amicase provides a source of free amino acids, which are among the essential products of the folate pathway. Thymidine concentrations were chosen to range from a partial to full rescue of TYMS activity. By alleviating selective pressure on the entire pathway, I sought to expose a broader range of

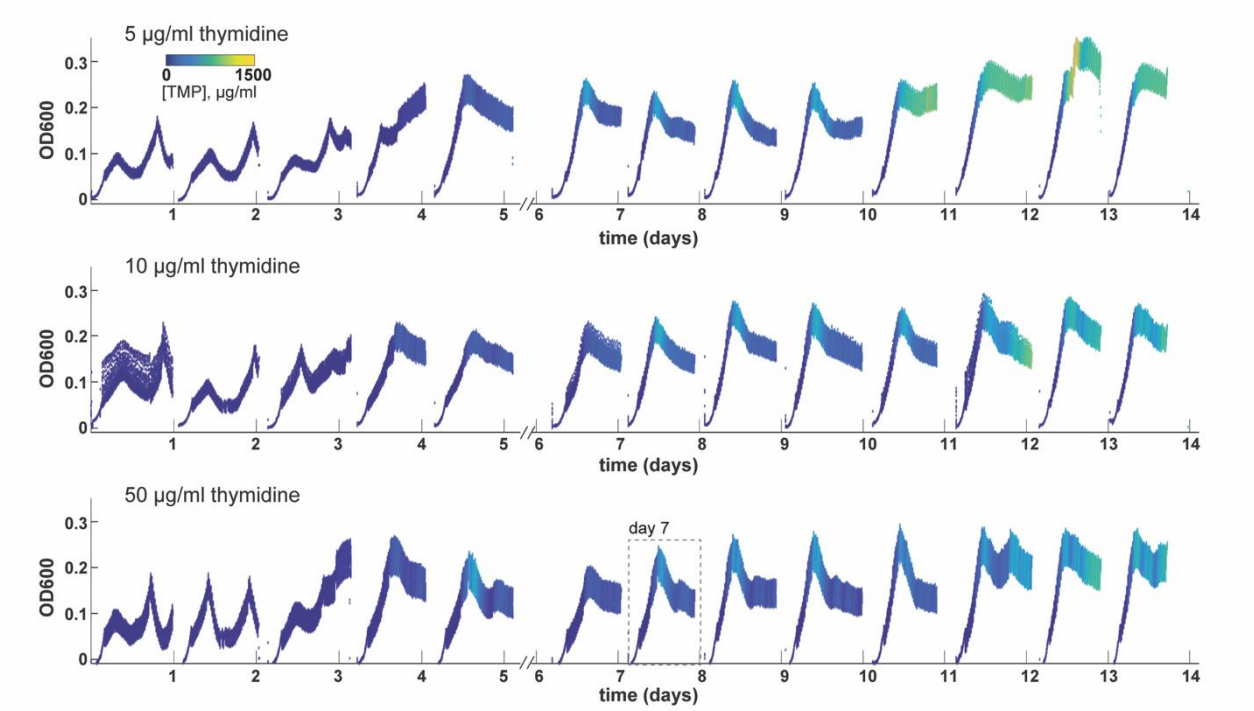

Figure 2.5 OD_{600} measurements and trimethoprim concentration over 13 days of forward evolution. The morbidostat was used to adapt three populations to trimethoprim stress in differing thymidine concentrations (indicated along the top left). Pixels are color coded according to trimethoprim concentration in the culture tube at the time of the density measurement. Discontinuities at day 5 are the result of a technical interruption; cultures we restarted using the previous day's glycerol stock. An enhanced view of day 7 in the 50 µg/ml thymidine condition is provided by Figure 2.4.

adaptive mutations without biasing the experiment toward a particular result. My experimental design was based on the common practice of conducting second site suppressor screens for essential genes under relatively permissive conditions [13]. One forward evolution population was evolved in each of the three thymidine concentrations. The strategy for modulating trimethoprim concentration in response to growth rate of each culture was adapted from Toprak et al (Figure 2.4) [9]. I observed steady increase in resistance over the course of 13 days (Figure 2.5), at which point the concentration of trimethoprim in my stocks began to approach the solubility limit of the drug. The adaptive progress in each culture can

be represented by the median trimethoprim concentration experienced in a given day. I plotted this quantity as a function of the number of generations elapsed (Figure 2.6). The results confirm the acquisition of a nearly 1000-fold increase in trimethoprim resistance by each experimental evolution population. Interestingly, the final adaptation event required more generations as thymidine supplementation increased (Figure 2.6). After 13 days of selection with trimethoprim, the experiment was terminated so that I could assess the potential drivers of adaptation through whole genome sequencing.

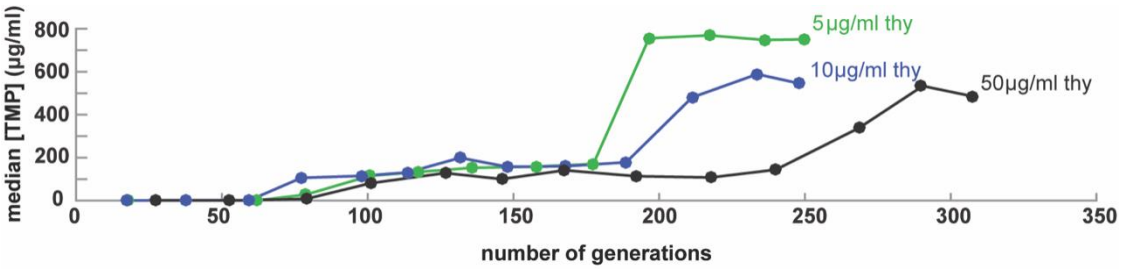

Figure 2.6 Trimethoprim concentrations permissive of growth throughout the forward evolution experiment. The plot indicates the median trimethoprim concentration in each culture tube as a function of the number of generations elapsed. Generation counts were estimated from the (piecewise) fold change in optical density over time.

2.3 Whole genome sequencing of evolved strains

The first step in mapping the genetic origin(s) of adaptation was to sequence genomes from the resulting populations. I chose ten clonal isolates (strains) from each experimental evolution condition for genotypic and phenotypic characterization. I chose to sequence individual genomes rather than the mixed population in order to preserve potentially relevant information about specific combinations of mutants. A shotgun (short-read) sequencing approach was used to obtain a genome and predicted mutations for each isolate. Read

quantity and depth varied across strains, but generally the average number of reads per base-pair exceeded 30x coverage. Read count, coverage, and statistics for each strain are reported in Table 2.1. Whole genome sequences of the selected isolates revealed that strains from all three evolution conditions acquired mutations to both DHFR and TYMS (in some cases only TYMS; Table 2.A1). Strains 4, 5, 7, and 10 from the 50 μg/ml thymidine condition displayed mutations in TYMS but no detectable changes at the *folA* (DHFR) locus or promoter region. Read count at the *folA* locus of these strains did not indicate any amplification events either. All other sequenced isolates obtained one of three amino acid substitutions in the DHFR protein coding sequence. In each case, the mutations to DHFR reproduce previously reported adaptation in an earlier morbidostat study of trimethoprim resistance [9]. The TYMS mutations observed included two instances of insertion sequence (IS1) mediated mutation, a frame shift mutation, the loss of two codons, and a non-synonymous active site mutation. These mutations are consistent with a loss-of-function in the TYMS protein. The ubiquity of mutations to DHFR and TYMS indicates that these two proteins play an important role in the evolution of trimethoprim resistance, and supports the hypothesis that they co-adapt in response to antibiotic stress.

	Condition	Strain	Coverage	Dispersion (σ^2/μ)	Reads (10^6)	Avg read length (BP)
Evolved Strains	5 thy	1	55	4.4	2.07	126
		2	40	3.4	1.52	125
		3	40	3.4	1.43	131
		4	35	3.5	1.46	114
		5	29	2.3	0.985	138
		6	38	3	1.45	122
		7	46	3	1.53	142
		8	68	4.1	2.30	137
		9	52	3.2	1.70	140
		10	52	4.2	1.80	134
	10 thy	1	32	3.4	1.13	135
		2	34	3.3	1.16	138
		3	34	3.4	1.19	135
		4	48	3.9	1.67	135
		5	42	3.6	1.46	133
		6	31	2.9	1.12	132
		7	29	2.9	1.03	131
		8	25	2.6	0.843	137
		9	31	2.9	1.07	135
		10	41	3.4	1.44	135
	50 thy	1	42	3.5	1.46	133
		2	35	3.4	1.32	126
		3	33	3	1.16	134
		4	21	3.1	0.879	116
		5	18	3.7	0.793	115
		6	25	3.5	0.990	121
		7	24	2.9	0.887	125
		8	31	3.2	1.13	130
		9	29	2.9	1.00	136
		10	24	3.2	0.936	126
Parent Strains	5 thy	1	44	3.1	1.53	134
		2	50	3.2	1.92	122
	10 thy	1	36	2.2	1.32	129
		2	40	2.5	1.45	130
	50 thy	1	35	2.3	1.18	137
		2	38	2.5	1.26	138

Table 2.1 Sequencing statistics of forward evolution strains. Ten clonal isolates (strains) were selected from the endpoint of each evolution condition for whole genome sequencing (WGS). Two clonal isolates were sampled from the corresponding parent cultures in order to identify variants already present in each population. Genomes were constructed by aligning short-length reads against a reference (see materials and methods for details). Total number of reads and average read length are displayed for each strain. Coverage refers to the mean number of reads mapped to each basepair in the genome. Dispersion indicates the normalized variance in coverage.

2.4 Materials and methods

2.4.1 Experimental model and subject details

The parent strain of the forward evolution experiment was *E. coli* MG1655 with a chromosomal green fluorescent protein (*egfp*) and chloramphenicol resistance (*cat*) cassette introduced at the P21 attachment site by phage transduction.

2.4.2 Forward evolution of TMP resistance using the morbidostat

The morbidostat/turbidostat apparatus was built as described by Toprak and colleagues [10]. To begin the experiment, the parent strain was grown overnight at 37 ˚C in Luria Broth (LB) with 30 μg/ml chloramphenicol [14] added for positive selection. The overnight culture was washed twice into M9 media the following day. All subsequent steps were performed at 30˚C in M9 minimal media supplemented with 0.4% glucose, 0.2% amicase, and 30 μg/ml of Cam. The washed overnight was used to inoculate three new cultures of M9 media further supplemented with either 5, 10, or 50 μg/ml thymidine (thy). These constitute the full experimental conditions for trimethoprim selection, and will henceforth be referred to as day 0 (with day 1 denoting the first period of morbidostat culture). Day 0 cultures were grown overnight in round-bottom tubes. The next morning, each culture was streaked onto agar plates: two colonies would be sampled from each plate to provide genomes of the parental strain. The remainders of the day 0 cultures were used to inoculate morbidostat tubes with the corresponding media and thymidine supplementation. Each population was inoculated at a starting density of approximately 0.005. Cultures were allowed to grow unperturbed until they surpassed an OD_{600} of 0.06, at which point they underwent periodic dilutions with fresh media. The dilution rate is described by the formula:

$$r_{dil} = f \ln \frac{V}{V + \Delta V}$$

where $V = 15ml$ is the culture volume, and $\Delta V = 3\,ml$ is volume added. A dilution frequency of $f = 3\,h^{-1}$ was chosen, resulting in r_{dil} = 0.55. Dilutions for each population were made using one of three media (labeled A, B, and C) with matching thymidine supplementation but different amounts of trimethoprim. Media A always contains no trimethoprim and is used whenever the culture is below an optical density of 0.15 or the growth rate dips below the dilution rate. Above this density, trimethoprim is introduced through the use of media B. In the event that the concentration of trimethoprim in the culture tube reaches 60% of the stock of media B, the program switches to media C containing 5-fold more trimethoprim. This process allows for continuous selection even as the population adapts. The initial drug concentrations of these media were 0, 11.5, and 57.5 μg/ml. Cycles of growth and dilution were sustained for a period of ~22 hours each day, at which point the run was stopped in order to make glycerol stocks, replenish media, and update TMP stock concentrations. If media C was used in a given day, then medias B and C are incremented by a factor of 5 for that population. Culture vials for the next day of evolution were filled with fresh media and inoculated with 300 μl from the previous day's culture. A schematic of the morbidostat and an illustration of the protocol for drug addition is provided by Figure 2.4. The complete trajectories of OD_{600} versus time for 13 days of experimental evolution are shown in Figure 2.5.

2.4.3 Genome preparation and sequencing

Two clonal isolates were chosen from each adapted day 0 culture, and ten isolates were randomly sampled from the endpoint of each evolution condition (36 strains in total).

Isolation of genomic DNA was conducted using the QIAamp DNA Mini Kit (Qiagen). The Nextera XT DNA Library Prep Kit (Illumina) was used to fragment and label each genome for sequencing. Paired end sequencing was performed using a v2 300-cycle MiSeq Kit (Illumina). Average read length and coverage is reported in Table 2.1.

2.4.4 *E. coli* genome assembly

Genome assembly and mutation prediction was conducted using the *bowtie2* [15] dependent program *breseq* [16]. The reference sequence for read alignment was a modification of the *E. coli* MG1655 complete genome (accession no. NC_000193) edited to include the GFP marker and chloramphenicol resistance cassette in the parent strain. *Breseq* predicted mutations are reported in Table 2.A1.

References

1. Schober, A.F., et al., *A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism.* Cell Rep, 2019. **27**(11): p. 3359-3370.e7.
2. Reynolds, K.A., R.N. McLaughlin, and R. Ranganathan, *Hotspots for allosteric regulation on protein surfaces.* Cell, 2011. **147**(7): p. 1564-75.
3. Lu, W., Y.K. Kwon, and J.D. Rabinowitz, *Isotope ratio-based profiling of microbial folates.* J Am Soc Mass Spectrom, 2007. **18**(5): p. 898-909.
4. Kwon, Y.K., et al., *A domino effect in antifolate drug action in Escherichia coli.* Nat Chem Biol, 2008. **4**(10): p. 602-8.
5. McGuire, J.J. and J.R. Bertino, *Enzymatic synthesis and function of folylpolyglutamates.* Mol Cell Biochem, 1981. **38 Spec No**(Pt 1): p. 19-48.
6. Kwon, Y.K., M.B. Higgins, and J.D. Rabinowitz, *Antifolate-induced depletion of intracellular glycine and purines inhibits thymineless death in E. coli.* ACS Chem Biol, 2010. **5**(8): p. 787-95.
7. Beverley, S.M., T.E. Ellenberger, and J.S. Cordingley, *Primary structure of the gene encoding the bifunctional dihydrofolate reductase-thymidylate synthase of Leishmania major.* Proc Natl Acad Sci U S A, 1986. **83**(8): p. 2584-8.
8. Lazar, G., H. Zhang, and H.M. Goodman, *The origin of the bifunctional dihydrofolate reductase-thymidylate synthase isogenes of Arabidopsis thaliana.* Plant J, 1993. **3**(5): p. 657-68.
9. Toprak, E., et al., *Evolutionary paths to antibiotic resistance under dynamically sustained drug selection.* Nat Genet, 2012. **44**(1): p. 101-5.
10. Toprak, E., et al., *Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition.* Nat Protoc, 2013. **8**(3): p. 555-67.
11. King, C.H., D.M. Shlaes, and M.J. Dul, *Infection caused by thymidine-requiring, trimethoprim-resistant bacteria.* J Clin Microbiol, 1983. **18**(1): p. 79-83.
12. Kriegeskorte, A., et al., *Inactivation of thyA in Staphylococcus aureus attenuates virulence and has a strong impact on metabolism and virulence gene expression.* MBio, 2014. **5**(4): p. e01447-14.
13. Forsburg, S.L., *The art and design of genetic screens: yeast.* Nat Rev Genet, 2001. **2**(9): p. 659-68.
14. Michener, J.K., et al., *Effective use of a horizontally-transferred pathway for dichloromethane catabolism requires post-transfer refinement.* Elife, 2014. **3**.
15. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
16. Deatherage, D.E. and J.E. Barrick, *Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq.* Methods Mol Biol, 2014. **1151**: p. 165-88.

