Coevolution Network Models Predict the Impact of Multiple Mutations on Protein Function

by

Violeta Beleva Guthrie

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

April, 2016

Abstract

Proteins often evolve new functions by acquiring a small number of mutations in an ancestral sequence not containing the phenotype. Modeling the functional effect of a mutation is, however, a nontrivial task, due to strong functional interdependencies.

Here, I used the recent evolution of the bacterial enzyme TEM β–lactamase under antibiotic selection as a model for genetic adaptation. I compiled a database of TEM β–lactamase sequences evolved under antibiotic resistance selective pressure and identified functional interactions between individual mutations/mutated residues. I built network models of coevolving residues (possible functional interactions), in which nodes are mutations and edges represent coevolution between two mutations. I reconstructed both the alignment and phylogeny-based mutation coevolution networks and assessed the utility of network-theoretical tools to derive information regarding role of individual mutations in the observed resistance.

Coevolution network analysis reveals key properties of mutations in evolution of antibiotic resistance, many of which were confirmed through extensive fitness measurements in the lab and by previous experimental studies of TEM β–lactamase

function. One finding is that mutations form densely connected clusters in the network corresponding to selection to different main classes of antibiotics or to different adaptive strategies within the same antibiotic class. Mutations that are central in the network tend to be either adaptive or compensate for effects of many other mutations.

By extending node centrality metrics to paths of mutations (connected nodes in the network) I was able to study properties of adaptive evolutionary trajectories in TEM. I found that central paths are enriched in non-negative functional interactions. Specifically, paths corresponding to triple mutants were experimentally shown to increase fitness from all or most of their constitutive single and double mutants. It was also shown that relative rankings of central paths and their constituent shorter paths can be used to predict the direction of fitness change in an evolutionary trajectory. In this way, this predictor of the effect of an evolutionary trajectory can be useful in anticipating evolution of antibiotic resistance.

In summary, my analysis of the combined functional effects of mutations in producing new biological activities should help anticipate evolution driven by a variety of clinically-relevant selections such as drug resistance, virulence, and immunity.

Thesis Committee

Primary Readers

Rachel Karchin (Primary Advisor)
Associate Professor
Departments of Biomedical Engineering and Oncology
Institute for Computational Medicine
Johns Hopkins University

Marc Ostermeier
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins University

Alternate Reader

Joel S. Bader
Associate Professor
Department of Biomedical Engineering
Johns Hopkins University

Acknowledgments

Below, I acknowledge the people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My most sincere thanks to my advisor Dr. Rachel Karchin for her encouragement and full support of my ideas for my graduate studies. She has taught me a great deal about conducting scientific research and has been a career role model for me with her enthusiasm, her depth and breadth of knowledge, and professionalism. Rachel not only gave me insightful advice regarding my research but also shared her personal experiences with me as guidance for my future directions.

I have also received a lot of useful advice, insightful ideas and suggestions, from my thesis committee members, Dr. Marc Ostermeier and Dr. Joel Bader. Many thanks for their contribution and patience during the completion of my Ph.D. project!

I am indebted to all current and past members of the Karchin group: Noushin Niknafs, Christopher Douville, Collin Tokheim, Rohit Bhattacharya, and Ashok Sivakumar, all exhibit a level maturity that goes well beyond their respective scientific career stages. They have inspired me over the past years, and I hope to continue to

work with them on exciting new projects during my upcoming post-doc in the Karchin group. Special thanks to Noushin, who has been the greatest friend at and outside of work, and I hope our friendship and collaborations will last well into the future!

In the past, I have also had the unique opportunity to learn from more senior lab members and like Hannah Carter, currently an assistant professor at UCSD, and David Masica, currently an assistant research professor at Hopkins. Hannah has become another career role model for me beyond Rachel, due to her motivation, hard work, and exceptional achievements. Other past Karchin lab members I am grateful for meeting and working with are Jean Fan (a great scientist and a talented photographer!), Grace Yeo, Xinyan Wang, Andrea Corredor, Dewey Kim, and Josue Samayoa.

I would like to thank our UCSC collaborators Dr. Manel Camps, Melissa Standley and Jennifer Allen. Melissa and Jennifer did all of the hard experimental work needed to test and validate my computational predictions in the lab. I have learned a lot about the properties of the TEM β–lactamase evolution and protein evolutionary principles from Manel. Every one of the countless brainstorming sessions and discussions we have had has truly propelled this work forward.

Finally, I would like to thank my family: My mother has been an inspiring woman of great intellect and abounding energy. While she did not choose a scientific career, she has insatiable curiosity and fearlessness when it comes to learning new things. My grandmother, who has been like a mother to me and the strongest woman I ever knew, has also been the greatest female role model I could ever have.

ACKNOWLEDGMENTS

My son George has infused new purpose to my life, giving me strength and happiness in difficult moments. My loving husband Daryl has steadily encouraged, supported, and helped keep me calm during all of those stressful Ph.D. years!

This research was supported in part by a National Science Foundation Advances in Biological Informatics (ABI) Innovation grant received in 2013 by Dr. Rachel Karchin and Dr. Manel Camps.

Dedication

To the most wonderful person in my life, my son George Guthrie.

Contents

CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

One of the fundamental principles in cell biology is that the amino acid sequence of proteins specifies their three-dimensional structure and biochemical function [1]. Improved understanding of the genetic basis of protein evolution should help predict the functional impact of mutations, which has critical clinical and biotechnological implications [2–4]. Proteins evolve to acquire new functions in a process known as adaptation. The genetic

1.1 Modeling the functional impact of single mutations

Until recently, most bioinformatic approaches for functional effects prediction have focused on single amino acid residue substitutions in a protein. Such approaches

have been applied to discovery of mutations that affect protein function in large-scale mutagenesis projects focused on a wide range of organisms [5, 6]. Bioinformatics methods have also been applied to the related problems of predicting a protein's function and predicting the location of functionally important protein regions (such as binding sites), based on sequence, evolutionary history, and/or structure [7–9].

Many bioinformatics methods for predicting mutation effects consider evolutionary history and/or biophysical properties of only single residue positions [6, 10]. Conserved positions in proteins are used to indicate sites that are key to maintaining structure and function [11]. On the other hand, variable sites occurring in otherwise conserved protein superfamilies are used to identify functional specialization [12]. The evolutionary history reconstructed from an alignment can provide additional information on the level of selection experienced by distinct protein sites. Applying the neutral theory of amino acid substitution, rates of molecular divergence are studied to estimate the level of positive or negative selection at a given site [13].

Phylogenetic trees are built on the assumption that similarities in morphological or molecular characteristics between any two organisms can be explained through a *common ancestor*. The principles of conserved and variable sites can be extended to the analysis of an inferred phylogenetic tree [14].

1.2 Modeling coevolution and functional dependencies between mutations

The assumption that different sites in a protein evolve independently cannot explain drastic changes in the functional effects of mutations that depend on what other mutations are present. In an extreme example, mutations that are deleterious and pathogenic in the human protein are present as the wild type protein residues in ortholog sequences of non-human species [15]. This suggests that other residues in the protein compensate for the deleterious effects of these mutations, and these functional interactions result in coevolution between these sites [16].

It is now well accepted that methods to predict the combined impact of multiple mutations will have great utility for protein engineers who seek to design proteins with new or improved functions [17, 18]. For example, such methods can contribute to the design of therapeutic regimens for diseases driven by bacteria or viruses, in which rapid evolution on short timescales generates drug resistance [19]. Introducing functional interactions as model parameters, greatly increases the complexity of models. For example in a regression model, the number of parameters needed to model pairwise (and beyond) interactions will lead to a number of parameters that far exceeds the number of experimental observations [19]. This problem is referred to as the curse of dimensionality [20].

Several statistical methods exist to identify direct pairwise functional interactions

between mutations. These methods include evolutionary trace (ET), statistical-coupling analysis (SCA), direct coupling analysis (DCA), and residue coevolution networks [7–9, 21–23]. ET [7] uses a phylogenetic tree to group protein sequences and rank the functional importance of amino-acid residues by correlating their evolution with divergence in the tree. Residues traced in this way are mapped onto a protein structure, and sites of clustering can be used to infer functionally important sites. SCA [8] relies on partitioning and perturbation of large and diverse multiple sequence alignments of homologous proteins to study higher-order interaction patterns. Direct-Coupling Analysis (DCA) [21, 23] combines covariance analysis with global inference analysis, adopted from use in statistical physics to distinguishes between directly and indirectly correlated residues, which in turn have been observed to accurately predict residue-residue contacts.

More similar to the study presented here are two previous studies of protein residue coevolution networks, based on large, diverse protein families. Both studies found that node connectivity and centrality had utility in predicting functionally important residues [9], and that protein specificity determining sites that do not cluster on the three-dimensional structure can still be found to coevolve due to complex functional constraints [22]. To my knowledge, none of these methods predict the impact of specific mutation trajectories in the coevolution network arising from higher-order interactions. However methods that predict such mutation trajectories and thus better describe selective pressures leading to increased function are needed in the evolutionary

biology community [24–26].

1.3 A mutation network model of protein evolution

Coevolution network approaches start with information on covariation from multiple sequence/phylogeny analysis, and position such pairwise interactions in the context of a network. Previously, most coevolution networks focused on large protein families and found essential structural constraints that maintain the function/structure of these families [9, 22, 27]. In reality, coevolving residues may not be co-localized in a protein structure and may represent complex evolutionary interactions or compensatory effects due to mutation pleiotropy [28, 29]. In contrast, the network coevolution models presented here are based on a collection of sequences closely related to a protein of interest, evolving under a defined selective pressure (Chapter 2). In this collection of evolutionarily-related sequences, sequences differ from each other by one or more point missense mutations. The network represents mutated positions (Chapters 3 and 4 and [28]) or specific mutations as nodes (Chapters 5 and 6). A link connecting two nodes corresponds to the strength of evolutionary interactions between two positions. In my work, such evolutionary interactions can be identified either at the level of aligned adaptive sequences (Chapters 3 and 4) or from co-occurrence in the same phylogenetic clade (Chapters 5 and 6).

A mutated position/mutation coevolution network has the following properties:

- A set of interacting mutations, of any order, is represented as a cluster or community in the network.
- An evolutionary trajectory is represented as a path through the network.
- Link weights represent the number of times two mutations occurred independently in the evolution of the protein family.
- Link direction can indicate preferred temporal ordering of a series of mutations during an evolutionary trajectory.
- Fewer parameters are required than in a regression that includes mutation interaction terms (pairwise interactions and beyond). The worst-case number of parameters for n mutations in the network is the maximum number of links between them, which is of order $O(n^2)$.

In this work I am able to identify communities of residue positions associated with different functional specificities (Section 3.4); expand pairwise interactions to adaptive evolutionary trajectories and predict fitness increasing combinations of mutations not previously encountered in natural evolution (Section 4.4.1); predict whether fitness increases or decreases in a specific evolutionary trajectory. While I use the bacterial TEM β-lactamase (Chapter 2) as a model system throughout this work, my network analysis could potentially be generalized to other proteins evolving under defined selective pressures.