Appendix 2

Name	Mutation		Strains	Annotation
folA	CCG to CTG	P21L	10thy: 7	dihydrofolate reductase
	CTC to CGC	L28R	5thy: 1-10; 50thy: 1, 6	
	TGG to AGG	W30R	10thy: 1-6, 8-10; 50thy: 2, 3, 8, 9	
thyA	IS1(+) +9bp	(627-635/795 nt)	50thy: 2, 3, 8, 9	thymidylate synthetase
	IS1(+) +9bp	(564-572/795 nt)	50thy: 4, 5, 10	
	Δ1	(535/795 nt)	10thy: 1-10	
	Δ6	(525-530/795 nt)	50thy: 7	
	TGG to AGG	W133R	50thy: 1, 6	
	Δ6	(64-69/795 nt)	5thy: 1-10	
dusB	IS1(+) +10bp	(573-582/966 nt)	10thy: 1-10	tRNA-dihydrouridine synthase B
	IS1(+) +8bp	(818-825/966 nt)	5thy: 1-10	
cynR	AAT to AAA	N272K	5thy: 10	transcriptional activator of cyn operon; autorepressor
	AAA to GAA	K271E	5thy: 6, 9, 10; 10thy: 4-6, 10	
	TTG to TTT	L267F	5thy: 3, 8, 10; 10thy: 1, 7, 8; 50thy: 6	
ychE/oppA	Δ1199	(+254/-485)	5thy: 2, 4, 5; 10thy: 2, 3, 7-10	UPF0056 family inner membrane protein/oligopeptide ABC transporter periplasmic binding protein
gadX	GAT to GGT	D38G	5thy: 5; 10thy: 2	acid resistance regulon transcriptional activator; autoactivator
	GCG to TCG	A37S	10thy: 2, 3, 5-10; 50thy: 8	
otsB/araH	A to G	(-136/+31)	10thy: 2, 3; 50thy: 3, 9	trehalose-6-phosphate phosphatase, biosynthetic/L-arabinose ABC transporter permease
	C to A	(-142/+25)	5thy: 3; 10thy: 2	
	G to A	(-164/+3)	10thy: 7; 50thy: 9	
yfbL/yfbM	G to A	(+31/-72)	5thy: 6, 7, 8; 10thy: 2, 5, 10	putative M28A family peptidase/DUF1877 family protein
betI	TCC to CCC	S182P	10thy: 1, 2, 4	choline-inducible betIBA-betT divergent operon transcriptional repressor
	ACC to CCC	T178P	10thy: 4	

Name	Mutation		Strains	Annotation
betI	GAT to GAA	D176E	10thy: 1, 4, 5	
	GAT to AAT	D176N	10thy: 2, 4	
ybaL	TAA to GAA	*559E	10thy: 8; 50thy: 10	inner membrane putative NAD(P)-binding transporter
	GTG to GGG	V555G	5thy: 5, 6	
cat/egfp	C to G	(+289/-204)	5thy: 3, 8; 10thy: 4, 5	chloramphenicol acetyltransferase/green fluorescent protein
	C to G	(+299/-194)	10thy: 5	
fis	TCG to TAG	S30*	50thy: 2, 3, 8, 9	global DNA-binding transcriptional dual regulator
lacI	CCC to CCA	P332P	10thy: 8	lactose-inducible lac operon transcriptional repressor
	ACC to CCC	T329P	5thy: 6; 10thy: 7, 8	
chaA	ACC to CCC	T10P	50thy: 7	calcium/sodium:proton antiporter
	GTA to GAA	V8E	10thy: 5	
	CAA to AAA	Q5K	10thy: 2	
csrA	TAA to TAC	*62Y	50thy: 4, 5, 10	pleiotropic regulatory protein for carbon source metabolism
citG	ACC to CCC	T255P	5thy: 2; 10thy: 3	2-(5"-triphosphoribosyl)-3'-dephosphocoenzyme-A synthase
ompF/asnS	G to T	(-529/+74)	10thy: 2	outer membrane porin 1a (Ia;b;F)/asparaginyl tRNA synthetase
	C to A	(-540/+63)	10thy: 8	
lpoB	CAA to CAC	Q38H	5thy: 6; 50thy: 9	OM lipoprotein stimulator of MrcB transpeptidase
sapA	ACC to CCC	T304P	5thy: 2; 50thy: 4	antimicrobial peptide transport ABC transporter periplasmic binding protein
yddE	CAA to CAC	Q14H	50thy: 4, 5	PhzC-PhzF family protein
	ACC to CCC	T12P	50thy: 4	
yghQ	GTG to GGG	V332G	5thy: 7; 10thy: 10	putative inner membrane polysaccharide flippase
	GGA to GGG	G323G	10thy: 10	
agaD	GGA to GGG	G120G	10thy: 1	N-acetylgalactosamine-specific enzyme IID component of PTS
	GCC to TCC	A126S	10thy: 1, 3	
rtcA	AGT to GGT	S215G	10thy: 3; 50thy: 9	RNA 3'-terminal phosphate cyclase
gntR	GAA to GGA	E147G	10thy: 4, 5	d-gluconate inducible gluconate regulon transcriptional repressor

Name	Mutation		Strains	Annotation
	GTG to GGG	V146G	10thy: 4	
yiaK	ACC to CCC	T309P	10thy: 7; 50thy: 6	2,3-diketo-L-gulonate reductase, NADH-dependent
	GAA to AAA	E313K	10thy: 7	
rrfB/murB	C to G	(+126/-175)	10thy: 2, 10	5S ribosomal RNA of rrnB operon/UDP-N-acetylenolpyruvoylglucosamine reductase, FAD-binding
ampC	GTA to GGA	V48G	10thy: 1, 8	penicillin-binding protein; beta-lactamase, intrinsically weak
thrC	CTC to ATC	L3I	50thy: 7	L-threonine synthase
dapB/carA	T to A	(+301/-155)	50thy: 6	dihydrodipicolinate reductase/carbamoyl phosphate synthetase small subunit, glutamine amidotransferase
paoC	CAA to AAA	Q72K	50thy: 9	PaoABC aldehyde oxidoreductase, Moco-containing subunit
acrR	IS1(+) +9bp	(320-328/648 nt)	50thy: 7	transcriptional repressor
ybdK	TGG to CGG	W263R	5thy: 5	weak gamma-glutamyl:cysteine ligase
dtpD/ybgI	T to A	(-84/-187)	10thy: 3	dipeptide and tripeptide permease D/NIF3 family metal-binding protein
ssuB	GGC to GGG	G44G	10thy: 3, 8	aliphatic sulfonate ABC transporter ATPase
	GTG to GGG	V43G	50thy: 10	
putP	GAT to GGT	D55G	50thy: 3	proline:sodium symporter
serX	A to G	(72/88 nt)	10thy: 8	tRNA-Ser
flgF	CAG to CGG	Q19R	50thy: 10	flagellar component of cell-proximal portion of basal-body rod
pabC	TAC to GAC	Y92D	10thy: 3	4-amino-4-deoxychorismate lyase component of para-aminobenzoate synthase multienzyme complex
dadX	ACC to CCC	T284P	5thy: 2	alanine racemase, catabolic, PLP-binding
oppF	CCG to CAG	P273Q	10thy: 3	oligopeptide ABC transporter ATPase
uspF/ompN	G to A	(-108/+33)	5thy: 7	stress-induced protein, ATP-binding protein/outer membrane pore protein N, non-specific
yneM/dgcZ	G to A	(+75/+144)	10thy: 3	inner membrane-associated protein/diguanylate cyclase, zinc-sensing
yebV/yebW	G to T	(+26/-79)	10thy: 7	uncharacterized protein/uncharacterized protein
araH	CAA to AAA	Q322K	10thy: 2	L-arabinose ABC transporter permease
mntH	GTG to GGG	V313G	10thy: 1	manganese/divalent cation transporter
xapR	ATG to ATA	M176I	5thy: 6	transcriptional activator of xapAB
uraA	ATT to GTT	I311V	50thy: 7	uracil permease

Name	Mutation		Strains	Annotation
relA	CAT to CAA	H518Q	5thy: 3	(p)ppGpp synthetase I/GTP pyrophosphokinase
ptrA	GAT to GAA	D38E	10thy: 1	protease III
	GAT to AAT	D38N	10thy: 1	
	CGT to CGA	R35R	10thy: 1	
rsmI	CAT to CAA	H235Q	10thy: 6	16S rRNA C1402 2'-O-ribose methyltransferase, SAM-dependent
gltF/yhcA	Δ4	(+90/-79)	50thy: 7	periplasmic protein/putative periplasmic chaperone protein
fis-yhdX	Δ9555		50thy: 7	fis, yhdJ, yhdU, acrS, acrE, acrF, yhdV, yhdW, yhdX
acrS	TAT to TTT	Y187F	10thy: 3	acrAB operon transcriptional repressor
secY	GTA to GGA	V274G	10thy: 2	preprotein translocase membrane subunit
xylF	GAA to AAA	E195K	50thy: 6	D-xylose transporter subunit
uhpT	GAA to GGA	E447G	50thy: 8	hexose phosphate transporter
pstA	GGT to GGG	G112G	50thy: 5	phosphate ABC transporter permease
	ATT to GTT	I106V	50thy: 5	
pyrB	ACC to CCC	T54P	50thy: 4	aspartate carbamoyltransferase, catalytic subunit

Table 2.A1 Complete list of non-synonymous mutations observed in the trimethoprim evolution experiment. Mutation predictions and gene annotations were produced by Breseq (see 2.4 Materials and methods). The strain column indicates experimental condition and clonal isolate(s) in which each mutation was observed. Location of each mutation in the amino acid sequence or intergenic region is indicated where applicable.

CHAPTER THREE

The Genetic Drivers of Trimethoprim Resistance

3.1 Background and introduction

The major goal of my trimethoprim (TMP) evolution experiment was to uncover the genetic basis of adaptation to the targeted inhibition of dihydrofolate reductase (DHFR). After 13 days of experimental evolution, whole genome sequencing was used to produce a map of the resultant mutations. Some fraction of these are expected to be non-adaptive, while others provide a competitive advantage [1]. Neutral or deleterious 'hitchiker' mutations can fix in a population by occurring alongside an advantageous allele [2]. In addition, it is impossible to completely decouple the selection for trimethoprim resistance from adaptation to the growth medium, and turbidostat environment. Mutations that optimize the ability of *E. coli* to harness glucose as a carbon source, for example, also have a chance to come to fixation. My experimental evolution populations began each day by growing to a threshold density without the addition of additional trimethoprim. This period of relaxed selection allows for the enrichment of such alleles. These could be described as 'generally adaptive' if they provide a competitive advantage even in the absence of trimethoprim inhibition. One cannot exclude the possibility that these would become prevalent in the same strains that acquired resistance-causing mutations. Therefore it is difficult, if not impossible, to discern between hitchhikers, resistance-causing, and generally adaptive variants from sequencing data alone [1]. Thus, targeted experiments are needed in order to ascribe specific phenotypic

consequences to the observed mutations, and to identify adaptive interaction(s) between DHFR and the other mutated genes. More generally, this work establishes appropriate experimental approaches necessary to interrogate the adaptive couplings of a given enzyme.

3.2 Phenotyping the thymidine dependence in the evolved populations

The *thyA* (TYMS) locus acquired the largest variety of mutations across the three populations of my trimethoprim evolution experiment, with particularly rich diversity in the 50 μg/ml thymidine condition. In total, these span substitutions near the active site, insertion, deletion, frame shift, and transposable element (IS1) mediated mutations. In all cases, the observed mutations seem likely to induce a reduction or loss-of-function in thymidylate synthase (TYMS). The observed genotypic diversity would make sense given the large target size of loss-of-function mutations. Additionally, thymidine auxotrophy has been observed in trimethoprim resistant clinical isolates from multiple genera of gram negative bacteria [3, 4]. To test for a reduction of TYMS activity in the evolved strains, I measured the dependence of growth rate on exogenously supplemented thymidine. I measured the total growth of all 30 endpoint strains across 8 different concentrations of thymidine. Each of the strains demonstrated a roughly monotonic increase with thymidine concentration (Figure 3.1). In all cases, the ability to grow in the absence of exogenous thymidine had been lost, indicating that they had become auxotrophs. These findings suggest that the *thyA* variants produced by my trimethoprim evolution experiment are an example of convergent evolution. Despite the apparent diversity, the mutations are phenotypically equivalent in that they inactivate TYMS and cause thymidine auxotrophy.

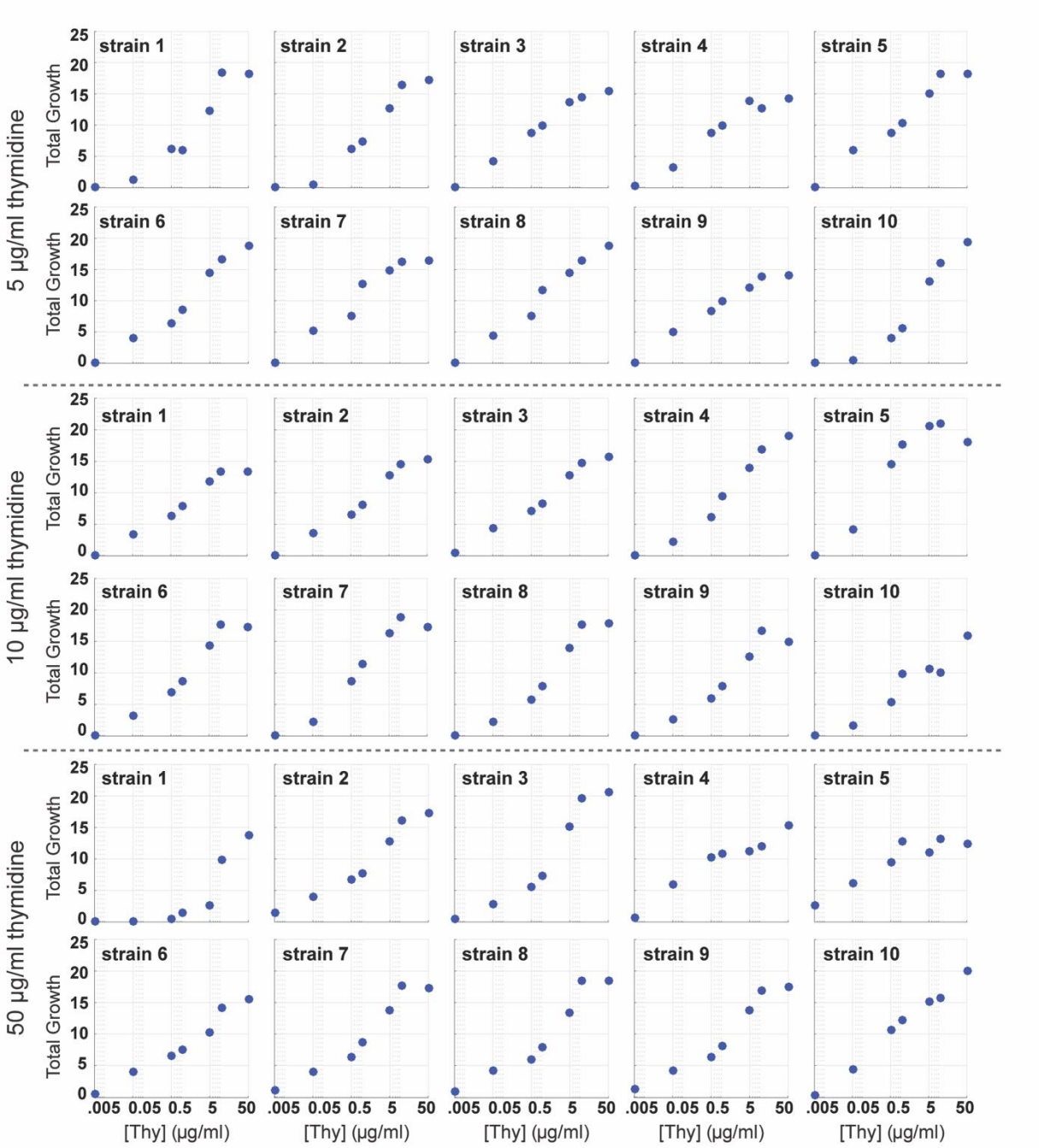

Figure 3.1 Thymidine dependence of the 30 evolved strains. Data shown indicate singlicate measurements. The y-axis denotes the positive integral of log(OD600) evaluated over 20 hours of growth. At low thymidine, total growth exhibits a monotonic increase as thymidine concentration is also increased. In a number of cases, this culminates with a plateau in which total growth flattens out. Each of the strains exhibits thymidine auxotrophy.

3.3 Trimethoprim stress is necessary to induce rapid *thyA* loss-of-function

Thymidine auxotrophy was observed in every clonal isolate harvested from the endpoint of the trimethoprim evolution experiment. It is possible that the presence of thymidine in the growth medium alone is sufficient to drive inactivation of the *thyA* locus. This could occur if the enzymatic activity of TYMS carried some metabolic cost which was counterbalanced by the necessity of thymidine production. TYMS converts 5,10-methylene tetrahydrofolate (THF) back into dihydrofolate (DHF). Dihydrofolate is not itself active in one carbon metabolism, requiring NADPH in order to be reduced [5]. Therefore, TYMS activity incurs an energetic toll on the cell in the form of NADPH consumption. As a result, a loss-of-function mutation in TYMS may provide a competitive advantage against wild-type in the presence of exogenous thymidine. So, is thymidine supplementation sufficient to drive TYMS loss-of-function on a similar timescale even in the absence of trimethoprim? To answer this question, I used continuous culture to facilitate sustained exponential growth in the presence of exogenous thymidine. For the purpose of direct comparison, the experimental conditions and parental *E. coli* strain matched that of my trimethoprim evolution experiment. Triplicate populations were grown in M9 minimal media supplemented with 0.4% glucose, 0.2% amicase and 50 μg/ml thymidine. This thymidine concentration provided the strongest rescue of auxotrophs and corresponds to the highest level of supplementation in the trimethoprim evolution experiment. Continuous exponential growth was achieved using the 'turbidostat' mode of the morbidostat/turbidostat apparatus [6]. In the turbidostat setting, cultures grow uninterrupted while below a target optical density. Each time the density surpasses that threshold, that culture is diluted with a fixed volume of fresh media. This

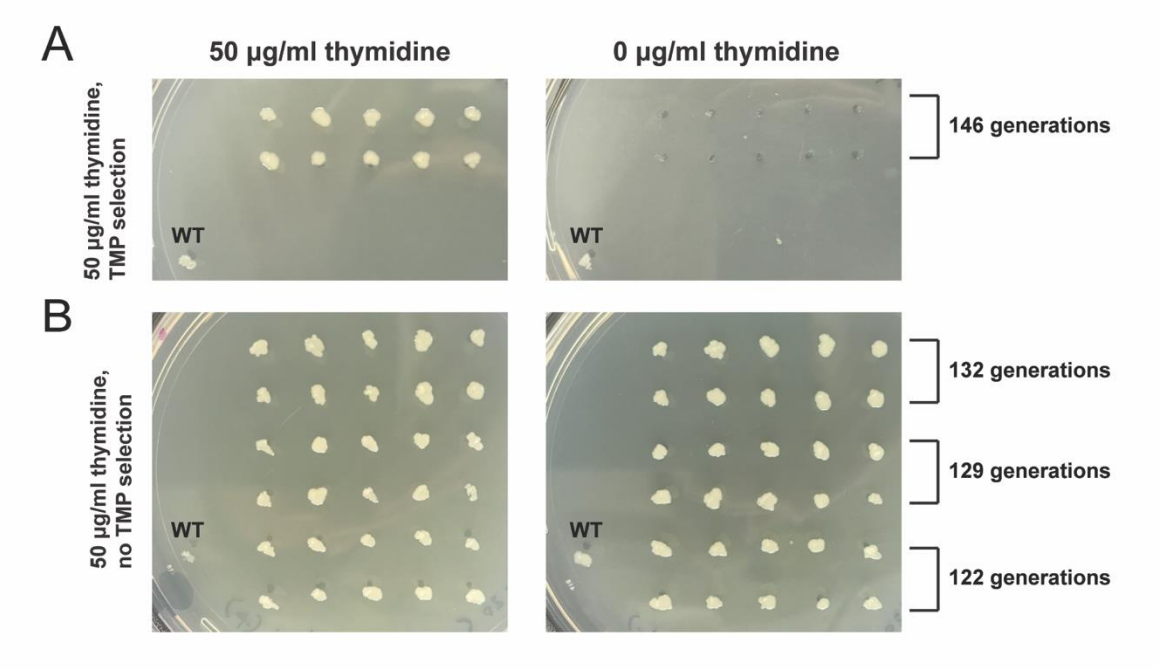

Figure 3.2 Lack of thymidine dependence after growth in the absence of TMP. A, Ten colonies from day 6 of the 50 µg/ml thymidine condition of the trimethoprim evolution experiment. Replica plating on 0 and 50 µg/ml thymidine demonstrates that they have become thymidine auxotrophs. **B,** Ten colonies from three replicate populations grown in 50 µg/ml thymidine without TMP selection. Cultures were grown until biofilm formation became prohibitive. Replica plating indicates that TYMS activity was retained in all strains.

program facilitates sustained exponential growth in a confined range of optical densities. The threshold was set at 0.15, corresponding to the target OD for TMP addition in the trimethoprim evolution experiment.

My three replicate cultures grew under these conditions for a period of five days, at which point biofilm formation became prohibitive. In a similar fashion to the trimethoprim evolution experiment, ten clonal isolates were harvested from each endpoint population for phenotyping. I used replica plating on Luria Broth (LB) supplemented with 0 and 50 µg/ml thymidine to screen for TYMS loss-of-function. Results indicate that all 30 strains grown without trimethoprim selection retained TYMS function. For comparison, I selected 10

colonies from day 6 of the 50 μg/ml thymidine condition of the trimethoprim evolution experiment. Replica plating demonstrated that thymidine auxotrophy had already fixed in the population by this time (Figure 3.2). To quantify the amount of evolutionary time elapsed in each case, I used optical density versus time to estimate the total number of generations. I found that day 6 of the corresponding trimethoprim evolution experiment condition constituted 10-20 more generations than those populations grown without selection (Figure 3.2). Thus, I conclude that the rapid acquisition and fixation of a TYMS loss-of-function variant was dependent on the presence of trimethoprim and not just thymidine.

3.4 Assessing the genetic drivers of resistance

Based on biochemical context and evolutionary statistics, I hypothesized that paired mutations in DHFR and TYMS were driving adaptation to trimethoprim stress. Consistent with this, TYMS acquired a loss-of-function mutation in a trimethoprim-dependent manner in every evolved population of my morbidostat forward evolution experiment. Additionally, one of three amino acid substitutions was observed in each *folA* (DHFR) locus in 27 out of 30 total strains (across all experimental conditions). All three mutations had been previously observed in trimethoprim resistant *E. coli* [7]. The affected residues are located in the substrate binding pocket. Biochemical characterization shows that these substitutions confer additional specificity to DHFR by either reducing the competitive binding of trimethoprim or increasing its affinity for the proper substrate [8, 9]. Mutations were not observed elsewhere in the pathway. These findings generally support the hypothesis that DHFR and TYMS function as an adaptive unit. However, the role of other common genetic variants in the

evolution of trimethoprim resistance cannot be excluded from sequencing data alone. A few genes outside of the folate pathway were found to mutate in multiple experimental conditions (Table 3.A1). Of particular interest were recurring mutations to the *gadX* and *dusB* loci. The *gadX* gene product is known regulator of the acid response system; trimethoprim has been shown to induce an acid response in *E. coli* which implicates the up-regulation of the *gadX* target genes *gadB/C* [10]. The repeated interruption of the *dusB* reading frame by IS1 mediated insertions is notable because *dusB* is located in an operon upstream of the *E. coli* global transcriptional regulator *fis* [11]. As previously mentioned, such mutations could be neutral or generally contribute to growth in this these continuous culture conditions. In order to quantify the genetic origin of adaptation to trimethoprim, I asked whether the mutation pairs observed at the *folA* and *thyA* loci are sufficient in order to recover the full resistance phenotype. If true, this would demonstrate that adaption was driven by these two genes and not mutations elsewhere in the genome.

I used lambda red recombineering to introduce representative pairs of *folA/thyA* mutations from the trimethoprim evolution experiment back into the ancestral wild-type background. I constructed four total genotypes, which were termed 'reconstitution strains.' Each reconstituted genotype included a TYMS loss-of-function (Δ25-26) with either a wild-type, P21L, W30R, or L28R DHFR allele (labeled R1-4 respectively). Erdal Toprak generously provided three strains containing only the DHFR single mutants for comparison [12]. These were produced through a different recombination protocol and thus contain additional chromosomal antibiotic markers but are otherwise identical. I phenotyped the ancestral strain, evolved strains, *folA* single mutants, and reconstitution strains for

trimethoprim resistance. In each case, phenotyping was conducting in M9 glucose media supplemented with 0.2% amicase and the thymidine concentration matching each respective strain's forward evolution condition. Reconstitution and DHFR single mutant strains were phenotyped in both the 5 and 5 μg/ml thymidine conditions. Triplicate growth measurements were made for each strain across an array of trimethoprim concentrations. The resulting dose response curves were used to estimate IC50, the drug concentration at which growth is half maximal. Most evolved strains obtained an IC50 between 700-1000 μg/ml, representing nearly a thousand-fold increase over the parental strains (Table 3.1). The entire genotype to phenotype mapping is illustrated in Figure 3.3. The reconstitution strains demonstrated a level of resistance that was equal or greater than that of the evolved strains when measured in the condition in which they occurred. Consistent with this, one of the evolved strains only contained mutations in DHFR and TYMS (Figure 3.3C evolved colony #1 in the 50 μg/ml thymidine condition). The resistance of DHFR single mutants (shown in red) was markedly lower than the corresponding DHFR/TYMS double mutant. The paired mutations featured in the reconstitution strains provide a greater level of resistance in combination than they do individually. Not only do these findings indicate that DHFR and TYMS coevolved in response to trimethoprim stress, but changes to these two loci are sufficient to reproduce the full resistance phenotype of the evolved populations.

	Condition	Strain	IC50 (μg/ml)	Std Err
Evolved Strains	5 thy	1	750	21
		2	720	24
		3	710	85
		4	910	11
		5	770	7.7
		6	790	7.3
		7	870	12
		8	840	10
		9	640	46
		10	830	81
	10 thy	1	890	63
		2	870	19
		3	1000	37
		4	970	25
		5	900	150
		6	1000	18
		7	830	120
		8	1100	37
		9	1000	140
		10	1000	31
	50 thy	1	1100	55
		2	780	29
		3	820	49
		4	NA	NA
		5	NA	NA
		6	1000	18
		7	820	77
		8	870	20
		9	760	65
		10	NA	NA

	Condition	Strain	IC50 (μg/ml)	Std Err
Parent Strains	5 thy	1	0.91	0.036
		2	1.1	0.043
	10 thy	1	0.86	0.0071
		2	0.95	0.11
	50 thy	1	1.2	0.037
		2	1.3	0.091
***folA* single mutants**	5 thy	WT	4.6	0.049
		P21L	41	0.5
		W30R	14	0.6
		L28R	370	4.9
	50 thy	WT	5.1	0.21
		P21L	40	0.72
		W30R	18	0.61
		L28R	420	8.9
Reconstitution Strains	5 thy	WT/Δ25-26	NA	NA
		P21L/Δ25-26	450	25
		W30R/Δ25-26	580	25
		L28RR/Δ25-26	1100	21
	50 thy	WT/Δ25-26	NA	NA
		P21L/Δ25-26	820	81
		W30R/Δ25-26	1000	45
		L28RR/Δ25-26	>1800	NA

Table 3.1 Trimethoprim resistance (IC50) for forward evolution strains. IC50 and standard error were determined from triplicate growth measurements taken across an array of trimethoprim concentrations. Measurements were conducted in the thymidine concentration indicated under the 'condition' tab. For forward evolved and parent strains, this means that IC50 was measured in the same condition as the initial selection using the morbidostat. Reconstitution strains are labeled based on their respective *folA/thyA* mutant pairs in an otherwise clean genetic background. Reconstitution and *folA* single mutant strains were measured the two indicated thymidine concentrations. An estimate of IC50 could not be obtained for most strains featuring a loss-of-function in TYMS paired with a wild-type DHFR (evolved strains 50thy-4,5,10; recon strain 1). These grew slowly in all TMP concentrations without a clear sigmoidal dose response.