Chapter 2

Model system: TEM β-lactamase evolution of antibiotic resistance

One of the most critical public health issues today is the evolution of microbial pathogens able to resist antimicrobial treatments [30–32]. Among the diverse antibiotic resistace strategies, some of the most common mechanisms include efflux pumps, which reduce the concentration of antibiotics inside the cell, and enzymes that modify, or otherwise metabolize antibiotics [33]. Some of the most prevalent antibiotic resistance enzymes are the β–lactamases [34]. These enzymes break down β–lactam antibiotics, such as penicillin and derivatives (e.g. ampicillin), cephalosporins (including (CTX)), monobactams, carbapenems and β–lactamase inhibitors [35]. β–lactam antibiotics act by interfering with bacterial cell wall synthesis by irreversibly binding to transpeptidases, enzymes that are involved in the cross-linking of the peptidoglycan layer of

bacterial cell walls. A wide variety of β–lactamases can be found in both gram-positive bacteria (where they are typically secreted) and gram-negative bacteria (where they are employed in the periplasmic space).

2.1 Evolution of antibiotic resistance in TEM β-lactamases

The TEM-1 β–lactamase was first isolated in the 1960s and named after the patient (Temoneira) providing the first sample [36]. Following the introduction of third generation cephalosporins in the 1980s, multiple variants with few differences in their amino acid sequence were isolated. Thus, as with other bacterial resistance evolution, resistance has arisen within few years of its first clinical use. The rapid emergence of antibiotic resistance is due to the selection for specialized traits that were already present in the environmental populations of bacteria [30]. Many of the most common resistance genes found in hospitals today are encoded on small plasmids that can be exchanged among different bacterial strains and species. This is one mechanism of horizontal transfer [37].

TEM-1 β-lactamase was one of the first antibiotic resistance enzymes for which it was demonstrated that amino acid substitutions could result in alteration of the resistance phenotype [38]. The clinical isolation of mutant TEM alleles as a result of the introduction of novel antibiotics at the beginning of the 1980s has provided

an extensive database of amino acid substitutions in the genes coding for TEM-1 mutants, which alter the genes' ability to provide antibiotic resistance [39, 40]. More than 200 derivatives of TEM-1 with aberrant amino acid sequences have been described today and catalogued in a public database [41]. This database includes both mutant sequences and the types of β–lactam antibiotics to which they are resistant. With this knowledge, TEM-1 has been used as a model system for the study of enzyme structure-function relationships, enzyme engineering, the in vitro evolution of antibiotic resistance and various fundamental evolutionary questions including the effects of fluctuating selective pressure, accessibility of evolutionary pathways, robustness, epistasis, and evolvability [38, 39, 42–54].

2.2 Properties of TEM as a model system for protein evolution

An important property of TEM β-lactamases as a model system for protein evolution is that there is a direct correspondence between the evolution of a new activity (such as extended-spectrum resistance) and bacterial survival [3]. Thus, computational predictions of the impact of multiple mutations can be systematically tested by introducing mutations of interest into the TEM gene and characterizing the bacteria carrying this mutated gene. Their survival, when exposed to extended spectrum antibiotics can be measured and used as a proxy for the protein's catalytic

activity. An established method for determining antibiotic resistance is the minimum inhibitory concentration (MIC), defined as the lowest concentration of an antimicrobial that will inhibit the visible growth of a microorganism after overnight incubation. MIC is generally regarded as the most fundamental *in vitro* measurement of the activity of an antimicrobial agent against an organism [55].

TEM β-lactamases have been evolved in the lab through alternating rounds of mutagenesis and selection for antibiotic resistance, primarily selection for increased MIC. Numerous such experiments have shown that *in vitro* evolution of TEM β-lactamase accurately mimics natural evolution [39, 44, 45, 53, 56–66]. Directed evolution experiments can also be used to predict the results of natural evolutionary processes, and to access new sets of mutations (sequence space) not previously observed in natural evolution of resistance [54]. These new (combinations of) mutations can also interact functionally, which provides additional sets of functional interactions to be studied.

Pervasive functional interactions, specifically sign epistasis [52] and pleiotropy, were identified in TEM when studying the accessibility of evolutionary trajectories from TEM-1 to the mutant containing mutations A42G, E104K, M182T, and G238S [49], as well as a promoter region mutation [44]. From this set of five mutations increasing fitness, only a 18 of all 120 possible trajectories (ordered combinations) were accessible. Most evolutionary paths would not be selected due to decreases it fitness at different stages of the trajectory. Overall numerous studies have used TEM

CHAPTER 2. MODEL SYSTEM: TEM β-LACTAMASE EVOLUTION OF ANTIBIOTIC RESISTANCE

β-lactamase as model system to show the importance epistasis in shaping protein evolution of new functions or characterize properties of mutations exhibiting epistatic interactions [51,52,67–71]. As such, TEM β-lactamase provides a rich system in which to study complex functional impacts of mutations during a protein's evolution of new functions.

Chapter 3

Alignment-based network of co-evolving positions in TEM

3.1 Introduction

In order to study how new biochemical activities arise during evolution, I compiled and aligned a database of clinically or experimentally derived TEM-1 β-lactamase mutant sequences.

My first assumption was that a majority of mutants present in the database would have undergone a degree of positive selective pressure; for sequences isolated in the clinic, the selection occurs via the β-lactam antibiotics that are administered to patients. In fact, the rapid evolution of β-lactamases in recent years has been linked to the widespread use of antibiotics [72, 73]. The experimentally derived sequences come

from directed evolution experiments, in which a mutation round is followed by selection of the mutants for a given level of resistance. A concordance between clinical and experimental TEM β-lactamase evolution has been well established [74]. Codon-based maximum likelihood phylogenetic analysis (in PAML, [75]) of the naturally occurring sequences further supports this assumption by showing enrichment of non-synonymous vs. synonymous mutations ($\omega > 1$) in most residue positions Table 3.1.

The second assumption was that frequent co-occurrence of a pair of mutated residue positions within the same sequences indicates a functional relationship between these positions. I constructed an undirected, weighted network representation of co-occurring residue pairs to map the potential functional interactions underlying the evolution of β-lactamase under antibiotic selective pressure. In this network model (shown in Figures 3.1 and 3.2), mutated residue positions are represented as nodes. Links connect pairs of nodes corresponding to residue pairs observed to be co-mutated in at least one TEM mutant sequence. In this representation, node size is proportional to weighted degree centrality, which shows how well a node is connected to its neighbors and how many neighbors it has (section 4.3).

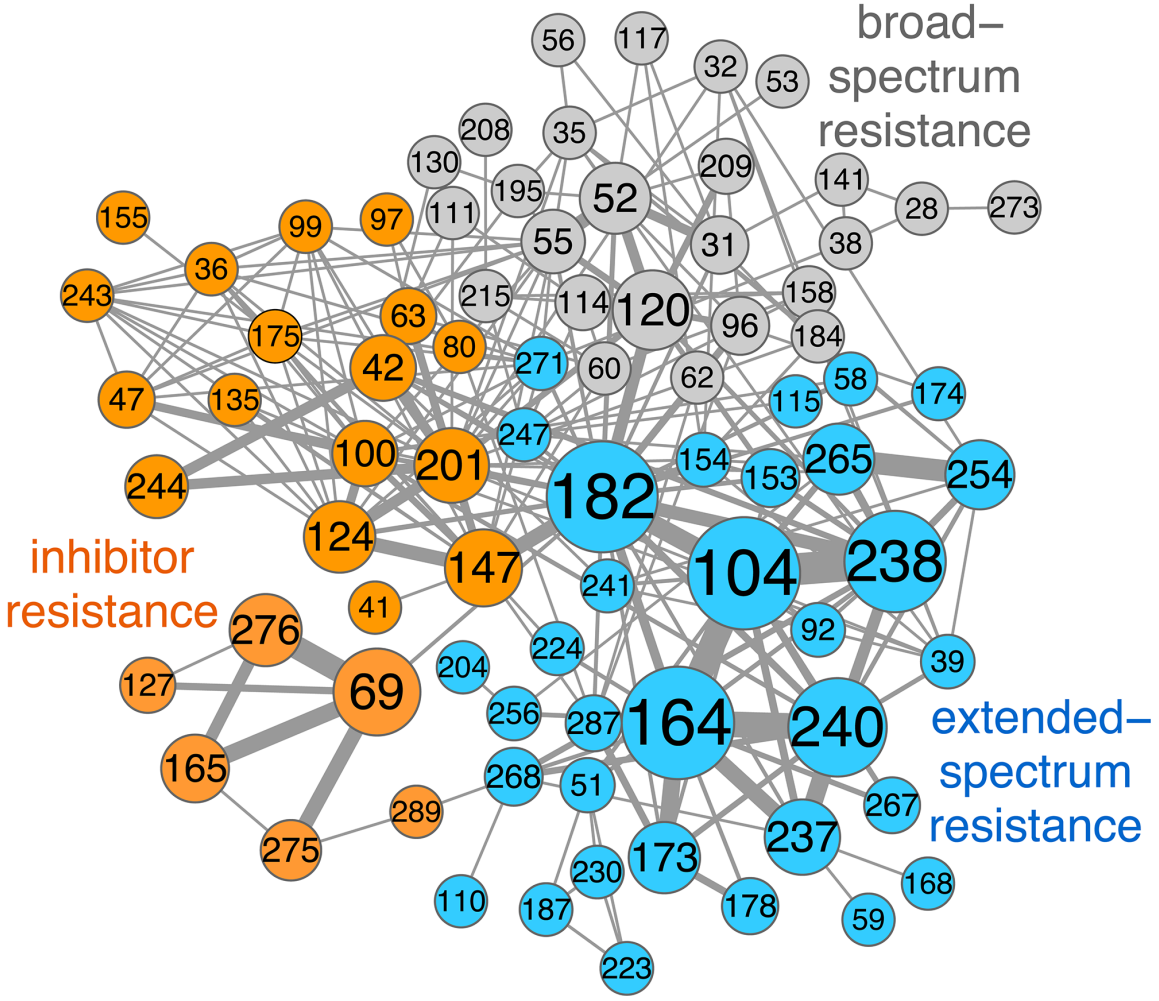

Figure 3.1: The TEM coevolution network and its communities: The network was constructed based on frequencies of co-occurring mutated residue positions in 363 mutant TEM β-lactamase sequences. Node size is proportional to how well connected a node is to its neighbors and how many neighbors it has (weighted degree centrality). Link thickness is proportional to the number of sequences in our database in which both positions are mutated, normalized by the number of sequences in which only one or the other position is mutated 3.2.4. Node (residue) numbers are shown in Ambler notation. The Clauset community-finding algorithm [76] identified three major communities, corresponding to three Bush-Jacobi β-lactamase phenotype classes: broad-spectrum antibiotic resistance or 2b (gray), extended-spectrum antibiotic resistance or 2be (blue) and inhibitor resistance or 2br (orange). Mutated positions with phenotypic effects documented in [74]: extended-spectrum resistance 51, 173, 237, 240, 39, 164, 104, 238, 153, 265, 92, 224; inhibitor resistance 165, 69, 275, 276, 244, 201; inhibitor and extended-spectrum resistance: 182 and 268.

3.2 Network construction

3.2.1 TEM mutant sequences alignment

The TEM mutant sequence database consists of sequences that have evolved under antibiotic selective pressure. This database includes clinical (n = 144 [77]) and laboratory evolved (n = 217 [39, 44, 45, 53, 56–66]) sequences.

Using TEM-1 as the reference sequence [77]) and the Ambler [78] amino acid residue numbering scheme for the class A β-lactamase superfamily to TEM, I constructed a multiple sequence alignment of naturally occurring and laboratory-evolved TEM mutants.

In order to examine the correspondence between network clustering patterns (section 3.4) and any common functional roles of mutations, I first annotated TEM sequences (but not individual mutations) by known phenotype class from the literature or the Lahey Clinic β-lactamase online database [41]. The phenotype class of naturally occurring TEMs is determined experimentally [55]. I was able to associate 380 out of 405 TEM naturally occurring or TEM laboratory-evolved mutant sequences in the database with a single major β-lactamase phenotype class (113 broad-spectrum, "2b", sequences, 201 extended-spectrum, "2be", sequences, and 49 inhibitor-resistant, "2br", sequences). There were also 17 sequences with a combined extended-spectrum antibiotics and inhibitor resistant phenotype class, "2ber", that were not used in the network. This was because it was not known whether the selection was for extended

spectrum, while starting with inhibitor resistance, or vice versa, i.e. selection for inhibitor resistance starting from extended spectrum resistance. I assumed that the resistance selection criterion used in the directed evolution experiments [39, 44, 45, 53, 56–66] determined the phenotype class of the TEM sequences coming from such experiments.

3.2.2 Identifying residues under selection in the TEM mutant sequence database

After constructing the mutant TEM sequence alignment, I analyzed the extent of positive selection in the naturally occurring TEM sequences. I compared the degree of positive selection at a residue position to the corresponding node's connectivity in the network. For this, I performed a PAML (codeml) analysis [75] for the naturally occurring TEM β-lactamase sequences. I used PHYLIP [79] to build a phylogenetic tree (gamma distribution, four classes, α parameter: 0.348). I used a log-likelihood test to compare the fit of codeml models 2 (three-classes of unselected/selected codon positions) and 1 (two-classes of unselected/selected codon positions) to the data, and found that model 2 was a better fit (χ^2 test, p-value $\ll 0.01$). Using model 2s three site classes, I found that out of 35 mutated amino acid residue positions in the network of naturally occurring TEM sequences, 11 were identified as strongly positively selected ($\omega \geq 8.4$) and 22 were positively selected (relaxed to $\omega \geq 0.8$) Table 3.1.

The top-ranking residues under positive selection tend to be well connected in the network, as per their high degree centralities. This means that they are frequently mutated together with other residues in the network, supporting the general belief that adaptive mutations are accompanied by many other mutations due to pleiotropy

Residue number	WT residue	Degree of positive selection ω (PAML)	Network node degree rank
104	E	10.5	1
164	R	10.5	2
240	E	10.5	3
182	M	10.5	4
238	G	10.5	5
69	M	10.5	6
237	A	10.5	7
275	R	10.5	10
244	R	10.5	20
153	H	10.4	12
165	W	9.9	9
223	S	4.3	15
224	A	4	25
55	K	3.1	21
187	A	3	15
230	F	3	15
268	S	2.1	14
221	L	2.1	27
49	L	2.1	31
276	N	1.7	8
51	L	1.6	11
280	A	1.6	29
226	P	1.4	18
196	G	1.3	23
175	N	1.1	31
179	D	1	34
289	H	1	22
215	K	1	35
248	A	1	18
92	G	0.9	33
163	D	0.8	24
127	I	0.8	13

Table 3.1 . . . continued

Residue number	WT residue	Degree of positive selection ω (PAML)	Network node degree rank
173	I	0.8	30
42	A	0.4	26
262	V	0.4	27

Table 3.1: Positive selection analysis for every codon in naturally occurring TEM sequences, in PAML 4 (codeml) [75]. Residue position number according to the Ambler system [78] (column 1); wild-type amino acid residue in TEM-1 (column 2); ω value (ratio of non-synonymous to synonymous nucleotide substitutions at a codon position) (column 3). Rows were sorted by decreasing ω value. Residues with $\omega \geq 8.4$ were classified by PAML as exhibiting strong positive selection, residues with $8.4 \geq \omega \geq 0.8$ were in the "relaxed" positive selection class, and residues with $\omega < 0.8$ were under no selection. The weighted node degree centrality 4.3 was computed in a network constructed with only sequences from clinical isolates (column 4).

3.2.3 Counting of co-selection events in TEM sequences

If two TEM residue positions appear altered in one or more *naturally occurring* TEM mutant sequences, these mutant sequences are considered distinct co-selection events. This approach does not account for any evolutionary relationships in TEM sequences (addressed in Chapters 5 and 6).

The directed evolution experiments included in my sequence database tend to consist of multiple rounds of selection in defined concentrations of β-lactams. In this way, only resistant TEM mutants that are selected in one selection round are used in the next round. In each subsequent round, TEM mutant sequences acquire additional random mutations and only sequences conferring the required level of

β-lactam resistance are then selected in the following round, etc.

As a conservative way to only include pairs of mutations that arose independently, I did not count the occurrence of a pair of mutations again when it continued to appear through the subsequent selection rounds of a laboratory evolution experiment. In addition, if a pair of mutations was already present in the library of sequences that was used to start the directed evolution experiment, that pair was counted only once, and only if it was present in the first selection round.

In the clinical samples database [77], the Q39K mutation arises independently only once, in the TEM-2 β-lactamase, as demonstrated by earlier studies of the TEM phylogeny [39]. Therefore, naturally occurring TEM mutants have either descended from TEM-1 directly, or through TEM-2. For the construction of the coevolution network, I was interested in how many times a mutation appeared and was selected for independently, therefore, I removed residue 39 from the alignments of naturally occurring TEMs, but any mutations in this residue found in in vitro evolution experiments were included in the model. This is an example of spurious correlation arising from the evolutionary history of TEM mutant sequences. In Chapters 5 and 6 this evolutionary history (phylogeny) was taken into account and Q39K did not have to be removed.

3.2.4 Network weights

To indicate the potential strength of the interaction, links within the network are weighted in proportion to the number of residue pair co-occurrence events. Specifically, two nodes (two mutated amino acid residue positions) are linked if mutations at both residues exist in at least one TEM sequence in the alignment. To estimate the number of times that mutations at two residue positions have coevolved, I counted independently selected mutation pairs (section 3.2.3). The weight w of each link is proportional to the number of sequences in which both positions are mutated, normalized by the number of sequences in which only one or the other position is mutated, which is a Jaccard-like index [80]:

$$w(M_i, M_j) = \frac{c(M_i,M_j)-(1-\varepsilon)}{c(M_i)+c(M_j)-c(M_i,M_j)} \tag{3.1}$$

where $c(M_i)$ and $c(M_j)$ are the number of times a the i^{th} and j^{th} column (residue position), respectively, are mutated in the alignment. $c(M_i, M_j)$ is the number of times both columns are mutated together, and $w(M_i, M_j)$ is the network weight of the link between nodes i and j (or residue positions i and j).

A correction term was included to ensure that mutated pairs, which occur in a single sequence together and never by themselves, are not overweighted. Without this term, these pairs would always have (the maximum) link weight of one. ε is the inverse of the number of aligned sequences used to construct the network (a heuristic

choice that works well in practice).

3.3 Global network properties

The weighted degree distribution of the network, i.e. the aggregate weight of the links incident on each individual node, reveals overall few highly connected nodes, with a majority of nodes exhibiting low connectivity (Figure 3.2).

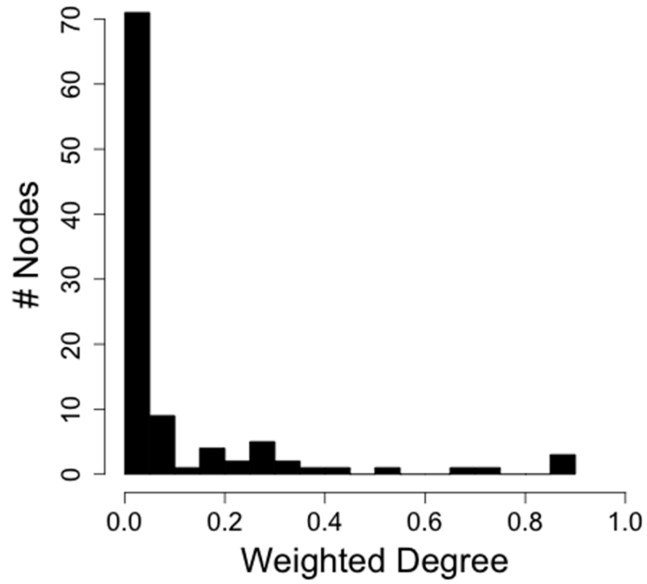

Figure 3.2: The weighted degree distribution of the TEM alignment-based coevolution network: The distribution of nodes by aggregate weight of links per node (weighted degree centrality, 4.3) is shown. Many nodes (residue positions) with high weighted degree are functionally important (Table 4.1). The distribution reveals that the network contains very few highly connected nodes, with a majority of the nodes exhibiting low connectivity. This topology is similar to that of scale-free networks [81], and is reminiscent of the connectivity distribution of other biological processes such as signaling or cellular differentiation.

3.4 Network communities and selective pressures

To identify highly connected subnetworks (communities) of mutated residue positions, I used the Community-Structure-Partition algorithm [76], implemented in the Graph Utilities Package in Mathematica 7.0 [82]. Communities with five or fewer nodes were merged onto one of the larger communities. The choice of a larger community onto which to merge the smaller community was determined by calculating the overall network modularity function [76] after a suggested merge. The merge that resulted in the highest network modularity was the one that was chosen.

The TEM coevolution network also has a modular structure, with a modularity score $Q = 0.522$, where $0 \leq Q \leq 1.0$; This modularity occurs in a hierarchical way, with larger communities and the communities within them (Figures 3.1 and 4.1). The Clauset community-finding algorithm [76] identified three major network communities (Figure 3.1). I found a clear correspondence between each of these communities and each of the β-lactamase phenotype classes defined by Bush and Jacobi [83]: (1) broad-spectrum antibiotic (2) extended-spectrum antibiotic (3) inhibitor resistance. These communities help identify the different sets of mutations that are selected for different resistance functions. While some mutations, like M182T, which is thermodynamically stabilizing, can be found in sequences with different resistance phenotypes, they are preferentially found in one single community. Functional interactions and

the resulting coevolution between mutations could be the reason for this preferential assignment [84, 85].

On a narrower level, within the two adaptive community networks (the extended-spectrum and inhibitor-resistant community networks), I found *subcommunities*, i.e. subnetworks of densely connected nodes. These *subcommunities* likely represent parallel strategies of adaptation within a community's phenotype class, namely trajectories leading to different local maxima within the fitness landscape (3.4 and 3.4).