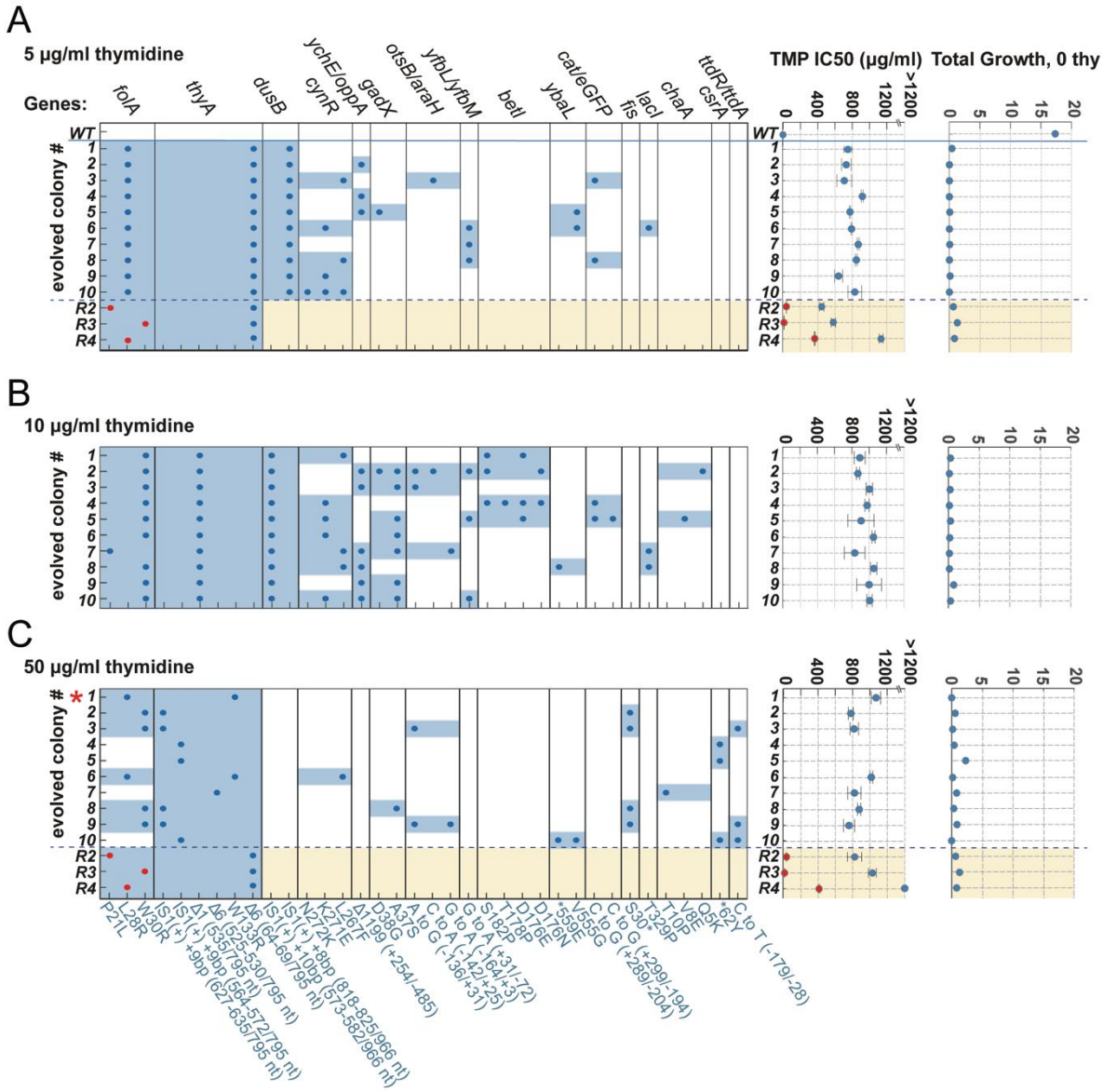

Figure 3.3 Genotype to phenotype map of the trimethoprim evolution experiment. Ten colonies (strains) were isolated from the endpoint of each forward evolution population (30 in total) for characterization. **A-C,** The leftmost panel of each section displays the mutations observed in the strains sampled from that evolution condition. Genes mutated in two or fewer strains across all conditions are excluded for brevity (as are synonymous mutations). See Table 2.A1 for a complete list of non-synonymous mutations. Gene names are labeled across the top edge of the mutation maps, while the specific nucleotide and amino acid changes are denoted along the bottom. If a given gene is mutated in particular strain, then the whole section corresponding to that gene is shaded blue for the sake of visual inspection. All but four strains acquired mutations to both *folA* and *thyA*, which encoded DHFR and TYMS. One strain which obtained mutations in only DHFR and TYMS is labeled with a small red star. Trimethoprim resistance and thymidine dependence phenotypes are displayed to the right of the mutation maps. Resistance is reported as an IC50, with standard error computed across triplicate measurements (see Table 3.1 for exact values). For comparison, three 'reconsitution strains' (**R2-4**), featuring representative pairs of mutations in *folA* and *thyA* have been included in panels A and C. Red dots correspond to the phenotype of the *folA* mutation taken alone. Thymidine dependence is represented by the positive integral of $\log(OD_{600})$ over time (see Figure 3.1 for complete set of measurements).

3.5 Materials and methods

3.5.1 Calculation of total growth for thymidine dependence and IC50 estimation

Growth was quantified as the positive integral of OD_{600} over time. This measure is sensitive to mutational or drug-induced changes in the duration of the lag-phase as well as exponential growth rate [7]. For each replicate, I defined a start time (t_0) at the end of lag phase for the reference condition (50 μg/ml thymidine or 0 μg/ml trimethoprim). Start time was chosen computationally as the last point before monotonic growth above the limit of detection. The log(OD_{600}) curves were vertically shifted such that the start time becomes zero and all subsequent values are positive. Curves were then numerically integrated using the trapezoid method over an interval of 15 hours in the case of thymidine dependence, and 10 hours for trimethoprim dose-response.

3.5.2 Measurement of growth as a function of exogenous thymidine concentration

All strains were grown overnight in LB + 5 μg/ml thymidine (thy), with the exception of those evolved in the 50μg/ml thy condition, which were supplemented with 50 μg/ml thy to ensure viability. Overnight cultures were washed twice into M9 media supplemented with 0.4% glucose and 0.2% amicase. These were used to inoculate a 96-well plate containing an array of thymidine concentrations. Cultures began at an OD_{600} of 0.005, and grew at 30°C over a period of 20 hours with periodic injection of distilled, deionized water to maintain culture volume against evaporative loss. Optical density was monitored in a Victor X3 plate reader.

3.5.3 Turbidostat culture without trimethoprim selection in 50 μg/ml thymidine

The *E. coli* MG1655 strain used for this experiment was identical to the trimethoprim evolution parent strain. To begin the experiment, the parent strain was cultured overnight at 37°C in Luria Broth (LB); 30 μg/ml of chloramphenicol [13] was added for positive selection. For subsequent steps and turbidostat growth, I used chloramphenicol selective M9 media supplemented with 0.4% glucose, 0.2% amicase, and 50 μg/ml thy. The overnight culture was washed twice with M9, and then back diluted for a second overnight adaptation at 30°C. The following day, the parent culture was used to inoculate three turbidostat tubes containing 17ml of M9 supplemented with 50 thy. The starting optical density of each culture was approximately 0.005. Each culture grew unperturbed while below a threshold OD_{600} of 0.15, at which point it was diluted with 2.4 ml of fresh media. Cycles of sustained exponential growth density continued for a period of ~22 hours a day, at which point the run was stopped in order to make glycerol stocks and replenish the media. Culture vials for the following day of continuous culture were filled with fresh media and inoculated with 300 μl of the culture from the previous day.

3.5.4 Construction of the reconstitution strains using scarless recombination

I followed the protocol for scarless genome integration using a modified λ-red system developed by Tas et al. [14]. In this method, a tetracycline [15] resistance cassette ("landing pad") is first integrated at the target set. The landing pad is then excised by the endonuclease I-SceI, and replaced with the desired mutation by λ-red mediated recombination. $NiCl_2$ is used for counterselection against cells retaining the tetracycline cassette. Tas et al. have provided a detailed protocol; here I will just give the specifics necessary for my experiments. For the λ-red machinery, I transformed the plasmid PTKRED (Addgene plasmid #41062)

[16] into electrocompetent *E. coli* MG1655 with a chromosomal *egfp/cat* resistance cassette (the forward evolution parent strain). To introduce the Δ25-26 TYMS mutation, I first recombined the tetA landing pad between genome positions 2,964,900 and 2,965,201 (genome NC000913). For the DHFR mutations (L28R, W30R, and P21L), the landing pad recombined between 49,684 and 49,990. Following landing pad insertion, cells were induced with 2mM IPTG and 0.4% arabinose, then transformed with 100ng of dsDNA PCR product containing the mutation of interest (with appropriate homology arms). This reaction underwent 3 days of outgrowth at 30°C in rich defined media (Teknova) with glucose substituted for 0.5% v/v glycerol. Media was supplemented with 6 mM or 4 mM $NiCl_2$ for counterselection against *tetA* at the *thyA* or *folA* locus respectively. The outgrowth culture was streaked onto agar plates and screened for the loss of tetracycline resistance daily (using LB supplemented with 50 μg/ml thy, 30 μg/ml spectinomycin, and +/- 5-10 μg/ml Tet). All genotypes were confirmed by Sanger sequencing of the complete *folA* and *thyA* open reading frame; for *folA* the promoter region was also sequenced.

3.5.5 Measurement of trimethoprim dose-response curves

All strains were grown overnight in LB + 5 μg/ml thymidine (thy), with the exception of those evolved in the 50 μg/ml thy condition, which were supplemented with 50 μg/ml thy to ensure viability. Overnight cultures were washed twice into M9 media supplemented with 0.4% glucose, 0.2% amicase, and the thymidine concentration matching their respective forward evolution condition (5, 10, and 50 μg/ml thy). Washed cells were back-diluted 1:10 then grown for 5.5 hours at 30°C. After adaptation, cultures were used to inoculate 96-well plates containing the same media along with serial dilutions of trimethoprim. Three replicates

were inoculated at a starting OD_{600} of 0.005 for each combination of strain and drug concentration. Optical density was monitored using a Tecan Infinite M200 Pro microplate reader and Freedom Evo robot at 30°C over a period of at least 12 hours.

3.5.6 IC50 estimation

Trimethoprim (TMP) resistance of each strain was quantified using its absolute IC50, which is the drug concentration (μg/ml) at which growth is half maximal. The relationship between growth and trimethoprim inhibition is modeled using the following four parameter logistic function:

$$Y = \frac{a - d}{1 + (X/c)^b} + d$$

where Y is growth, X denotes TMP concentration, a is the asymptote for uninhibited growth, d is the limit for inhibited growth, c provides the concentration midway between a and d, and b captures sensitivity [17]. The above model was fit to growth versus TMP concentration using MATLAB. Absolute IC50 is the concentration X^* for which growth $Y(X^*) = a/2$.

References

1. Lang, G.I. and M.M. Desai, *The spectrum of adaptive mutations in experimental evolution.* Genomics, 2014. **104**(6 Pt A): p. 412-6.
2. Hartfield, M. and S.P. Otto, *Recombination and hitchhiking of deleterious alleles.* Evolution, 2011. **65**(9): p. 2421-34.
3. King, C.H., D.M. Shlaes, and M.J. Dul, *Infection caused by thymidine-requiring, trimethoprim-resistant bacteria.* J Clin Microbiol, 1983. **18**(1): p. 79-83.
4. Kriegeskorte, A., et al., *Inactivation of thyA in Staphylococcus aureus attenuates virulence and has a strong impact on metabolism and virulence gene expression.* MBio, 2014. **5**(4): p. e01447-14.
5. Green, J.M. and R.G. Matthews, *Folate Biosynthesis, Reduction, and Polyglutamylation and the Interconversion of Folate Derivatives.* EcoSal Plus, 2013.
6. Toprak, E., et al., *Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition.* Nat Protoc, 2013. **8**(3): p. 555-67.
7. Toprak, E., et al., *Evolutionary paths to antibiotic resistance under dynamically sustained drug selection.* Nat Genet, 2012. **44**(1): p. 101-5.
8. Abdizadeh, H., et al., *Increased substrate affinity in the Escherichia coli L28R dihydrofolate reductase mutant causes trimethoprim resistance.* Phys Chem Chem Phys, 2017. **19**(18): p. 11416-11428.
9. Tamer, Y.T., et al., *High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection.* Mol Biol Evol, 2019. **36**(7): p. 1533-1550.
10. Mitosch, K., G. Rieckh, and T. Bollenbach, *Noisy Response to Antibiotic Stress Predicts Subsequent Single-Cell Survival in an Acidic Environment.* Cell Syst, 2017. **4**(4): p. 393-403 e5.
11. Bradley, M.D., et al., *Effects of Fis on Escherichia coli gene expression during different growth stages.* Microbiology, 2007. **153**(Pt 9): p. 2922-40.
12. Palmer, A.C., et al., *Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes.* Nat Commun, 2015. **6**: p. 7385.
13. Michener, J.K., et al., *Effective use of a horizontally-transferred pathway for dichloromethane catabolism requires post-transfer refinement.* Elife, 2014. **3**.
14. Tas, H., et al., *An Integrated System for Precise Genome Modification in Escherichia coli.* PLoS One, 2015. **10**(9): p. e0136963.
15. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-2492.
16. Kuhlman, T.E. and E.C. Cox, *Site-specific chromosomal integration of large synthetic constructs.* Nucleic Acids Res, 2010. **38**(6): p. e92.
17. Sebaugh, J.L., *Guidelines for accurate EC50/IC50 estimation.* Pharm Stat, 2011. **10**(2): p. 128-34.

Appendix 3

E. coli gene	Abbreviation	Name	Uniprot ID	Mutation	Description
folA	**DHFR**	dihydrofolate reductase	P0ABQ4	coding	Catalyzes the reduction of THF to DHF, an essential reaction for de novo glycine and purine synthesis, and for DNA precursor synthesis.
thyA	**TYMS**	thymidylate synthase	P0A884	coding	Catalyzes the reduction of dUMP to dTMP while utilizing 5,10-methylene THF as the methyl donor and reductant in the reaction, DHF as a by-product
dusB	**DUSB**	tRNA-dihydrouridine synthase B	P0ABT5	coding	Catalyzes the synthesis of 5,6-dihydrouridine via the reduction of the C5-C6 double bond of uridine on target tRNA.
cynR	**CYNR**	HTH-type transcriptional regulator CynR	P27111	coding	Positively regulates the cynTSX operon for cynate metabolism, and negatively regulates its own transcription.
ychE	**YHCE**	UPF0056 membrane protein YhcE	P25743	intergenic (/*oppA*)	Putative inner membrane protein.
oppA	**OPPA**	periplasmic oligopeptide-binding protein	P23843	intergenic (*ychE/*)	A component of the oligopeptide permease, a binding protein-dependent transport system.
gadX	**GADX**	HTH-type transcriptional regulator GadX	P37639	coding	Positively regulates the expression of about fifteen genes involved in acid resistance such as gadA, gadB and gadC. Depending on the conditions (growth phase and medium), can repress gadW.
otsB	**TPP**	Trehalose-6-phosphate phosphatase	P31678	intergenic (/*araH*)	Removes the phosphate from trehalose 6-phosphate (Tre6P) to produce free trehalose. Also catalyzes the dephosphorylation of glucose-6-phosphate (Glu6P) and 2-deoxyglucose-6-phosphate (2dGlu6P).
araH	**ARAH**	L-arabinose transport system permease protein AraH	P0AE26	intergenic (*otsB/*)	Part of the binding-protein-dependent transport system for L-arabinose. Probably responsible for the translocation of the substrate across the membrane.
yfbL	**YFBL**	uncharacterized protein YfbL	P76482	intergenic (/*yfbM*)	N/A
yfbM	**YFBM**	protein YfbM	P76483	intergenic (*yfbL/*)	N/A

E. coli gene	Abbreviation	Name	Uniprot ID	Mutation	Description
betI	**BETI**	HTH-type transcriptional regulator BetI	P17446	coding	Repressor involved in the biosynthesis of the osmoprotectant glycine betaine. It represses transcription of the choline transporter BetT and the genes of BetAB involved in the synthesis of glycine betaine.
ybaL	**YBAL**	Putative cation/proton antiporter YbaL	P39830	coding	Putative antiporter.
cat	**CAT**	chloramphenicol acetyltransferase		intergenic (/*eGFP*)	Chloramphenicol resistance marker introduced by phage transduciton.
eGFP	**EGFP**	enhanced green fluorescent protein		intergenic (*cat/*)	Enhanced green fluorescent protein introduced by phage transduction.
fis	**FIS**	DNA-binding protein Fis	P0A6R3	coding	Activates ribosomal RNA transcription, as well other genes. Plays a direct role in upstream activation of rRNA promoters. Binds to hundreds of transcriptionally active and inactive AT-rich sites.
lacI	**LACI**	lactose operon repressor	P03023	coding	Repressor of the lactose operon. Binds allolactose as an inducer.
chaA	**CHAA**	sodium-potassium/proton antiporter ChaA	P31801	coding	Sodium exporter that functions mainly at alkaline pH. Can also function as a potassium/proton and calcium/proton antiporter at alkaline pH.
csrA	**CSR**	carbon storage regulator	P69913	coding	A key translational regulator that binds mRNA to regulate translation initiation and/or mRNA stability, initially identified for its effects on central carbon metabolism.
ttdR	**TTDR**	HTH-type transcriptional activator TtdR	P45463	intergenic */ttdA*)	Positive regulator required for L-tartrate-dependent anaerobic growth on glycerol. Induces expression of the ttdA-ttdB-ygjE operon.
ttdA	**L-TTDα**	L(+)-tartrate dehydratase subunit alpha	P05847	intergenic *(ttdR/*)	Catalyzes the oxidation of (R,R)-tatrate to oxaloacetate.

Table 3.A1 Functional annotations for commonly mutated genes. Each of these genes were mutated at least 3 times across all conditions of the trimethoprim evolution experiment. Descriptions were paraphrased from UniProt.

CHAPTER FOUR

A Phylogenetically Aware Model of Positional Coevolution

4.1 Background and introduction

4.1.1 Understanding the constraints on protein sequence through models of coevolution

The previous chapters of my thesis explore the application of evolutionary statistical approaches in identifying multi-protein modules that shape the adaptation of cellular systems. The models of coevolution utilized in that work were based exclusively on a low-dimensional representation of each gene: their presence or absence (co-occurrence), and their position on the chromosome (synteny). However, the evolution of complex systems such as central metabolism can also occur through more subtle changes to the amino acid sequences of its proteins. For example, my case study of the metabolic enzymes dihydrofolate reductase (DHFR) and thymidylate synthase (TYMS) demonstrated how a constraint on metabolite concentration might drive reciprocal changes to the amino acid sequence. Nevertheless, even my own experiment involved only 300 generations of evolution, and selected for the complete inactivation of TYMS. These results represent a short period of selection and a relatively coarse change compared to substantial evolutionary record accessible through genomics. In order to map the functional constraints between interacting protein sequences in a way that is not dependent on choice of model organism or environment, it is necessary to define an appropriate model of inter-gene protein sequence coevolution.

Various methods have been developed for the statistical analysis of coevolution within individual protein families and between the sequences of interacting proteins [1-3]. Many of the efforts toward modeling inter-protein sequence coevolution have emphasized on the prediction of structural contacts between proteins that physically bind [4, 5]. By construction, these approaches are expected to "miss" coevolution between proteins that do not directly bind. In contrast, mirror-tree analyses have been shown to be a general indictor of functional couplings which are not limited to the case of physical interaction [6]. Similar to synteny and co-occurrence, this method produces an interaction score which quantifies the degree of evolutionary coupling between proteins families as single scalar value. Moreover, the mirror-tree analysis can incorporate an explicit model to control for the effect of phylogeny. In existing mirror-tree analyses, the site-specific origin of this coevolutionary signal is neglected. Developing a framework that can unify the analysis of residue-residue coevolution with the prediction of evolutionary interaction is an open problem, which I explore in this chapter.

4.1.2 Basic principles of the mirror-tree analysis

Mirror-tree is a coevolutionary analysis which uses information in the coding sequence to infer interactions between protein families. It was introduced by Pazos and Valencia as a tool for predicting which combinations of proteins engage in a macromolecular complex [7]. More recent work applying the principals of mirror-tree has broadened its biological scope. Using a derivative method called evolutionary rate covariation (ERC), Clark and colleagues demonstrated a similar magnitude of sequence coevolution between physically binding proteins and those that only interact genetically [6]. Many instances of

sequence coevolution in the absence of known binding come from metabolic and biosynthetic pathways. For example, three enzymes in galactose metabolism exhibit some of the strongest coevolutionary signal in the entire analyzed proteome (Gap1p, Gal7p, and Gal10p; >4000 proteins). These proteins are analogous to my case study of dihydrofolate reductase (DHFR) and thymidylate synthase (TYMS) in that they catalyze sequential steps of their respective pathway and are collinear on the chromosome. Clark et al report that elevated ERC signal significantly overlaps with correlated codon adaptation and thus co-expression [6]. The potential implication of gene-synteny as a second mode of coevolution between galactose enzymes supports the notion that a single adaptive interaction may imprint itself on multiple evolutionary observables.

While the application and technical details of mirror-tree methods have developed over time, the basic premise remains the same. In the mirror-tree analysis, protein families are represented by their sequence similarity matrix. This two-dimensional array is computed from a multiple sequence alignment (MSA), and describes the percent identity of the amino acid sequence across all pairs of species that contain it. The information in a similarity matrix provides the basis for estimating phylogenetic trees [8]. In fact, the only difference between base mirror-tree and ERC lies in imposing a tree structure on the data and explicitly estimating branch lengths. The objective of mirror-tree methods is to statistically assess whether a pair of protein families is undergoing sequence change across the same pairs of species or branches of a phylogenetic tree. The mechanistic origin of this signal has been debated, but it’s thought to arise due to both site-specific constraints and a shared fluctuation of selective pressures between coupled proteins [9-11].

The first step of computing the mirror-tree score between a pair of protein families is to filter their sequences as to only include species that contain both orthologs. For simplicity of explanation, the case where multiple paralogous sequences exist for in the same species will be neglected here (see 4.6.5 Materials and methods for details). S^A represents the MxM sequence similarity matrix for family A as defined across a set of M species. The upper-triangle of S^A contains $M_{\text{pair}} = 0.5M(M-1)$ total elements where each possible pair of species is represented exactly once. Since the object of mirror-tree is to compare identity change across species, this upper triangle is reshaped into a one-dimensional vector denoted in $|s^A\rangle$ bra-ket notation. The base implementation of mirror-tree computes the Pearson correlation between two such similarity vectors. I will denote the standard score as $|\hat{s}^A\rangle = \left(|s^A\rangle - |\mu_{s^A}\rangle\right)/\sigma_{s^A}$, where μ_{s^A} and σ_{s^A} are the mean and standard deviation of $|s^A\rangle$ respectively. The term $|\mu_{s^A}\rangle$ is used to represent a vector with the same dimensions as $|s^A\rangle$ where every element is equal to the mean. The Pearson correlation between two families A and B can then be written using the following inner product:

$$r^{AB} = \frac{\langle \hat{s}^A | \hat{s}^B \rangle}{M_{\text{pair}}}$$

By definition, this value ranges from -1 to 1 and captures the magnitude and direction of linear association between the two similarity vectors. Negative values indicate that one protein family is being conserved while the other is varying, and vice versa. This is an extreme case and not expected to occur in the base implementation of mirror-tree due to the phylogenetic relationship of the samples. Positive values occur when the conservation or variation of one protein family is coordinated with the other. Positive values with the highest

magnitude are taken as an indicator of functional coupling, Pazos and Valencia used this measure to predict physical interactions in their 2001 study of the *E. coli* proteome with modest success [7].

4.1.3 Incorporating phylogenetic information into mirror-tree

By using Pearson correlation, each species pair is treated as a Bernoulli sample (independent trial). Since individual species are related through the structure of a phylogenetic tree, this choice trades off accuracy for simplicity. Sato and colleagues demonstrate a way to incorporate phylogenetic information into the model without compromising linearity [12, 13]. They accomplished this using the statistical tool of partial correlation. Conceptually, this amounts to measuring the association between two variables while controlling for any confounding factor(s). For the purpose of this work I will focus on the use of phylogeny as the only control variable, although Sato and colleagues also consider a higher-order partial correlation where association between proteins A and B is measured while controlling for mutual dependence on all other protein families in in the dataset ($N-2$ in an analysis of N proteins) [14]. As with individual protein families, phylogenetic information is encoded in an MxM similarity matrix defined across the same set of Mspecies as the considered protein pair. A number of numerical models of phylogeny may be used, including a simple average sequence similarity across all analyzed protein families contained in those Mspecies. Other examples include the sequence similarity of 16s ribosomal RNA as well as housekeeping genes (such as the subunits of RNA polymerase), which have been shown to be a stable indictor of genome divergence across species [15]. Each of the phylogenetic models considered by Sato et al were shown to improve performance over base

mirror-tree [12, 13]. The upper-triangle of the phylogenetic similarity matrix P is reshaped into a one dimensional vector which I'll denote $|p\rangle$. To compute partial correlation, one first applies the following operation to each protein family:

$$|\varepsilon^A\rangle = |s^A\rangle - \frac{\langle s^A|p\rangle}{\||p\|^2}|p\rangle$$

Geometrically, this operation obtains the component of $|s^A\rangle$ that is orthogonal to phylogeny $|p\rangle$ by subtracting its projection; the projection onto $|p\rangle$ along with the orthogonal component form a right triangle where $|s^A\rangle$ is the hypotenuse. The Pearson correlation between orthogonal components $|\varepsilon^A\rangle$ and $|\varepsilon^B\rangle$ is then computed from an inner product of their standard scores $|\hat{\varepsilon}^A\rangle$ and $|\hat{\varepsilon}^B\rangle$ as above. This value constitutes the partial correlation $r^{AB\cdot p}$ between families A and B with the effect of $|p\rangle$ removed, and provides a more effective indicator of functional relationships.

4.2 Derivation of positional mirror-tree

Canonical mirror-tree is a coarse-grained measure of coevolution which condenses the data of multiple sequence alignments into a single interaction score. Previous studies have looked at origin of the mirror-tree signal with respect to the protein sequence, albeit not to the resolution of mapping all sites individually. Kann and colleagues investigated the role of positions located in the binding interface of physically interacting proteins [10]. While their results indicate that sequence coevolution is enriched at the binding interface, the rest of each protein still carried a significant signal of coevolution. Congruently, Hakes et al found that restricting their mirror-tree analysis of yeast proteins to the relevant surface or binding

sites did not lead to any improvement in interaction prediction [9]. However, these only represent one possible view of how site-specific interactions might underlie the full-length mirror-tree score. Important interactions need not be constrained to the surface or binding interface of two proteins. Indeed, work on predicting histidine kinase-response regulator interaction partners found that residue pairs carrying the signal of sequence coevolution were not restricted to those that are in direct contact at the interface [16].

In order to construct a mapping from the full-length mirror-tree tree score to the individual positions of each alignment, one begins with the similarity vector which is the building block of the analysis. The sequence similarity vector of a given protein family A can be represented as an average across all of the positions in the protein sequence:

$$|s^A\rangle = \frac{1}{L_A}\sum_{i=1}^{L_A}|{s_i}^A\rangle$$

In this expression, index i enumerates the positions in an aligned protein sequence of length L_A. Congruently, $|{s_i}^A\rangle$ is a binary vector which contains a 1 if the amino acid identity at position i is conserved in a given species pair and a 0 if it is not. The next ingredient in mapping mirror-tree to the resolution of site-specific contributions is to note that the inner product used used to compute $|\varepsilon^A\rangle$ and $r^{AB}{}_{\text{partial}}$ is bilinear. This means that the function $\langle x|y\rangle$ is linear with respect to its inputs x and y. Thus, $|\varepsilon^A\rangle$ can also be understood as an average across the positions of the alignment.

$$|\varepsilon^A\rangle = \frac{1}{L_A}\sum_{i=1}^{L_A}\left(|{s_i}^A\rangle - \frac{\langle {s_i}^A|p\rangle}{\|p\|^2}|p\rangle\right)$$

I will analogously refer to the individual terms of this sum as $|{\varepsilon_i}^A\rangle$. If $|{s_i}^A\rangle$ records the transitions of position i across species, then $|{\varepsilon_i}^A\rangle$ captures the degree to which that pattern of transitions deviates from the phylogenetic relationships between species. Moving forward to the mirror-tree score itself, one can make the following substitution:

$$r^{AB\cdot p} = \frac{\left\langle \varepsilon^A - \mu_{\varepsilon^A} \middle| \varepsilon^B - \mu_{\varepsilon^B} \right\rangle}{M_{\text{pair}} \sigma_{\varepsilon^A} \sigma_{\varepsilon^B}}$$

$$= \frac{1}{L_A L_B M_{\text{pair}} \sigma_{\varepsilon^A} \sigma_{\varepsilon^B}} \sum_{i=1}^{L_A} \sum_{j=1}^{L_B} \left\langle {\varepsilon_i}^A - \mu_{{\varepsilon_i}^A} \middle| {\varepsilon_j}^B - \mu_{{\varepsilon_j}^B} \right\rangle$$

While $\mu_{{\varepsilon_j}^A}$ refers to the mean of the position-specific $|{\varepsilon_i}^A\rangle$, it should be noted that the standard deviation σ_{ε^A} is not linear with respect to the positions and is thus computed across the full length of the alignment. However, the summation term of the expanded equation is equal to the covariance of positional vectors $|{\varepsilon_i}^A\rangle$ and $\left|{\varepsilon_j}^B\right\rangle$. Thus, the mirror-tree score is composed of a normalized sum across all possible covariances between the positions of alignments A and B. Based on the equations above, I defined the following positional coevolution matrix.

$${C_{ij}}^{AB\cdot p} = \frac{1}{M_{\text{pair}} \sigma_{\varepsilon^A} \sigma_{\varepsilon^B}} \left\langle {\varepsilon_i}^A - \mu_{{\varepsilon_j}^A} \middle| {\varepsilon_j}^B - \mu_{{\varepsilon_j}^B} \right\rangle$$

The dimensions of this matrix are $L_A x L_B$, and the average across its elements is exactly equal to the canonical mirror-tree score. As such, this simple series of substitutions is enough to produce a linear mapping between the full-length mirror-tree score and scaled positional covariances. This work provides roadmap for studying sequence coevolution between proteins at the resolution of individual positions using a framework that has already been

verified to predict diverse functional relationships. However, it also invites the question of whether position-resolution information can be used to improve the performance of mirror-tree as a whole.