Community associated with broad-spectrum resistance in TEM

The broad-spectrum antibiotic community includes mutations previously reported as nearly neutral or as preserving the parental TEM-1 phenotype, since catalytic efficiency for broad-spectrum β-lactams has evolved to perfection in TEM-1 [86]. The extended-spectrum community contains mutations at eight positions that are known to extend the substrate spectrum of the enzyme: 39, 51, 104, 164, 173, 237, 238, 240 [39, 49, 59, 62, 65, 68, 74, 87–93], as well as four stabilizing mutations: 153, 182, 224, 268 [49, 53, 74, 91, 94, 95].

Community associated with extended spectrum resistance in TEM

This extended spectrum community contains two large subcommunities, which are discussed in detail in 4.2. Central to each subcommunity is one position involved in substrate recognition, 164 and 238 respectively. R164H/S/C mutations are thought to

lead to the collapse of the Ω-loop, creating greater active site accessibility (Figure 3.3); G238S on the other hand, appears to increase affinity for the substrate and/or cause repositioning of the Ω-loop (Figure 3.3). These two mutations were shown to represent alternative evolutionary solutions, leading to parallel, divergent mutation trajectories with different fitness optima and are known to exhibit negative epistasis [68].

Community associated with inhibitor resistance in TEM

The inhibitor resistant network comprises two communities corresponding to two distinct mechanisms disrupting inhibitor binding at the active site [81]. One involves positions 69 and 276, which are strongly connected in one community, and the other one involves 244, which is in a separate community Figure 3.1. Likewise, the inhibitor community contains five positions known to confer inhibitor resistance: 69, 165, 244, 275, 276 [56, 74, 96–101] and three enhancer stabilizing mutations: 147, 201, 275 [48, 53, 74, 95, 99, 102, 103].

Communities associated with new resistance function in TEM

The observed segregation of residue positions according to the selection driving their evolution is remarkable given that no phenotype class information was used to construct the network. This effect is consistent with previously described antagonistic pleiotropy between different resistance phenotypes [104]. Within the two adaptive communities (extended spectrum and inhibitor resistance) community annotation

largely matched phenotypic data: five mutant positions were correctly classified as inhibitor resistance mutations and 12 positions were accurately classified as extended-spectrum mutations (Figure 3.1, legend).

TEM mutations at the interface between adaptive communities

Interestingly, mutations that are known or suspected to contribute to both inhibitor and extended spectrum antibiotic resistance (182, 268, 201) are at the interface between the two communities. Positions 100 and 147 are similarly located at this interface. These are positions with likely compensatory, thermodynamically stabilizing mutations [53, 66, 95, 102] that have also been found in extended-spectrum evolution experiments [39] [65] [68]. They may also belong to the dual resistance phenotype category, as experimental data on inhibitor resistance evolution is scarce. The only clearly misclassified mutant positions are 175 (involved in extended-spectrum resistance [105] but classified as inhibitor resistance) and 130 (an inhibitor resistance mutation classified as broad-spectrum). In the case of the catalytic site residue 130, the misclassification was due to the fact that the S130G mutation confers resistance to inhibitors on its own and therefore rarely co-occurs with other mutations. Its assignment to the broad-spectrum community is based on a single co-occurrence event in the database.

Mapping adaptive communities of residues to the TEM tertiary structure

In summary, I find that selective pressures leave recognizable footprints on the TEM network's connectivity. Furthermore, the amino acid positions within network modules are not necessarily physically close in the protein's tertiary structure, as interactions are defined genetically (functionally) rather than physically. To illustrate this point, Figure 3.3 maps nodes (mutant positions) belonging to the three major communities in the TEM coevolution network onto the tertiary structure of the TEM enzyme (PDB ID: 1ero). It is apparent that neither community is physically localized to a defined area of the protein.

3.5 Conclusions

Here, I used co-occurrence in the sequence alignment of TEM sequences of three main resistance phenotypes (broad, extended-spectrum, and inhibitor-resisance), as an indicator of potential functional interaction. Pairwise interactions were visualized using a network representation where each node is a mutant position, and each link represents occurrence of two mutated positions in the same sequence. The resulting undirected, weighted network has a few highly connected nodes and a majority of nodes exhibiting low connectivity (Figure 3.2). This connectivity property [107] is reminiscent of the link distribution in networks representations of other biological processes, such as cell signaling or differentiation, where it helps in buffering noise

caused by random variation within the system. In the case of proteins, it may contribute to robustness to mutation.

Communities in this network correspond to the three distinct phenotypic categories. The observed segregation of residue positions according to the selection driving their evolution is remarkable given that no phenotype class information was used to construct the network. This effect is consistent with previously described antagonistic pleiotropy between different resistance phenotypes [104]. Within the two communities with non-ancestral phenotype (extended-spectrum and inhibitor resistance) I found that community annotation largely matched phenotypic data, while the amino acid positions within network modules are not necessarily physically close in the protein's tertiary structure.

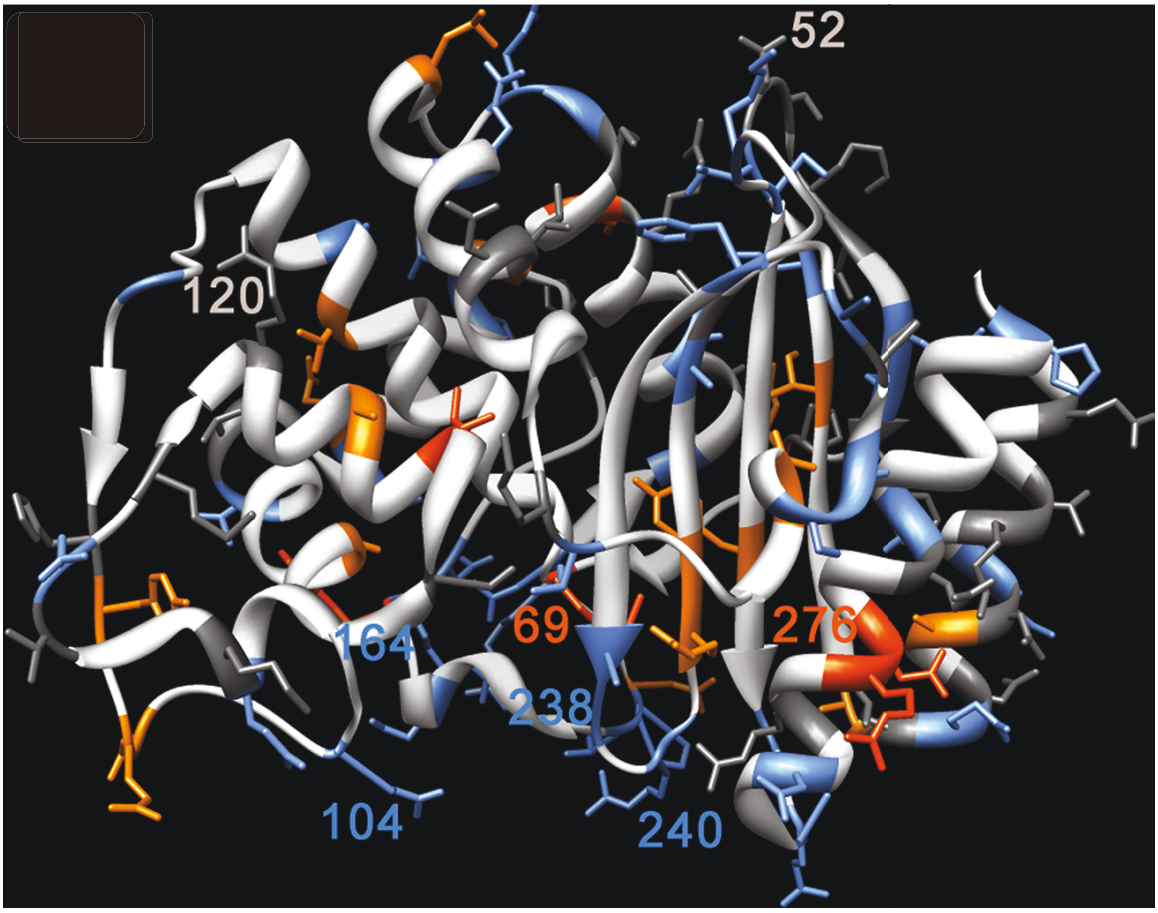

Figure 3.3: Locations of amino acid residues in the TEM coevolution network mapped onto the TEM tertiary structure (PDB 1ero). Residues in the TEM coevolution network (Figure 3.1) are colored by community membership: gray (broad-spectrum resistance), blue (extended-spectrum resistance) and orange (inhibitor resistance). The communities do not map to distinct regions of the tertiary structure. Image created with UCSF Chimera [106].

Chapter 4

Alignment-based network of extended spectrum resistance evolution

4.1 Introduction and overview.

In Chapters 3 and 4, I constructed and analyzed the structure of a network of positions found to be mutated in all TEM β-lactamases evolved under *multiple* (natural or laboratory) selective pressures. The network communities segregate residue positions into groups related by the function which was selected, i.e. extended spectrum or inhibitor resistance. In this chapter, I focus on a network model of mutant positions evolved under a *single* selective pressure. Extended-spectrum antibiotic resistance is

the best-represented resistance phenotype class in my TEM mutant sequence database, both in clinically isolated and laboratory evolved sequences.

Adaptive trajectories and information flow through the network.

Since the evolution of TEM to acquire extended spectrum antibiotic resistance is an adaptive process, mutations that become fixed are the result of a positive selection. Therefore, adaptive evolutionary trajectories can be conceptualized as a successful combination of functional milestones. In this scenario, the evolution of new biochemical activities involves transfer of information within the network, where each node is a potential functional milestone. I reasoned that efficient information transfer would improve the chances of generating mutant combinations with high fitness. Thus, within the context of a network based on sequences selected only for a given in phenotype, every edge in the network should represent a favorable evolutionary interaction. Furthermore, if I assume that every mutant position represents a potential functional milestone, adaptation involves information transfer across the network [9].

Network communities and alternative adaptive trajectories.

I constructed a coevolution network based on the sequence alignment of extended spectrum TEM sequences only, similarly to Section 3.2 and analyzed the community structure as in Section 3.4. In this case of a single selective pressure, I found that distinct communities of mutated positions tend to represent alternative strategies of

adaptation [68].

Adaptive trajectories and central network paths

In order to find significant adaptive trajectories, I applied graph-theoretical metrics to find the most central network paths. Specifically, I focused on shortest path betweenness centrality (Section 4.3), which I used to measure the importance of a path for information transfer across the network. It is assumed here that central paths in the network are of likely special significance for adaptation.

I started with the most experimentally tractable evolutionary trajectories (ones involving three mutations) and identified the central network paths. The particular significance of the corresponding evolutionary trajectories identified by my analysis is demonstrated because they frequently increase CTX resistance over constituent double mutation pairs. Even though most of these trajectories had been previously described, the ability to identify them implies that this analysis has predictive value, as it had no information about the original sequence context of the co-occurring pairs of mutations.