4.3 Application to a focused test set

In order to illustrate my implementation of positional mirror-tree and make initial observations, it was necessary to define a small collection of proteins to serve as a focused test set. For this purpose, I elected to focus on proteins that are known to physically interact, which is the most well-studied context of inter-protein residue-residue coevolution [4, 5, 17-19]. I chose a set of 4 protein complexes from *E. coli* which were referenced by either our analysis of gene synteny or the work by Ovchinnikov and colleagues applying direct coupling analysis to physically-interacting proteins [4]. This test set, with gene names shown in italics, consists of the following complexes:

1. Cytochrome bd oxidase subunits *cydA* and *cydB* [20]
2. The flagellar motor switch *fliG*, *fliM,* and *fliN* [21]
3. Ribonucleoside-diphosphate reductase subunits *nrdA* and *nrdB* [22]
4. Tryptophan synthase which consists of *trpA* and *trpB* [23]

As a computational control, I used our existing analyses to produce maps of coevolution in this test set according to synteny and co-occurrence (Figure 4.1A-B). Each of the complexes are easily resolvable by both of our coarse-grained statistical measures, with one exception being that the *nrdA*/*nrdB* gene pair only produces a weak co-occurrence signal. Next, I implemented a phylogenetically-corrected full-length mirror-tree method, and applied it to

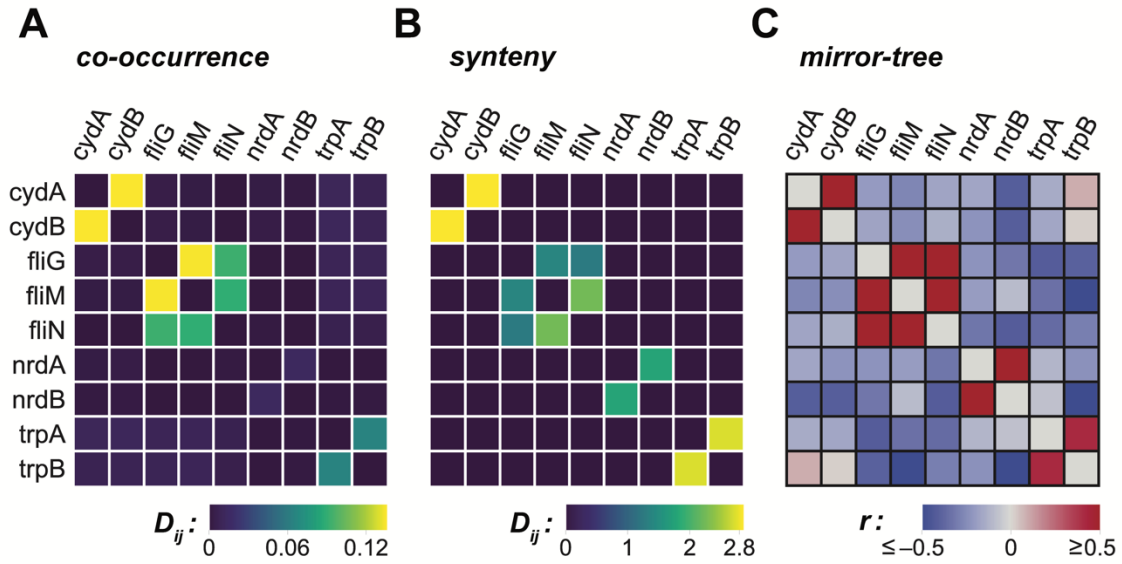

Figure 4.1 Statistical coevolution in a test set of 4 physical complexes. *E. coli* gene names are labeled along the left and top of axes of the figure. Gene products with a known physical interaction share a prefix, such as *cyd-*. **A-B** Statistical coevolution according to analyses of gene co-occurrence and synteny computed across 1445 complete bacterial genomes. Coupling between gene pairs in the test set is reported as a relative entropy D_{ij}^{intra}, shown by pixel intensity. **C,** Statistical coevolution between amino acid sequences according to mirror-tree. Coupling between gene products is reported as a partial correlation, where the effect of phylogeny has been excluded. Positive values indicate coevolution orthogonal to the known phylogenetic relationships among the samples (species). Values along the diagonal would be equal to 1 by definition, so these are excluded from each heatmap

these complexes (Figure 4.1C). The result is a signed heatmap of coevolution. Negative values show that a pair of protein families exhibits less correlation in sequence change than would be expected from the phylogenetic relationships among the samples. These can be interpreted as indicating that two proteins undergo sequence change between distinct branches of the phylogenetic tree (even though a tree is not being explicitly constructed). Negative partial correlation values are observed between almost every pair of non-interacting proteins in the dataset. Positive values indicate that two proteins exhibit correlation orthogonal to phylogeny, which would constitute repeated events of coevolution across

parallel branches of the phylogenetic tree. These are taken as ‘true’ coevolution and a prediction of evolutionary coupling. The results of the full-length mirror-tree analysis agree well with synteny, co-occurrence, and the known functional relationships among these test proteins.

In this work, I derived a mapping from the full-length mirror-tree signal onto the individual positions of the amino acid sequence alignment. When applied to a pair of protein families, this produces a signed residue-level coevolution matrix that describes the degree of

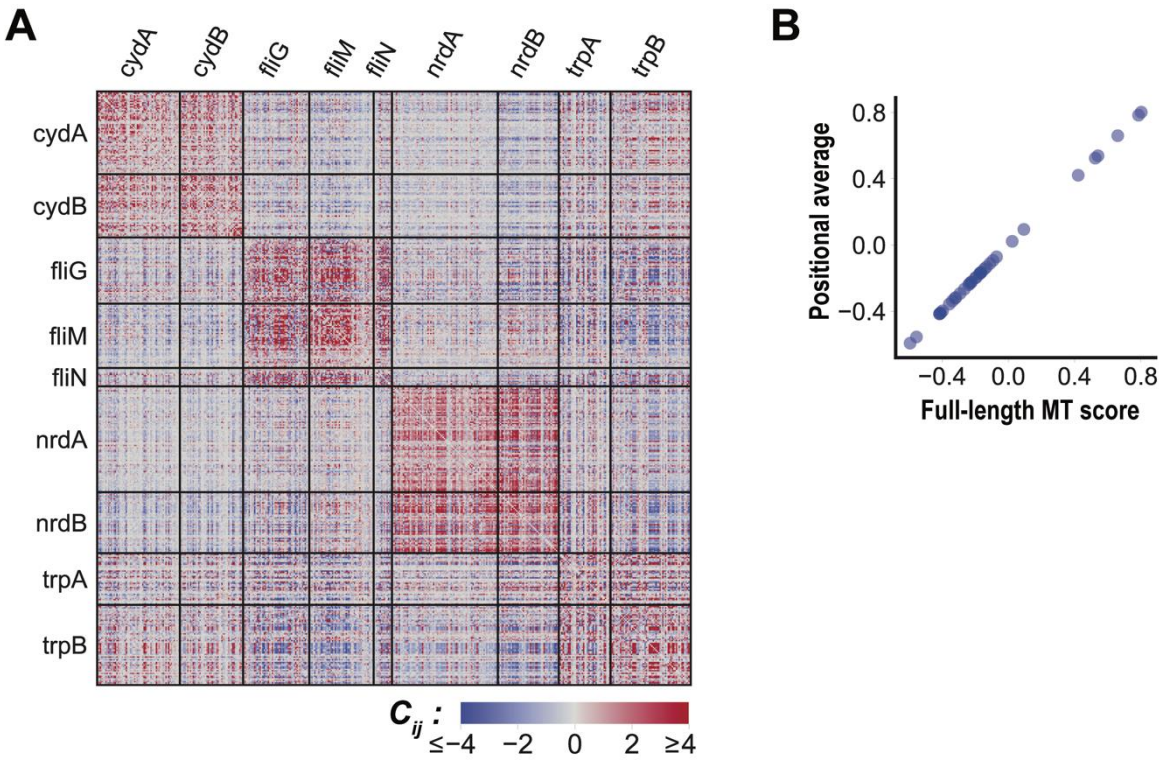

Figure 4.2 Intra- and inter- protein coevolution according to positional mirror-tree. A, Positional mirror-tree matrices for all individual proteins and protein pairs in my test set of 4 complexes. Matrix blocks corresponding to each protein and protein pair are separated by black dividers. Gene names are labeled across the top and left of the heatmap. The values $\boldsymbol{C_{ij}}$ constitute scaled positional covariances computed using the mirror-tree framework. Positive values indicate evolutionary coupling. **B,** A reconstruction of the full-length mirror-tree score for all protein pairs. The x-axis displays the full-length mirror-tree score computed for each protein pair in the dataset (Figure 4.1C). The y-axis displays an average taken across all values in the corresponding positional mirror-tree matrix. These data illustrate that the full-length mirror-tree interaction score can be exactly and additively reconstructed from information in the positional matrices.

evolutionary coupling between all pairs of positions across the two alignments. The interpretation of the sign of these values is the same as in the case of the full-length mirror-tree analysis; positive values indicate evolutionary coupling beyond the expectation from phylogeny. I computed both intra- and inter- protein positional coevolution using this method for all of the proteins in my toy system (Figure 4.2A). The signal within each resulting positional coevolution matrix is heterogeneous. In the context of interacting proteins, this means that not all pairs of positions carry the signal of evolutionary coupling equally. As described before, the values in each positional coevolution matrix constitute scaled positional covariances and are distributed around the full-length mirror-tree interaction score. In Figure 4.2B, I demonstrate that the average value of each positional coevolution matrix can be used to exactly reconstruct the full length mirror-tree scores within my dataset. This framework will provide a basis for the future study of positional coevolution between interacting protein sequences.

4.4 Comparison of positional mirror-tree to statistical coupling analysis

How does positional mirror-tree compare to existing measures of protein sequence coevolution? Potential choices for comparison include statistical coupling analysis (SCA) and direct coupling analysis (DCA). The SCA positional coevolution matrix is derived from covariance in amino acid identity and weighted by single-site conservation [24], while DCA fits amino acid covariance terms to an underlying model assuming pairwise positional interaction [2]. These operations are computed across species rather than species pairs. As a result, neither SCA nor DCA can easily accommodate a linear model of phylogenetic information. Since positional mirror-tree is also based on the direct observation of

covariance, SCA is the more natural comparison. I computed the SCA measure of coevolution for all inter-protein position pairs in my test set. These constitute the SCA-version of the off-diagonal blocks in my positional mirror-tree matrix (Figure 4.2A). I plotted SCA values against mirror-tree on two-dimensional axes for 6 of the total 36 protein pairs (Figure 4.3). Because each inter-protein coevolution matrix contains on the order of 10^4-10^5 values, the data are visualized as a kernel density estimate (created using Seaborn v0.9.0). In these plots, shading indicates the density of points (position pairs) in two-dimensional space. To quantify the linear association between SCA and mirror-tree, I used Pearson correlation (r-values and best-fit lines shown in black). The resulting r-values range between 0.4-0.7, indicating a direct linear relationship. Because positional mirror-tree incorporates phylogenetic information to produce a signed measure of coevolution, some disagreement is expected. For methods such as SCA utilizing single-site and joint amino acid frequencies, positional covariance due to phylogeny is not directly discernable from "true" coevolution [25]. Thus, one would expect some points to exhibit high SCA values but non-positive mirror-tree values. When I restrict my correlation analysis to only consider position pairs with positive positional mirror-tree scores, the linear association between SCA and positional mirror-tree increases to approximately 0.8 for each protein pair (Figure 4.3; orange text and best-fit lines). These results show that despite vast mathematical differences between the two methods, position pairs that coevolve strongly according to mirror-tree are also likely to coevolve according to SCA.

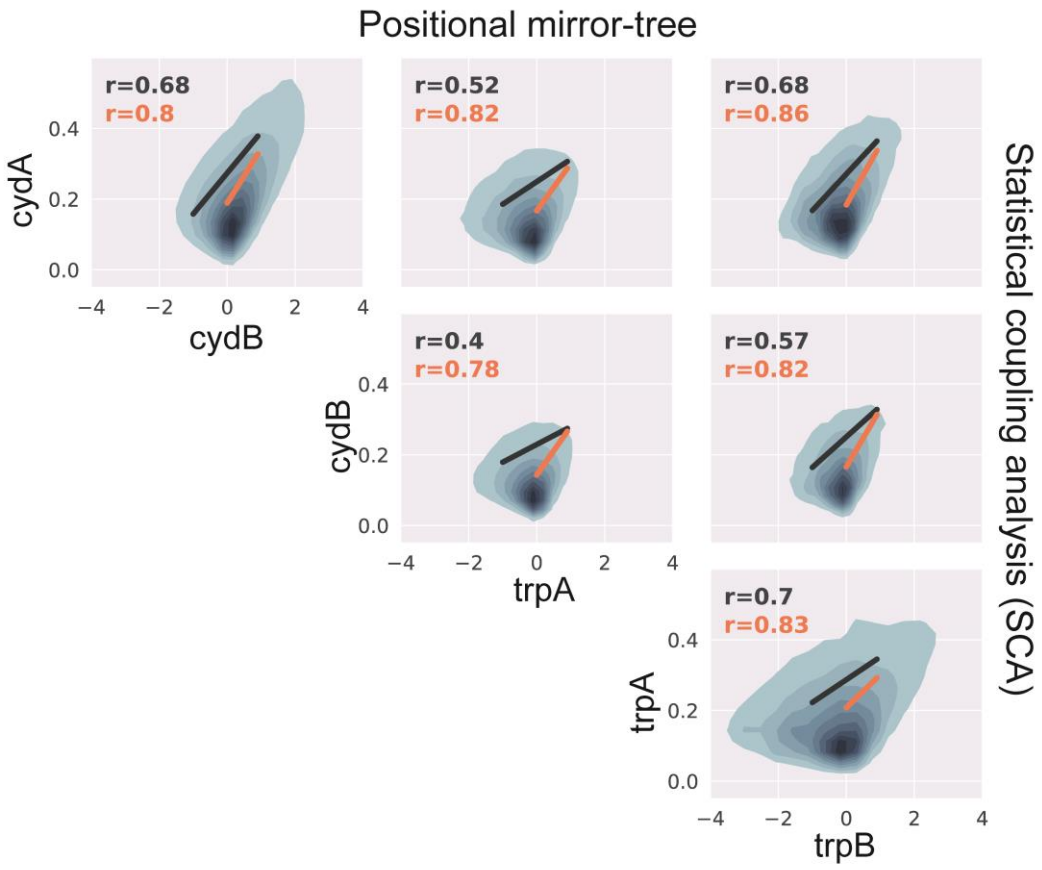

Figure 4.3 Statistical comparisons between SCA and positional mirror-tree. Each plot represents the inter-protein coevolution matrix for a single protein pair according to positional mirror-tree (x-axis) versus SCA (y-axis). Because each inter-protein coevolution matrix contains on the order of 10^4-10^5 position pairs, the data are displayed as a kernel density estimate (sklearn v0.9.0). The depth of shading within each contour indicates the density of data points encompassed within. Pearson correlation and a best fit line were computed for all the data (black) and restricting to the case where positional mirror-tree is positive (orange).