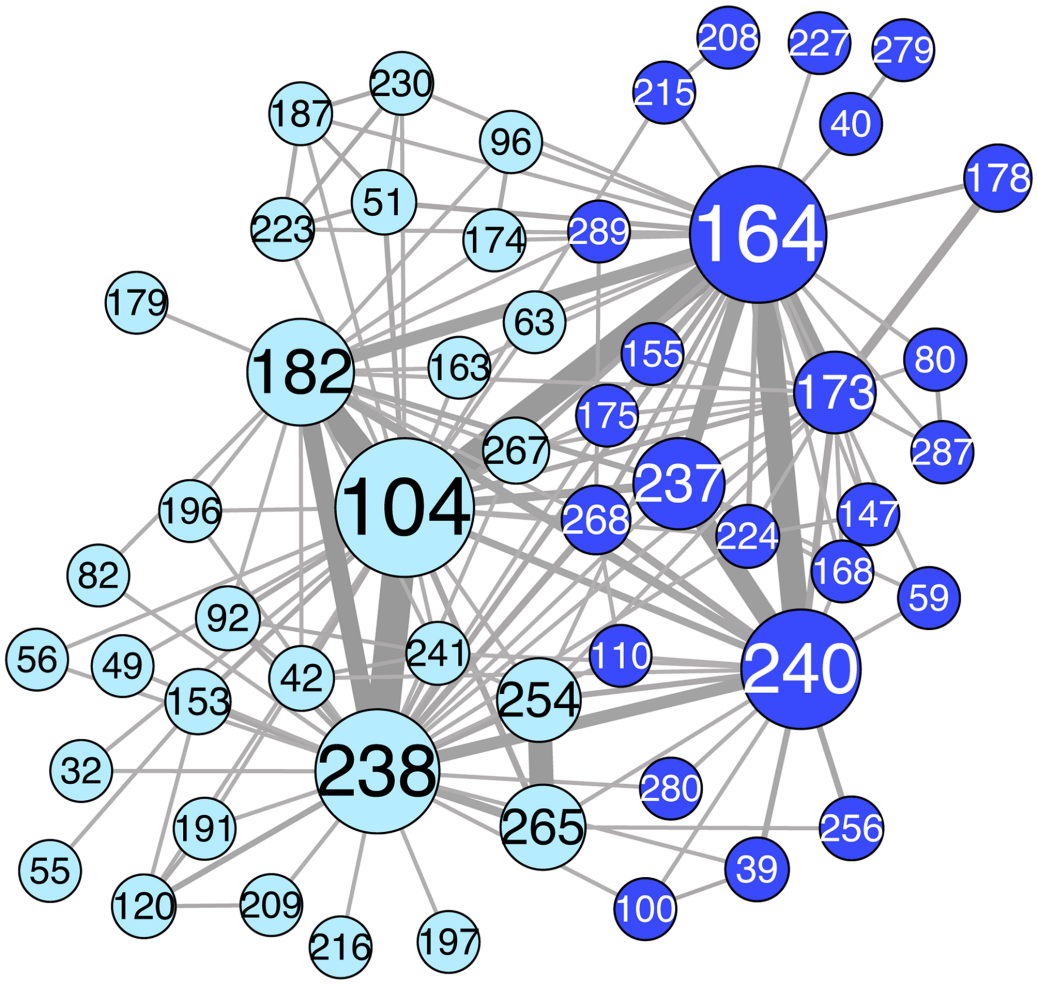

Figure 4.1: The extended-spectrum TEM network and its two communities: The network was constructed in the same way as Figure 3.1, but here I only used sequences associated with extended-spectrum antibiotic resistance. I identified two large communities, the first containing the active-site residue 238 (light-blue), and the second containing the active-site residue 164 (dark-blue). Node size is proportional to how well connected a node is to its neighbors and how many neighbors it has (weighted degree centrality 4.3). Link thickness indicates how frequently two residues (nodes) are mutated in the same sequence, normalized by the number of sequences in which only one or the other position is mutated as in Section 3.2.4. Image created with CytoScape [108].

4.2 Properties of the extended-spectrum network

I constructed an undirected, weighted coevolution network (as in Section 3.2), this time using only TEM mutant sequences conferring extended-spectrum resistance: A total of 201 naturally occurring and laboratory-evolved sequences were in the extended-spectrum database.

Extended spectrum network modularity.

The extended-spectrum resistance network contains two large communities (Figure 4.1). Central to each community is one position involved in substrate recognition, 164 and 238 respectively. R164H/S/C mutations are thought to lead to the collapse of the Ω-loop, creating greater active site accessibility (Figure 4.2A); G238S on the other hand, appears to increase affinity for the substrate and/or cause repositioning of the Ω-loop (Figure 4.2B).

Mutations R164S and G238S were shown to represent alternative evolutionary solutions, leading to parallel, divergent mutation trajectories with different fitness optima [68]. Specifically, divergent evolution appeared as a contingency effect of trajectories involving the mutually antagonistic G238S or R164S mutations. The first mutation in an adaptive trajectory thus significantly impacted the composition of subsequent evolutionary trajectories. In the extended spectrum resistance network

analysis, this divergent evolution is represented by the two communities defined by residues 164 and 238. Most nodes have strong connections (high-weight links) to one of these communities and much weaker connections (very low-weight or absent links) to the other community. For example, position 237 is strongly linked to 164, but is weakly connected to nodes from the 238 community. This non-uniform node connectivity agrees with a laboratory evolution study [68], which reported that E104K is preferentially selected in G238S trajectories, while E240K is more frequently found in R164S trajectories. Therefore, the network can be used to make inferences on evolutionary contingency effects, at least for the two main fitness peaks present in extended-spectrum evolution. The observation that other residue positions are frequently linked with both 164 and 238 in the network, even if I typically find a preference for one or the other, indicates that the evolutionary divergence associated with the two fitness peaks is only partial.

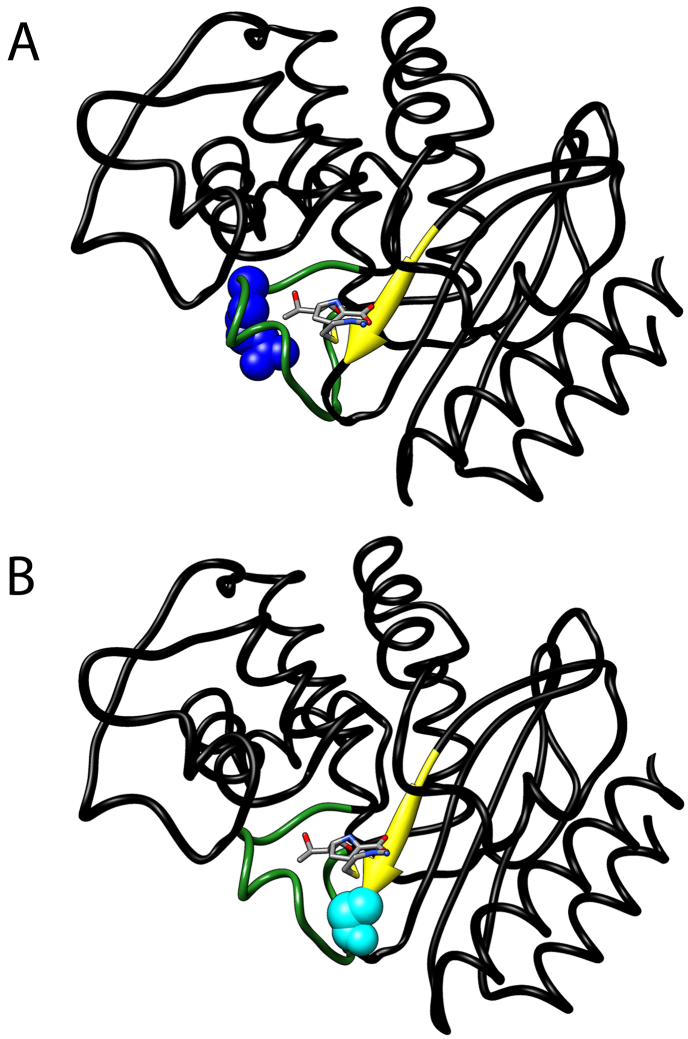

Figure 4.2: (A) Mutations at residue 164. An arginine to serine (or arginine to histidine) substitution at position 164 (blue spheres) has been hypothesized to collapse the critical Ω-loop (green) in the active site, thus opening the active site to β-lactams with larger side chains [50, 74, 109] (PDB ID 1zg6 [110]). The ligand (shown in stick representation) is an N-Formimidoyl-Thienamycine pseudo-substrate from PDB ID 1jvj [111]. (B) Mutations at residue 238. A glycine to serine (or glycine to alanine) substitution at position 238 has been hypothesized to expand the active site by either repositioning the B3 β-strand (positions 235-240) [112] (yellow) or by tilting the Ω-loop (green) (positions 161-179) [113] that connects the two sub-domains of the protein. Mutations at both positions are associated with increased resistance to third generation cephalosporins [112, 114].

Mapping adaptive communities of residues to the TEM tertiary structure

In the previous section, I found that as selective pressures leave recognizable footprints on the general TEM network's connectivity, leading to communities of nodes with common functional effects. However, here too the amino acid positions within network modules are not necessarily physically close in the protein's tertiary structure, as interactions are defined genetically (functionally) rather than physically. Figure 4.3 maps nodes (mutant positions) belonging to the two major communities in the TEM extended spectrum network onto the tertiary structure of the TEM enzyme (PDB ID: 1ero). Again, neither community appears to be physically localized to a defined area of the protein.