4.5 Distinct communities of coevolving positions in the mirror-tree matrix

A key result in the statistical coupling analysis of individual protein families is the identification of collectively coevolving communities of amino acids, termed ‘sectors’ [1, 24]. The identification of sectors through statistical analysis has provided insights as to how specific positions in the protein structure mediate properties such as allostery [26], substrate specificity [27], and thermodynamic stability [1]. One particular work on the S1A

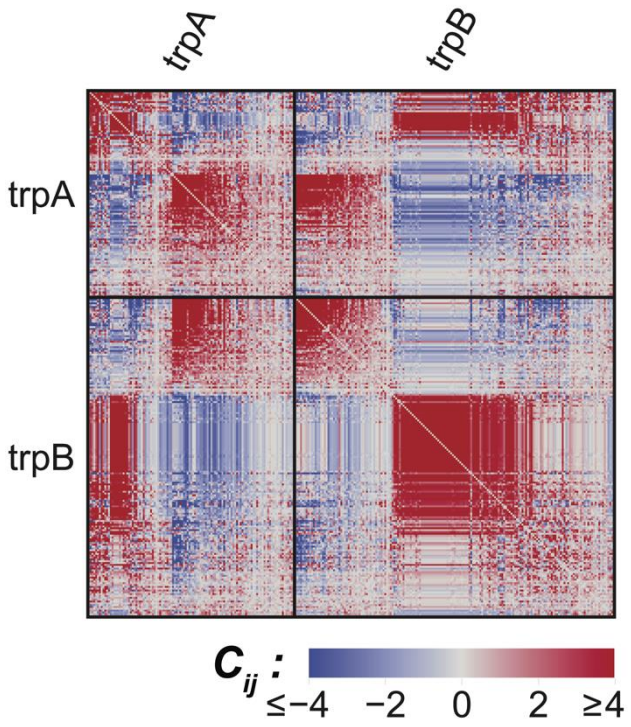

Figure 4.4 Distinct collections of coevolving positions in the *trpA* and *trpB* sequence alignments. I used agglomerative clustering to identify 2 distinct coevolving communities of positions within each of the *trpA* and *trpB* intra-protein coevolution matrices (diagonal blocks). Data presented constitute the positional mirror-tree matrices of the *trpA*/*trpB* pair with the axes (positions) having been sorted according to cluster identity. Positions exhibit positive evolutionary coupling to other positions in the same cluster, but are negatively coupled between clusters. Similarly, each cluster identified within *trpA* demonstrates positive evolutionary coupling with one cluster from *trpB*, and is decoupled from the other.

superfamily of serine proteases identified multiple independent sectors within the same protein, each associated with a distinct aspect of protein function [1]. The statistical association observed between SCA and positional mirror-tree motivates the question of whether distinct coevolving communities of positions can be resolved from within positional mirror-tree matrices as well. I applied agglomerative clustering (sklearn v0.20) to the intra-protein positional mirror-tree matrices of tryptophan synthase subunits *trpA* and *trpB*. Both intra- and inter- protein coevolution matrices are plotted as before, but positions have been

sorted based on cluster (Figure 4.4). Much like independent SCA sectors, the data indicate two distinct clusters of positions exhibiting modular coevolution. Positional mirror-tree scores are largely positive among positions within the same cluster and negative between clusters. Despite the fact the clusters were defined based only on intra-protein coevolution, the same pattern persists in the inter-protein coevolution matrix between interacting subunits encoded by *trpA* and *trpB*. Each cluster of *trpA* is positively coupled to exactly one cluster from *trpB* and negatively coupled to the other. These findings raise questions about whether community detection is relevant to the analysis of inter-protein coevolution as well: can the community detection be used to improve our capacity to predict adaptive couplings using mirror-tree? Further work is needed in order to better understand the prevalence of distinct coevolving communities within positional mirror-tree matrices, and their potential biological or mechanistic implications.

4.6 Materials and Methods

4.6.1 Software and data analysis

Data analysis was performed with custom scripts written with the AnacondaV2.4.0 data science distribution of Python 2.7 [28]. Some of the code for alignment processing was adapted from the Python-based Statistical Coupling Analysis software package (pySCA) [24].

4.6.2 Multiple sequence alignment generation and preprocessing

Alignments were acquired from eggNOG, a public resource of orthologous groups at different taxonomic levels [29]. The alignments used in this work were obtained from the

'trimmed' tab for the eubacterial taxa of each orthologous gene group. Trimmed alignments from eggNOG have already been filtered to remove highly gapped positions. Before analysis, alignments were further processed to remove highly gapped positions and sequences. First, positions with a gap frequency over 20% were filtered. Sequences whose proportion of gapped positions remained greater than 30% after the initial filtering step were subsequently removed from the alignment. Highly gapped sequences provide less information about amino acid coevolution and can indicate an incorrectly identified paralog.

4.6.3 Estimating phylogenetic similarity

Estimation of phylogenetic similarity between species pairs was conducted using an all-protein average, based on work by Sato et al [12]. I computed the average sequence similarity across all proteins in the test dataset represented in a given species pair. In order to ensure that this average encompassed some protein families that did *not* interact, I restricted my dataset to only include species that contained at least 5 of the 9 total proteins. The all-protein average is intended to estimate genome similarity and thus the phylogenetic relationships among species pairs. This information was incorporated into the mirror-tree analysis as the $|p\rangle$ vector.

4.6.4 Empirical down-weighting of redundant species

The available protein sequences and complete genomes do not constitute an even or random sampling of bacterial phyla. To control for the effect of oversampling particular clades, I implemented a reciprocal species weighting scheme. This strategy was first applied to direct coupling analysis of multiple sequence alignments by Morcos et al [2], and has since adopted

by other theoretical frameworks [24]. The basic idea is to combine redundant species into a single *effective* species. The species weight w_g for a given species g is defined:

$$w_g = \frac{1}{\left|\{h: P_{gh} > (1-\delta)\}\right|}$$

The expression in the denominator indicates the total number of species h whose phylogenetic similarity to g is above the threshold $1-\delta$. The work presented here uses $\delta = 0.1$, meaning that species with phylogenetic similarity above 90% are combined. Utilization of species weights renders a new definition of sample size, referred to here as the number of effective species. The sample size for a pair of protein families is simply the sum of all species weights for species that contain an ortholog of both proteins, denoted:

$$M_{eff} = \sum w_g$$

I wished to apply this principle to the sequence similarity vector, where dimensions represent species pairs rather than individual species. The natural generalization of the above heuristic is to weight each species pair g, h by the product of its respective individual species weights $w_g w_h$. This results in the analogous number of effective species pairs which is $0.5 M_{eff}(M_{eff}-1)$. The weighting scheme was applied to all dimension-reduction operations computed across species pairs, such as the average or inner product of similarity vectors. The result is that the effect of overrepresented species is numerically mitigated.

4.6.5 Treatment of paralogous sequences

Some proteins in the test dataset contained more than one paralog in a given species. With respect to species weighting, these sequences are considered as within a phylogenetic distance of $\delta = 0$ from one another. When computing the mirror-tree score between two

protein families, the dimensions of their respective similarity vectors need to match exactly. This is accomplished by filtering the set of species in each alignment down to those that contain both proteins. Paralogous sequences complicate this process slightly because multiple rows of the alignment can correspond to a single species. For illustration, consider the case where protein family A contains 3 sequences in a given species, while protein family B contains only 2 paralogs. The objective is to filter each alignment such that corresponding rows are drawn from the exact same species. Because my test set only concerns protein families that interact through physical binding, I assumed a 1:1 matching between paralogs across protein families. Since interacting proteins are often found in the same operon, I matched each paralogous sequence from family B to the sequence from A that was closest on the chromosome. For simplicity, I used the numerical gene IDs contained in each eggNOG alignment as a proxy for chromosomal proximity, since these indicate the ordering of genes on the chromosome. As a result, each alignment row from family B was paired with exactly one sequence from family A, and the remaining 'orphan' sequence from family A was dispensed of for the purpose of this comparison. There are a number of other potential implementations that could handle this slight complexity, but given the rarity of paralogs in my test set it is expected to have a minimal effect on the results.

References

1. Halabi, N., et al., *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-86.
2. Morcos, F., et al., *Direct-coupling analysis of residue coevolution captures native contacts across many protein families.* Proc Natl Acad Sci U S A, 2011. **108**(49): p. E1293-301.
3. de Juan, D., F. Pazos, and A. Valencia, *Emerging methods in protein co-evolution.* Nat Rev Genet, 2013. **14**(4): p. 249-61.
4. Ovchinnikov, S., H. Kamisetty, and D. Baker, *Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information.* Elife, 2014. **3**: p. e02030.
5. Croce, G., et al., *A multi-scale coevolutionary approach to predict interactions between protein domains.* bioRxiv, 2019: p. 558379.
6. Clark, N.L., E. Alani, and C.F. Aquadro, *Evolutionary rate covariation reveals shared functionality and coexpression of genes.* Genome Res, 2012. **22**(4): p. 714-20.
7. Pazos, F. and A. Valencia, *Similarity of phylogenetic trees as indicator of protein-protein interaction.* Protein Eng, 2001. **14**(9): p. 609-14.
8. Fitch, W.M. and E. Margoliash, *Construction of phylogenetic trees.* Science, 1967. **155**(3760): p. 279-84.
9. Hakes, L., et al., *Specificity in protein interactions and its relationship with sequence diversity and coevolution.* Proc Natl Acad Sci U S A, 2007. **104**(19): p. 7999-8004.
10. Kann, M.G., et al., *Correlated evolution of interacting proteins: looking behind the mirrortree.* J Mol Biol, 2009. **385**(1): p. 91-8.
11. Lovell, S.C. and D.L. Robertson, *An integrated view of molecular coevolution in protein-protein interactions.* Mol Biol Evol, 2010. **27**(11): p. 2567-75.
12. Sato, T., et al., *The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.* Bioinformatics, 2005. **21**(17): p. 3482-9.
13. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein–protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-2492.
14. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-92.
15. Zeigler, D.R., *Gene sequences useful for predicting relatedness of whole genomes in bacteria.* Int J Syst Evol Microbiol, 2003. **53**(Pt 6): p. 1893-900.
16. Bitbol, A.F., et al., *Inferring interaction partners from protein sequences.* Proc Natl Acad Sci U S A, 2016. **113**(43): p. 12180-12185.
17. Hopf, T.A., et al., *Sequence co-evolution gives 3D contacts and structures of protein complexes.* Elife, 2014. **3**.
18. Feinauer, C., et al., *Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon.* PLoS One, 2016. **11**(2): p. e0149166.

19. Gueudre, T., et al., *Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis.* Proc Natl Acad Sci U S A, 2016. **113**(43): p. 12186-12191.
20. Giuffre, A., et al., *Cytochrome bd oxidase and bacterial tolerance to oxidative and nitrosative stress.* Biochim Biophys Acta, 2014. **1837**(7): p. 1178-87.
21. Delalez, N.J., et al., *Signal-dependent turnover of the bacterial flagellar switch protein FliM.* Proc Natl Acad Sci U S A, 2010. **107**(25): p. 11347-51.
22. Yokoyama, K., U. Uhlin, and J. Stubbe, *Site-specific incorporation of 3-nitrotyrosine as a probe of pKa perturbation of redox-active tyrosines in ribonucleotide reductase.* J Am Chem Soc, 2010. **132**(24): p. 8385-97.
23. Weyand, M. and I. Schlichting, *Crystal structure of wild-type tryptophan synthase complexed with the natural substrate indole-3-glycerol phosphate.* Biochemistry, 1999. **38**(50): p. 16469-80.
24. Rivoire, O., K.A. Reynolds, and R. Ranganathan, *Evolution-Based Functional Decomposition of Proteins.* PLoS Comput Biol, 2016. **12**(6): p. e1004817.
25. Maddison, W.P. and R.G. FitzJohn, *The unsolved challenge to phylogenetic correlation tests for categorical characters.* Syst Biol, 2015. **64**(1): p. 127-36.
26. Reynolds, K.A., et al., *Evolution-based design of proteins.* Methods Enzymol, 2013. **523**: p. 213-35.
27. Raman, A.S., K.I. White, and R. Ranganathan, *Origins of Allostery and Evolvability in Proteins: A Case Study.* Cell, 2016. **166**(2): p. 468-480.
28. *Anaconda Software Distribution*. 2015.
29. Huerta-Cepas, J., et al., *eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences.* Nucleic Acids Res, 2016. **44**(D1): p. D286-93.

CHAPTER FIVE

Conclusions and Future Directions

5.1 A two-enzyme adaptive unit in bacterial folate metabolism

In my case study of bacterial folate metabolism, comparative genomic analyses of gene synteny and co-occurrence suggested a sparse pattern of evolutionary coupling in which most enzymes in the pathway do not coevolve. Several groups of 2-3 proteins exhibited strong evolutionary coupling with one another but remained decoupled from the rest of the pathway. A notable example is the enzyme pair dihydrofolate reductase (DHFR) and thymidylate synthase, which catalyze sequential biochemical steps but are not known to physically interact. Experimental analyses of *E. coli* support the interpretation that these two enzymes behave as a relatively independent adaptive unit. The activities of DHFR and TYMS are coupled by a shared constraint on metabolite concentrations. Reducing DHFR activity causes an accumulation of the toxic intermediate dihydrofolate (DHF) and a fitness defect, both of which can be abrogated by the reciprocal inactivation of TYMS. Under conditions of trimethoprim stress, adaptation was driven by a combination of mutations to DHFR and TYMS without requiring compensatory modification to other folate genes. Additional experimental data collected by our lab supports the idea that DHFR and TYMS are more tightly coupled to one another than any other enzymes in the pathway [1]. These results expose the potential for an intermediate level of organization in cellular systems in between the scale of individual genes and the entire pathway. Further, they motivate careful interpretation of

not only the presence of evolutionary coupling between genes, but its absence as well. It becomes less obvious that membership in the same KEGG pathway is a specific enough target for interaction prediction approaches; instead, the goals of comparative genomics and the type of interactions being predicted (shared biochemical intermediate, physical complex, and/or adaptive interaction) should be carefully considered.