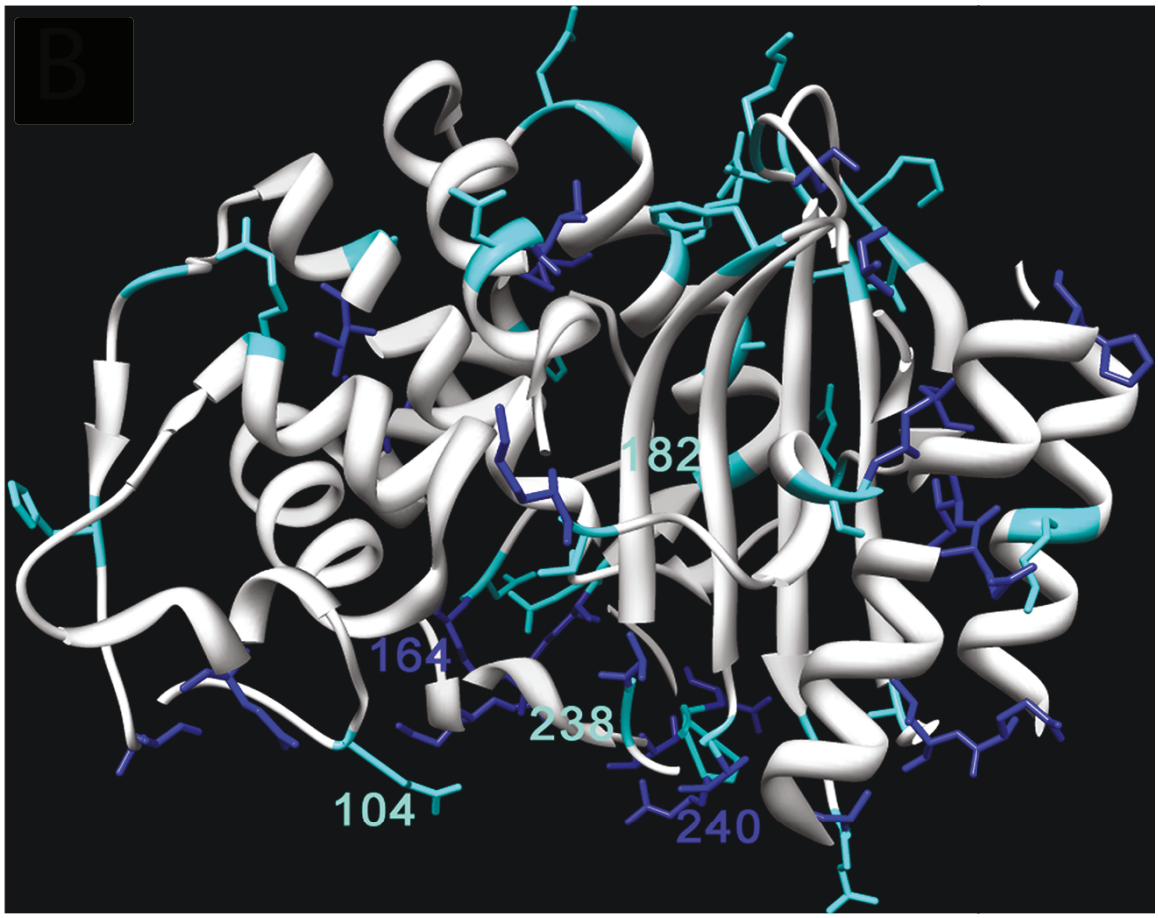

Figure 4.3: Locations of amino acid residues in the TEM extended spectrum network communities on the TEM tertiary structure (PDB 1ero). Residues in the TEM extended spectrum adaptation network (Figure 3.1) are colored by community membership: light blue (community containing the active-site residue 238) and dark blue (community containing the active site residue 164). The communities do not map to distinct regions of the tertiary structure. Image created with UCSF Chimera [106].

4.3 Network node centralities analysis

I reasoned that by analyzing the connectivity of the TEM β-lactamase coevolution network, I could extract functional information about amino acid residue positions

in this enzyme. I focused the analysis on the extended-spectrum community, which is the adaptive community network based on the largest number of available mutant sequences.

I used three standard graph-theoretical node centrality metrics to identify important residue positions in the undirected, weighted network: (weighted) degree centrality, closeness centrality, and betweenness centrality.

Weighted degree centrality

The degree of a node is a local measure of this nodes importance in a network. Specifically, in a network (graph) $G(V, E)$, the importance of the node (vertex) v is only defined by its immediate set adjacent nodes or neighbors $N(v)$ (Equation 4.1) [115]. For a network with a node/vertex set V, degree centrality is the degree normalized by the total number of remaining nodes $|V| - 1$:

$$C_D(v) = \frac{\deg(v)}{|V|-1} = \frac{\sum_{u \in N(v)} w_{uv}}{|V|-1} \tag{4.1}$$

where w_{uv} corresponds to the weight of association between residue positions u and v, as defined in Equation 3.1.

Shortest path network centralities

The next two centralities, the weighted *closeness* 4.3 and the weighted *betweenness* centralities are based on shortest paths, a.k.a. *geodesics* in the network. A network

path in $G(V, E)$ is a sequence of vertices $P = (v_1, v_2, ..., v_n), v_i \in V$, such that v_i is adjacent to v_{i+1} for $1 \leq i \leq n$. Letting $e_{i,i+1} \in E$ be the edge incident on both v_i and v_{i+1}, and given a real-valued weight function $f : E \to \mathbb{R}$, the length of the path P_{uv} between vertices $u = v_1$ and $v = v_n$ is $\sum_{i=1}^{n-1} f(e_{i,i+1})$. The shortest path P^*_{vw} minimizes this length, and the distance between u and v, $d_G(v, w)$ is the length of this shortest path:

$$l(P_{uv}) = \sum_{i=1}^{n-1} f(e_{i,i+1}) \tag{4.2}$$

$$P^*_{uv} = \text{argmin}_{P_{uv}} l(P_{uv}) \tag{4.3}$$

$$d_G(v, w) = l(P^*_{vw}) \tag{4.4}$$

Because in my network, weights represent strength of association rather than distance, the weight function is defined as

$$f(e_{uv} \in E) = \frac{1}{w_{uv}} \tag{4.5}$$

where e_{uv} is the network edge between nodes u and v and w_{uv} is the corresponding weight of association between residue positions u and v, as defined in Equation 3.1.

The use of network paths to assess the functional importance of corresponding residue positions follows the assumption that adaptation involves information transfer across the network (Section 4.1). By applying *shortest paths* algorithms, I furthermore

assume that adaptive trajectories are optimal or parsimonious and that evolution proceeds sequentially using the minimum number of mutations that lead to increased fitness. In Section 5.6.2, I consider algorithms in which evolution is not assumed to proceed by an optimal overall trajectory. Instead, trajectories are optimized at each individual step, and so are the corresponding paths in the network, which are not necessarily the shortest ones.

Weighted closeness centrality

Closeness centrality is a global measure of a node's importance in a network is defined as:

$$C_C(v) = \frac{1}{\sum_{w \in V \setminus v} d_G(v,w)} \tag{4.6}$$

where $d_G(v, w)$ (Equation 4.4) is the shortest path distance between nodes v and w in the graph G. The summation in the denominator is over all nodes w in the set V of all nodes in the network (excepting v) that are reachable from v.

Weighted betweenness centrality

Betweenness centrality, an alternate global measure of a node's importance in a network is defined as:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4.7}$$

where $\sigma_{st}(v)$ is the total number of distinct shortest paths connecting all pairs of network nodes (s, t) that pass through node v, and σ_{st} is the number of all distinct

shortest paths connecting node s to node t in the network. The inequality requirements ensure that only paths that pass through the node of interest v, but do not start or end at it, are counted. In order to find *all* the shortest paths in the network between a given pair of nodes, I developed a simple bidirectional search algorithm. The length of the shortest path was calculated using the weighted distance definition as in the denominator as in Equation 4.6.

I interpret node betweenness centrality as a measure of information flow through a given node from the entire community network [22]. Again, this assumes that information (related to selection for a certain function) flows along optimal trajectories (combinations of mutations) during the adaptive process. In Chapter 6, I also apply a betweenness centrality based on random walks rather than shortest paths to address a model of adaptation in which new functions are not not required to evolve in the most parsimonious way.

Node centralities and functional effects of mutated positions

Mutated residues that are highly ranked by the network centrality metrics have known functional impact previously described in the literature. While many of the mutations known to contribute to extended-spectrum resistance are highly frequent, the network also ranks highly the less frequent mutations with known contributions (Table 4.1).

Residue Number*	Count within database	Node Degree Rank	Node Closeness Rank	Node Betweenness Rank	Described function	References
104	48	1	1	1	The long K side chain of E104K mutants interacts directly with carboxylic acid group of the substrate.	[39, 116]
164	48	2	2	2	Collapses the Ω-loop, resulting in an active site with greater accessibility.	[90]
238	38	3	3	3	Expands the active site either by repositioning the B3 β-strand or by tilting the Ω-loop	[117, 118]
240	31	4	5	4	Interacts with substrate; possibly stabilizing.	[89, 118, 119]
182	27	5	4	5	Increases the thermodynamic stability of the protein; could suppress misfolding and aggregation caused by other mutations.	[50, 94, 95]
265	20	7	8	9	Unknown mechanism. Possibly important for enzyme stability.	[45, 57]
237	9	6	7	8	Introduces another H-bond with carbonyl group of the substrate's β-lactam ring.	[38, 89]
173	5	9	6	6	Likely mildly adaptive, specific to certain extended spectrum cephalosporins.	[59, 61, 65]
120	3	17	14	8	Unknown mechanism. Possibly important for enzyme stability.	[53, 102]
254	3	8	10	NA	Unknown mechanism. Possibly stabilizing.	[39, 45, 53]
51	2	15	17	7	Unknown mechanism. Possibly important for both enzyme activity and stability.	[92]
268	2	10	11	8	Unknown mechanism. Possibly stabilizing.	[53, 63, 64, 91]

Table 4.1: The mutated residue positions most important for TEM extended-spectrum antibiotic resistance, according to measures from network theory (node centralities): Weighted degree (4.3), closeness (4.3), and betweenness 4.3 centraty ranks are shown. *Residue numbers are based on the Ambler TEM β-lactamase numbering scheme. [78].

4.4 Central network paths and predicted evolutionary trajectories

Each link in the TEM coevolution network represents a potential step within an adaptive evolutionary trajectory. Once individual coevolution links between pairs of mutated residues are put into the context of network, the definition can be expanded to adaptive trajectories of any length. Although, by construction, all two-node paths have been seen in natural or laboratory evolution, by defining longer paths within the network, I should be able to derive evolutionary trajectories consisting of more than two mutations.

In *single-node* shortest-path betweenness centrality, a node's importance to the overall connectivity of the network is measured by the number of shortest paths that pass through it. In the generalization from a single node to a path of connected nodes, I define the betweenness of a path P_{uv} between two nodes u and v as the number of shortest paths that pass through P_{uv} but do not start or end at u or v.

$$C_B(P_{uv}) = \sum_{\substack{\{s,t,u,v\}\in V \\ \{s,t\}\cap\{u,v\}=\varnothing}} {\sigma_{st}(P_{uv})}/{\sigma_{st}} \tag{4.8}$$

4.4.1 Central network paths to identify adaptive evolutionary trajectories

I chose to analyze two-edge (three-node) shortest paths, each of which represents an evolutionary trajectory that produces a triple mutant sequence, because they are the most tractable to enumerate and explore. I identified evolutionary trajectories of special significance for adaptive evolution based on shortest path betweenness-centrality a metric that can be interpreted to measure the efficiency of information transfer through the network.

I investigated the significance of betweenness centrality as an indicator of potential adaptive evolution. Below I show that: (1) the triple mutant trajectories listed in Table 4.2 as of potential special significance for adaptation are enriched for triple mutants that have been previously reported; (2) the reported triple mutant combinations consistently increase extended-spectrum resistance over constituent double mutants, confirming they resulted from a functional selection; (3) using reported triplet mutants as a proxy for increased resistance, I can estimate the success rate of the coevolution network path betweenness centrality metric (Equation 4.8). This success rate is considerably higher than what would be anticipated based on the simple assumption that the most successful triplet combinations consist of the most frequent single mutations in the database. Together, these three lines of evidence strongly support the predictive value of the extrapolation to triple mutant evolutionary

trajectories.

Nonzero betweenness centrality triplets frequently identify triple mutants associated with extended-spectrum resistance.

A subset of all possible three-node paths in the network (48 out of 214) had a shortest path betweenness centrality greater than zero. These triple mutant trajectories are listed in Table 4.2, ranked in descending order of betweenness centrality. Shown is also the number of times (count) that each residue position in the trajectory was seen mutated in the 201 extended-spectrum resistant TEM sequences in the database. Note that many nonzero betweenness trajectories consist of at least one infrequent mutation and therefore would not have been predicted as critical based on frequency alone. Note also that these 48 triplets consist of combinations of only 16 residue positions out of a total of 55 residue positions in the network. These positions could be of special significance for the evolution of extended-spectrum β-lactamase resistance.

In addition to listing nonzero shortest path betweenness centrality trajectories, Table 4.2 also shows which of these trajectories were previously reported in clinical or experimental studies. Trajectories are listed in descending order of betweenness centrality value. I noted that this list is rich in triple mutant combinations that have been previously described in clinical or experimental reports, with 23 previously described out of the 48 predicted paths. In addition, I found a strong association between the chance of having been previously reported and the corresponding shortest

path betweenness centrality value: while all of 10 top-ranked triplet paths are already known; only 1 of the 6 paths with the lowest positive betweenness centrality (value of 1) is known.

Evolutionary Trajectory	Betweenness centrality	Database count*	Previously reported
238_104_164	96	48,48,38	TEM-008, TEM-134, [68]
173_164_104	92	48,48,5	[39]
182_104_164	66	27,48,48	TEM-043,TEM-063, [68]
240_164_104	62	31,48,48	TEM-046
268_240_164	41	2,31,48	TEM-136, [68]
120_238_104	39	3,38,48	[65] , [68]
39_240_164	32	1,31,48	[68]
237_164_104	28	9,48,48	TEM-130, [68]
104_238_153	23	48,38,9	TEM-021, [68]
240_164_173	22	31,48,5	TEM-132, [68]
104_164_40	18	48,48,1	
238_104_51	16	38,48,2	
215_104_164	15	48,38,20	TEML-136
104_238_265	15	2,48,48	[39, 57, 68]
39_240_238	12	1,31,38	
182_104_51	11	27,48,2	
173_164_51	9	5,48,2	
215_104_238	8	2,48,38	
182_238_120	7	27,38,3	[65]
240_164_51	6	31,48,2	
224_164_173	6	3,48,5	[59]
173_164_237	6	5,48,9	[39, 68]
224_164_240	5	3,48,31	
173_164_40	4	27,38,20	
182_104_215	4	27,48,2	
182_238_153	4	5,48,1	[68]
240_238_153	4	31,38,9	
182_238_265	4	27,38,9	[68]
51_164_40	3	20,38,31	
40_164_240	3	2,31,9	
224_164_251	3	3,48,2	
51_164_237	3	1,48,31	
268_240_237	3	2,48,1	TEM-136, [68]
265_238_240	3	2,48,9	[68]

Table 4.2: Prediction of critical triple mutant evolutionary trajectories ...

Evolutionary Trajectory	Betweenness centrality	Database count*	Previously reported
39_240_237	2	3,38,9	[68]
39_240_268	2	2,38,3	
120_238_153	2	31,38,3	[65]
240_238_120	2	3,48,9	
120_238_265	2	1,31,2	
268_238_120	2	2,38,9	
268_238_153	2	3,38,20	
224_164_237	2	1,31,9	[68]
224_164_40	1	2,38,20	
237_164_40	1	9,48,1	
51_104_215	1	48,48,3	
104_164_224	1	20,38,9	[68]
265_238_153	1	3,48,1	
268_238_265	1	2,48,2	

Table 4.2: Prediction of critical triple mutant evolutionary trajectories in the extended-spectrum antibiotic resistance community. Triple mutant trajectories are shown as an ordered list of three residue positions, where an ordered pair represents a link in the network. The shortest path betweenness centrality is listed for each triple mutant trajectory, in descending order. I interpret the betweenness centrality of a trajectory as a representation of information flow through this path for the entire extended spectrum resistance network. The count shows the number of times that each residue position in the trajectory was seen mutated in the 201 extended-spectrum resistant TEM sequences in the database. Note that many trajectories consist of at least one infrequent mutation and therefore would not have been predicted as critical based on frequency alone. Some of the triple mutants have been seen either alone or in combination with other mutations in clinical isolates, in laboratory-evolved isolates that were included in the database, or in laboratory-evolved isolates that were not in the network database.

Reported extended-spectrum resistant mutants increase extended-spectrum resistance.

I interpreted the occurrence of a given path (evolutionary trajectory) in clinical isolates or published laboratory evolution experiments as an indication of likely fitness

advantage, i.e. of likely increased resistance to extended-spectrum β-lactam antibiotics. This interpretation was experimentally confirmed, using CTX as a representative extended-spectrum β-lactam antibiotic as done previously in similar studies [51, 53, 68]. Site-directed mutagenesis of TEM β-lactamase was used to obtain TEM β-lactamase mutants. Resistance to CTX was determined using a gradient plate assay [120].

48 out of a possible 214 three-node shortest paths in the extended spectrum community network had nonzero betweenness centrality, so the experiments were focused on the corresponding 48 triple mutants. Because all triplets represent a mutational trajectory and are therefore ordered, I compared the activity of each triplet to each possible trajectory (i.e. initial pair of mutations) that led to it.

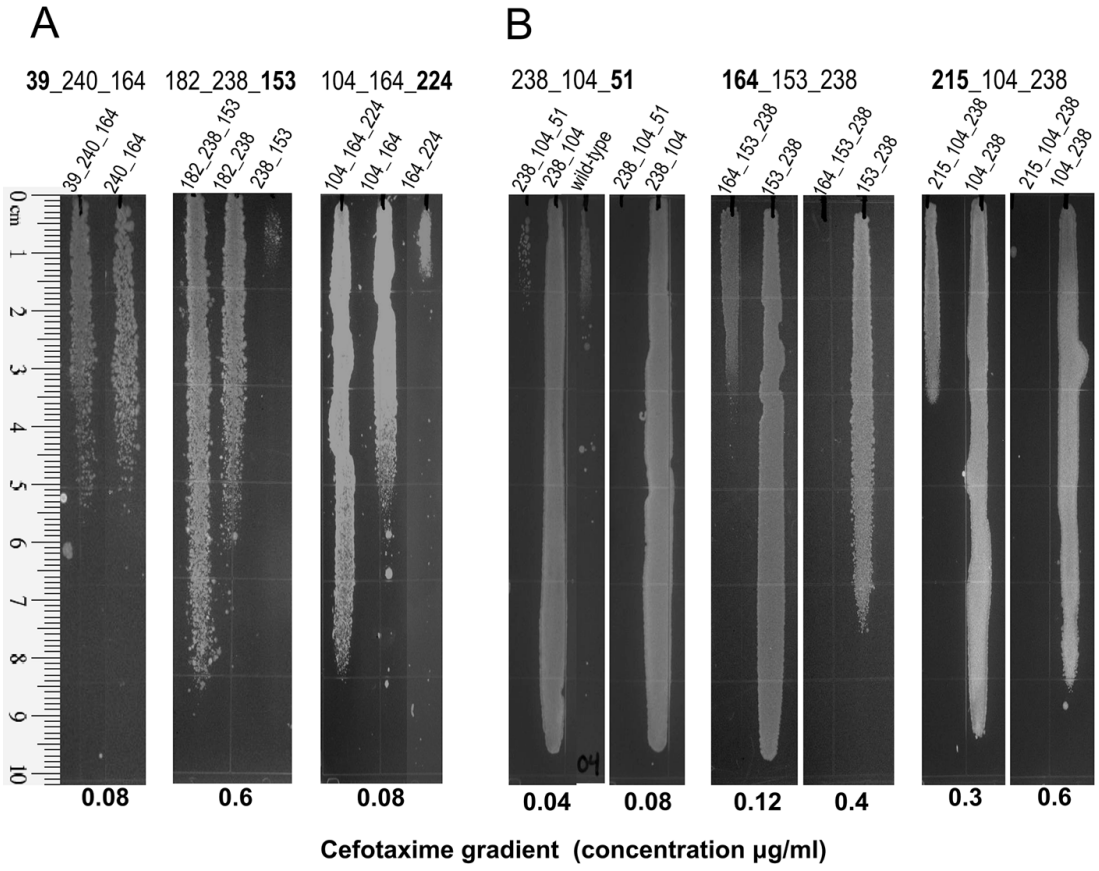

Figure 4.4: Cefotaxime plate growth assays for selected clones. Cultures of cells expressing the β–lactamase mutants listed at the top of the gradients were stamped on LB plates containing a CTX gradient. The direction of the gradient is from top (minimal concentration) to bottom (maximal concentration). The maximal concentration of the gradient is listed at the bottom. Note that in part B more than one concentration is shown to cover the wide range of resistance phenotypes of the panel of mutants being tested. (A) Two mutant triplets predicted to be of special significance by my analysis but that were not present in the sequence database used to build the network but were subsequently reported in [68], and a third triplet also predicted by my analysis but that showed only a marginal increase. Only the doublet with the highest level of resistance is shown. (B) Triplets with the strongest negative functional interactions. The mutation responsible for the negative effect is highlighted in bold.

15 triple mutants that span a range of shortest path betweenness centrality values were tested by measuring growth (in centimeters) along an LB agar plate containing a CTX gradient. Of these 15 triple mutant trajectories, 9 had already been described, and 6 were new. The results (Table 4.3) show that observed mutants consistently increased resistance over both ordered, constitutive pairs: 8 out of the 9 previously reported triple mutants. By contrast, none of the non-observed mutant sequences I tested improved on both constitutive double mutants. These results confirm the intuitive notion that combinations of mutants that increase fitness are more likely to have been selected during evolution of TEM β-lactamase under extended-spectrum antibiotic selection and therefore reported.

Triplet	Betweenness centrality	Reported?	Triplet resistance [cm]	Pair 1	Pair 1 resistance [cm]	Triplet change from pair 1 [cm]	Pair 2	Pair 2 resistance [cm]	Triplet change from pair 2 [cm]
104_164_173	92	Y	16.50	104_164	8.42	8.07	164_173	6.95	9.54
182_104_164	66	Y	16.86	182_104	2.82	14.04	104_164	8.42	8.44
39_240_164*	32	Y	9.10	39_240	2.16	6.94	240_164	9.48	-0.38**
104_238_153	23	Y	17.65	104_238	16.83	0.82**	238_153	11.5	6.15
240_164_173	22	Y	17.48	240_164	9.48	8.01	164_173	6.95	10.53
104_164_40	18	N	5.06	104_164	8.42	-3.36	164_40	2.13	2.93
238_104_51	16	N	1.88	238_104	16.83	-14.95	104_51	1.65	0.23**
104_238_265	15	Y	19.40	104_238	16.83	2.57	238_265	10.84	8.56
39_240_238	12	N	9.59	39_240	2.16	7.43	240_238	12.04	-2.45
182_104_51	11	N	2.54	182_104	2.82	-0.28**	104_51	1.65	0.89
173_164_51	9	N	1.79	173_164	6.95	-5.16	164_51	1.9	-0.11**
215_104_238	8	N	11.26	215_104	2.39	8.87	104_238	16.83	-5.57
182_238_153*	4	Y	17.95	182_238	16.17	1.78	238_153	11.5	6.45
120_238_153	2	Y	14.36	120_238	7.22	7.14	238_153	11.5	2.86
104_164_224*	1	Y	9.31	104_164	8.42	0.89***	164_224	3.9	5.41

Table 4.3: Extended-spectrum network triple mutant trajectories reistance measurements. Each mutant trajectory is shown as an ordered list of three mutated residue positions (column 1). Ordered pairs of mutated residue positions represent a link in the extended-spectrum network. The shortest path betweenness centrality is listed for each trajectory (column 2). 9 of the 15 tested trajectories were reported in clinical or directed evolution isolates (column 3). The level of CTX resistance (an indicator of extended-spectrum antibiotic resistance) is shown in centimeters of linear growth on a 0.04 μg/ml CTX gradient. The level of resistance is shown for each triple mutant trajectory (columns 1 and 4) and its two ordered constituent double mutants (columns 5 and 6, and 8 and 9). The differences representing the improvement in resistance conferred by the triple mutant trajectory with respect to each double mutant, is shown in columns 7 and 10. Trajectories marked with * had not been reported when this work was done and were not included in input to the network. They were subsequently reported [68]. (**) Triplet improvement over pair is outside the margin of standard error (for the number of replicates (n) refer to Table B.1). (***) Improvement is outside the margin of standard error if the variability between gels is subtracted out (Figure 4.4).