DHFR and TYMS represent a first case study, and, by construction, comparative genomic analyses are not expected to predict all possible interactions. Analysis of synteny and co-occurrence are computed from an average across thousands of species, encapsulating hundreds of millions of years of evolutionary divergence. Idiosyncratic interactions that are particular to a specific environment (including those never encountered in evolutionary) or model organism are also expected to exist. This nuance is illustrated by recent, not-yet published work by Joao Rodrigues and Eugene Shakhnovich [2]. Very much analogous to my own study, the authors engineered a reduction of function into DHFR and then adapted *E. coli* to higher levels of fitness while titrating molecular products of the folate pathway out of the medium. Expectedly, a number of trials resulted in a reversion of the deleterious mutation to DHFR, which is a trivial solution to this evolutionary problem. All other repeats in the experiment were dominated by a loss of TYMS function as the first step of adaptation. However, due to the fact that their media contained thymine instead of thymidine, this was followed by the inactivation of a second locus. Phosphopentomutase, encoded by *deoB,* is an enzyme involved in the thymine salavage pathway. By inactivating *deoB*, the salvage of exogeneous thymine is diverted away from glycolysis and toward the production of dTMP. Thus, the inactivation

of *deoB* only becomes relevant after TYMS acquires a loss-of- function and is contingent on this particular environmental condition. We expect that the adaptive units predicted by synteny and co-occurrence, such as DHFR and TYMS, represent well-conserved, core interactions that can sometimes be elaborated on under particular environmental conditions or in particular species. We propose that if perturbation is made to one gene of the adaptive unit, the first and most common adaptive mutations will also occur within the unit (whether in the directly affected gene or elsewhere). Further, mutations within the unit should largely suffice to restore function, with mutations outside the pair having more subtle or idiosyncratic effects.

If it is generally possible to decompose metabolic pathways into smaller, adaptive subunits, then this would suggest a route to identify building blocks for biosynthetic engineering. Such adaptive units could provide fundamental insights as to how cells maintain homeostasis and evolve in the face of changing environments. For example, thymidine synthesis is the rate-limiting step for DNA synthesis of eukaryotic cells. Transcription of the DHFR and TYMS genes is greatly upregulated through a common transcription factor at the G_1/S cell cycle transition [3]. *In silico* modeling of eukaryotic folate metabolism reveals that computationally increasing the activities of DHFR and TYMS 100-fold results in increased synthesis of thymidine but only modest changes to the concentration other folates. Thus, decoupling the DHFR/TYMS pair from the remainder of the pathway may represent a strategy for ensuring modularity of the different metabolite pools. Substantial further work is necessary to go beyond this case study and comprehensively test the relationship between the modules identified by

comparative genomics, the pattern of functional dependency in the cell, and adaptation to environmental changes.

5.2 Genome wide analysis of synteny and co-occurrence reveals additional coevolving pairs

With this goal in mind, we extended our analyses of synteny and co-occurrence to consider all gene families represented in *E. coli*. The computational methods for this work are described in detil by Schober et al [1], but in brief: We defined gene families using the Clusters of Orthologous Groups of proteins (COGs) by Koonin and colleagues. COGs represented in *E. coli* were filtered based on their sample size in 1445 complete bacterial genomes in order to ensure adequate statistical representation. All possible pairs among the remaining gene families were analyzed based on synteny and co-occurrence 2095 COGs, ~500,000 pairs in total). To compare our analysis to existing functional annotations, we mapped metabolic proximity from KEGG and [4] and (2) the set of high-confidence binding interactions in *E. coli* reported by the STRINGdb onto the considered gene pairs [5]. Consistent with intuition and prior work, co-evolving gene pairs show enrichment for physical complexes, enzymes in the same metabolic pathway, and more specifically, enzymes with a shared metabolite (Figure 5.1A-B). To identify gene pairs that are strongly evolutionarily coupled to each other but decoupled from the remainder of the genome, we constructed scatterplots of the resulting data (Figure 5.1C-D). These plots indicate the strength of coupling within each gene pair (as a relative entropy, along the x-axis) compared to the strongest external coupling involving just one

of the two genes. Gene pairs that fall below the diagonal (dashed line) are more tightly coupled to one another than they are to any other gene in the dataset. These scatterplots provide a simple graphical method for predicting two-gene coevolving units (258 by synteny, 194 by co-occurrence, such as the DHFR/TYMS pair indicted in red. Our analysis will need to be extended in order to identify larger communities within the complete network of coevolutionary relationships [6]. Many of the predicted adaptive units are engaged in the same physical complexes, while others share a metabolic intermediate like DHFR and TYMS.

In this plot, the degree of pairwise modularity is indicated by the distance of each point from the diagonal. By construction, points above the dashed lane have an interaction with at least one other gene that is similar to or greater than the coupling within the pair. The highly modular regime where regime where $D_{ij}^{\text{intra}} > 1.0$ and $D_{ij}^{\text{exter}} < 0.5$ on the synteny plot, we observe a few other pairs with experimental evidence for adaptive coupling. For example, the *accB/accC* gene pair encodes two of the four subunits of acetyl-CoA carboxylase, which catalyzes the first enzymatic step of fatty acid biosynthesis. Previous work shows that overexpressing either *accB* or *accC* individually causes a reduction in fatty acid biosynthesis; however, overexpressing the two genes in stoichiometric amounts rescues this defect [7, 8]. Similar constraints on relative expression have also be observed for the *selA*/*selB* and *tatB*/*tatC* gene pairs [9, 10]. The *tatB/tatC* genes encode subunits of the TABCE twin arginine translocation complex, while *selA/selB* are involved in selenoprotein biosynthesis but not known to bind physically. In conclusion, the statistical pattern of modular coevolution that we observed

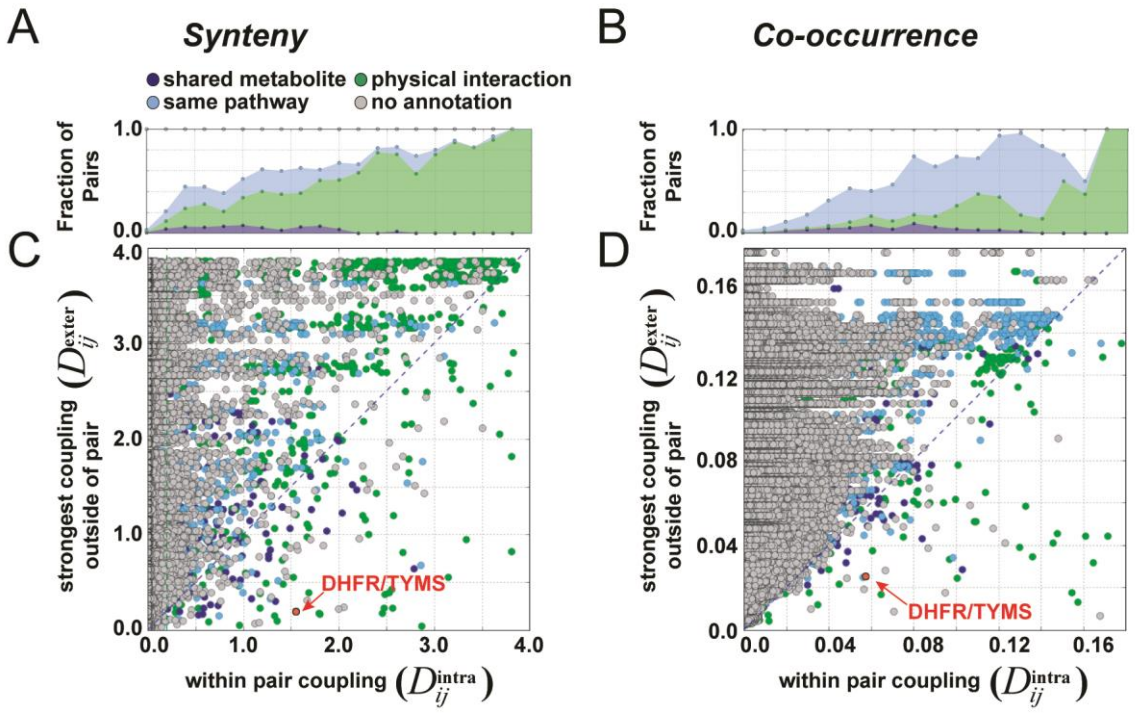

Figure 5.1 Genome wide analyses of coevolution in *E. coli*. Data represent an analysis of 2095 ortholog families (COGs) found in *E. coli* across 1445 bacterial genomes; computational methods are described in Schober et al [1]. **A-B,** Enrichment of physical and metabolic interactions as a function of evolutionary coupling, according to synteny and co-occurrence respectively. **C-D,** Scatterplot of evolutionary coupling for all analyzed gene pairs. Each point represents a unique pair of orthologous genes (COGs). Coupling within the pair is shown on the x-axis, while the strongest external coupling involving just one member of that gene pair is indicated along the y-axis. Thus, gene pairs below the dashed line are more strongly coupled to one another than they are to any other family of ortholgs in the analysis. Color-coding reflects annotations from the STRING database (physical interactions) or KEGG (metabolic pathways): green represents physical binding, while pairs in dark blue or light blue are not annotated as physical interactions but are found in the same pathways. Pairs colored dark blue share a metabolic intermediate. The DHFR and TYMS gene pair is highlighted in red on each plot.

in the folate pathway is also relevant to many other cellular contexts. The gene pairs below the diagonal now serve as a starting place for more deeply understanding the hierarchy of evolutionary couplings within cellular systems. We propose that adaptive modularity may be a general property of cellular systems, and the data presented here provide the necessary computational hypotheses to test this claim more generally.

5.3 Using coevolution to study the functional constraints on the amino acid sequence

Modification to the amino acid sequence of a protein can mediate subtle or even innovative changes to its function and thus the behavior of a pathway as a whole. However, the functional constraints between interacting proteins that shape the evolution of protein sequences are still poorly understood. In order to produce the most complete description of adaptive coupling between proteins, we seek to infer these constraints from models of coevolution. Given the mostly-direct relationship between the amino acid sequence and protein function, doing so has the potential to provide insight into the mechanistic forces that underlie such interactions. Mirror-tree is a computational method which compares the evolutionary history of two protein families and produces an interaction score based on covariance in their coding sequences [11]. Unlike a number of other approaches, the mirror-tree framework is able to explicitly incorporate information about the phylogenetic relationships between its samples (species) [12, 13]. Phylogenetically aware mirror-tree has been shown to be a reliable indicator of functional interaction, but does not explicitly describe coevolution between individual positions of the amino acid sequence. In the final component of my thesis work, I derived an additive mapping between the mirror-tree interaction score and the contribution of individual sequence positions. The resulting matrix, termed positional mirror-tree, describes coevolution between all position pairs between the amino acid sequences of two protein

families. By taking an average across all of the values of the positional mirror-tree matrix, one can exactly reconstruct the canonical mirror-tree interaction score.

For illustrative purposes, I applied this new measure to a toy system consisting of several well-known physical complexes. Somewhat surprisingly, I found that positional mirror-tree is numerically similar to an established model of sequence coevolution called statistical coupling analysis (SCA), despite substantial differences in their mathematical construction. However, a fraction of the position pairs that appear to coevolve according to SCA are represented as phylogenetic noise in the positional mirror-tree framework. In addition, positional mirror-tree is unique in its ability to quantify the total interaction between protein families. The mirror-tree matrix is sometimes able to resolve multiple distinct collectively coevolving groups of positions within the same protein family. Previous studies guided by SCA has shown that such communities can modularly determine specific aspects of protein function such as substrate specificity or thermal stability [14]. Further computational and experimental work is needed to test the biological significance of the positional mirror-tree matrix and its relationship to other existing models of sequence coevolution. Given the existence of multiple coevolving communities within some protein families, intelligently utilizing position-resolution information may provide a strategy for improving the predictive power of mirror-tree. Sequence coevolution should prove to be a powerful companion to methods such as synteny and co-occurrence in our quest to understand the adaptive interactions between proteins which shape the cell. The work presented here provides the theoretical framework and experimental motivation for doing so.

References

1. Schober, A.F., et al., *A Two-Enzyme Adaptive Unit within Bacterial Folate Metabolism.* Cell Rep, 2019. **27**(11): p. 3359-3370.e7.
2. Rodrigues, J.V. and E.I. Shakhnovich, *Adaptation to mutational inactivation of an essential E. coli gene converges to an accessible suboptimal fitness peak.* bioRxiv, 2019: p. 552240.
3. Bjarnason, G.A., et al., *Circadian expression of clock genes in human oral mucosa and skin: association with specific cell-cycle phases.* Am J Pathol, 2001. **158**(5): p. 1793-801.
4. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
5. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
6. Newman, M.E.J., *Networks : an introduction*. 2010, Oxford ; New York: Oxford University Press. xi, 772 p.
7. Abdel-Hamid, A.M. and J.E. Cronan, *Coordinate expression of the acetyl coenzyme A carboxylase genes, accB and accC, is necessary for normal regulation of biotin synthesis in Escherichia coli.* J Bacteriol, 2007. **189**(2): p. 369-76.
8. Janssen, H.J. and A. Steinbuchel, *Fatty acid synthesis in Escherichia coli and its applications towards the production of fatty acid based biofuels.* Biotechnol Biofuels, 2014. **7**(1): p. 7.
9. Bolhuis, A., et al., *TatB and TatC form a functional and structural unit of the twin-arginine translocase from Escherichia coli.* J Biol Chem, 2001. **276**(23): p. 20213-9.
10. Rengby, O., et al., *Assessment of production conditions for efficient use of Escherichia coli in high-yield heterologous recombinant selenoprotein synthesis.* Appl Environ Microbiol, 2004. **70**(9): p. 5159-67.
11. Pazos, F. and A. Valencia, *Similarity of phylogenetic trees as indicator of protein-protein interaction.* Protein Eng, 2001. **14**(9): p. 609-14.
12. Sato, T., et al., *The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships.* Bioinformatics, 2005. **21**(17): p. 3482-9.
13. Sato, T., et al., *Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions.* Bioinformatics, 2006. **22**(20): p. 2488-92.
14. Halabi, N., et al., *Protein sectors: evolutionary units of three-dimensional structure.* Cell, 2009. **138**(4): p. 774-86.