Betweenness centrality vs. random choice of frequently occurring mutations

The experimental results show that observed triple mutants consistently increase CTX resistance. Thus, I reasoned that to be reported as having extended-spectrum resistance is a viable proxy for having increased fitness. By this logic, the predictive success rate of my method is 23 out of 48. To demonstrate that this success rate is not due to chance, I ran a simulation in which I randomly selected 48 triple mutants only from TEM residue positions previously reported in association with extended-spectrum antibiotic resistance. I sampled these positions according to their mutation frequency in the database. The 10,000 random sets of 48 triple mutants selected in this way followed a normal distribution, as expected by the central limit theorem. This simulation produced an average success rate of 12.8 ± 3.08 observed triplets out of 48. Since the success rate of 23 out of 48 is well outside the range of standard error, this analysis was able to extrapolate triple mutant trajectories from pairs of coevolving mutations more accurately than simply combining mutations of high frequencies and thus can be inferred to have predictive value.

4.4.2 Central network paths and pairwise functional interactions between mutated residues

Next, I investigated whether links connecting co-occurring pairs in this network represent synergistic functional interactions. The individual vs. combined effects of the mutations in the mutant triplets from Table 4.3 were tested, and the results are listed in Table 4.4: A difference between adding the individual fitness effects of two mutations (M1+M2) and the combined fitness effect of the double mutant (M1_M2) indicates of either synergistic (positive difference) or antagonistic (negative difference) interactions.

Of the mutation pairs in Table 4.4, six have been previously reported as having synergistic functional interactions. In agreement with previous reports, the experimental tests here show significant positive interactions in five known synergistic interaction cases, with the exception of E104K_M182T. The experimental test shows no synergy, but the mutations' individual effects combine in an additive way. I also found two new examples of synergistic interactions involving I173V (E104K_I173V and II73V_E240K),

Five examples of antagonistic interactions, not previously reported, were experimentally identified. The high count of negative interactions in the tested pairs is surprising given that each connected pair of nodes represents pairs of mutations that co-occur in at least one sequence. I assume that the reported sequences containing these mutations with negative functional effects must have additional mutations pro-

ducing an overall increase in resistance. Therefore, these mutations were selected for based on the specific sequence context in which they occurred. Similarly, I found a number of significant antagonistic interactions in the triple mutants tested (8 out of 27; Section 4.4.3). Thus, even in a network model representation, intrinsically biased in favor of synergistic interactions, I frequently find antagonistic interactions among linked mutation pairs. This observation highlights the pervasiveness of antagonistic pairwise interactions in TEM extended-spectrum resistance evolution. The reason these residues are identified as functionally associated could be because they tend to only occur with other mutations in the background. Therefore, while their combined effect is negative when tested against the wild type (TEM-1) context, they may exhibit synergy in the context of additional mutations. Network paths or communities containing such residue pairs could provide insights into the context relevant to these interactions. In summary, links within the network are rather more indicative of *potential functional* dependencies than specifically of interactions that are *positive* in the context of the wild type. Examining the non-wild-type sequence context in which such negative interactions occur would point to important complex compensatory mechanisms. Additional analysis of the network communities in which negatively interacting pairs of residues occur, could provide further insights into such complex compensatory effects.

M1	M1 growth [cm]	M2	M2 growth [cm]	M1_M2 growth [cm]	M1_M2-(M1+M2) [cm]	Significant functional interaction*
Q39R	2.09	G238S	9.61	7.76	-2.35	
Q39R	2.09	E240K	1.58	2.16	0.08	
L40W	2.08	R164H	3.43	2.13	-1.79	antagonistic
L51P	1.93	E104K	2.20	1.65	-0.89	antagonistic
L51P	1.93	R164H	3.43	1.90	-1.87	antagonistic
E104K	2.2	H153R	2.17	2.73	-0.05	
E104K	2.2	R164H	3.43	8.42	4.38	synergistic**
E104K	2.2	I173V	2.10	10.84	8.13	synergistic
E104K	2.2	M182T	2.15	2.82	0.06	**
E104K	2.2	K215E	1.90	2.39	-0.12	
E104K	2.2	A224V	1.86	1.90	-0.57	
E104K	2.2	G238S	9.61	16.83	6.61	synergistic**
R120S	1.94	G238S	9.61	7.22	-2.74	antagonistic
H153R	2.17	G238S	9.61	11.50	1.31	
R164H	3.43	I173V	2.10	6.95	3.01	synergistic**
R164H	3.43	A224V	1.86	3.90	0.20	
R164H	3.43	E240K	1.58	9.48	6.06	synergistic**
I173V	2.1	E240K	1.58	3.62	1.53	synergistic
M182T	2.15	G238S	9.61	16.17	6.00	synergistic**
K215E	1.9	G238S	9.61	6.34	-3.58	antagonistic
G238S	9.61	E240K	1.58	12.04	2.44	
G238S	9.61	T265M	N/A	10.84	N/A	

Table 4.4: Experimentally determined functional interactions between single mutations in the extended-spectrum antibiotic resistance community network. Mutated residues (columns 1 and 3) and their individual CTX resistance levels (columns 2 and 4) are compared to resistance levels when they occur together in the same sequence (column 5). The level of CTX resistance (an indicator of extended-spectrum antibiotic resistance) is shown in centimeters of linear growth on a 0.04 μg/ml CTX gradient. The difference between the combined effect (column 5) and the sum of the individual effects (column 2 + column 4), which represents synergistic or antagonistic functional functional interactions, is shown in column 6. In column 7 a significantly synergistic or antagonistic interactions are reported when the difference in column 6 the margin of standard error for a given number of replicates (Table B.1)
** Six interactions were previously reported as synergistic.

4.4.3 Central network paths and functional interactions in triple mutants

The level of resistance of pairs of mutations present in nonzero betweenness centrality trajectories to their constituent mutations is shown in Table 4.5. When I compared the effect of single mutations on mutation pairs in triple mutant trajectories, I found 8 significantly antagonistic functional interactions versus 19 synergistic ones. Overall, the analysis revealed a surprising number of antagonistic interactions: 22 out of 60 tested interactions had a negative trend, which was statistically significant in 13 cases. Thus, while links in the network represent potential functional interactions, these links are not necessarily indicative of synergistic functional interactions. In fact, their interactions are frequently antagonistic. Because all the pairs tested co-occurred in at least one TEM sequence, I inferred that the interaction was synergistic in the context of original sequence, i.e. in the presence of additional mutations, similarly to the effect observed with pairwise interactions (Section 4.4.2). Here too, analysis of the network communities in which negatively interacting pairs of residues occur, could provide further insights into the underlying complex compensatory mechanisms.

Table 4.5: Epistasis within mutant triplets

M1	M1 growth [cm]	M2	M2 growth [cm]	M1_M2 (M1+M2) [cm]	M1_M2– functional effect	Significant
Q39R	2.09	E240K_R164H	9.48	9.10	-0.88	
Q39R	2.09	E240K_G238S	12.04	9.59	-2.95	negative
L40W	2.08	E104K_R164H	8.42	5.06	-3.85	negative

Table 4.5: Epistasis within mutant triplets (cont.)

M1	M1 growth [cm]	M2	M2 growth [cm]	M1_M2 (M1+M2) [cm]	M1_M2– functional effect	Significant
L51P	1.93	M182T_E104K	2.82	2.54	-0.62	
L51P	1.93	I173V_R164H	6.95	1.79	-5.5	negative
L51P	1.93	G238S_E104K	16.83	1.88	-15.29	negative
E104K	2.20	R164H_L40W	2.13	5.06	2.32	positive
E104K	2.20	R164H_A224V	3.90	9.31	4.80	positive
E104K	2.20	K215E_G238S	6.34	11.26	4.31	positive
E104K	2.20	I173V_R164H	6.95	16.49	8.93	positive
E104K	2.20	G238S_T265M	10.84	19.40	7.95	positive
E104K	2.20	G238S_H153R	11.50	17.65	5.54	positive
R120S	1.94	G238S_H153R	11.50	14.36	2.51	
R120S	1.94	E240K_G238S	12.04	12.92	0.53	
H153R	2.17	R120S_G238S	7.22	14.36	6.56	positive
H153R	2.17	E104K_R164H	8.42	11.5	2.50	positive
H153R	2.17	E104K_I173V	10.84	2.65	-8.77	negative
H153R	2.17	M182T_G238S	16.17	17.95	1.20	
H153R	2.17	E104K_G238S	16.83	17.65	0.24	
R164H	3.43	Q39R_E240K	2.16	9.10	5.10	positive
R164H	3.43	E104K_H153R	2.73	11.5	6.93	positive
R164H	3.43	M182T_E104K	2.82	16.86	12.20	positive
R164H	3.43	E104K_I173V	10.84	16.49	3.81	positive
R164H	3.43	H153R_G238S	11.5	6.00	-7.34	negative
I173V	2.10	R164H_L51P	1.90	1.79	-0.62	
I173V	2.10	E104K_H153R	2.73	2.65	-0.59	
I173V	2.10	E104K_R164H	8.42	16.49	7.56	positive
I173V	2.10	E240K_R164H	9.48	17.48	7.49	positive
M182T	2.15	E104K_L51P	1.65	2.54	0.33	
M182T	2.15	E104K_R164H	8.42	16.86	7.88	positive
M182T	2.15	G238S_H153R	11.50	17.95	5.89	positive
K215R	1.90	E104K_G238S	16.83	11.26	-5.88	negative
A224V	1.86	E104K_R164H	8.42	9.31	0.62	
G238S	9.61	E104K_L51P	1.65	1.88	-7.79	negative
G238S	9.61	Q39R_E240K	2.16	9.59	-0.59	
G238S	9.61	K215R_E104K	2.39	11.26	0.85	
E240K	1.58	Q39R_R164H	3.70	9.10	5.41	positive
E240K	1.58	R164H_I173V	6.95	17.48	10.54	positive
E240K	1.58	R120S_G238S	7.22	12.92	5.71	positive
T265M	N/A	E104K_G238S	16.83	19.40	N/A	

Table 4.5: Epistasis within mutant triplets (cont.)

M1	M1 growth [cm]	M2	M2 growth [cm]	M1_M2 (M1+M2) [cm]	M1_M2– functional effect	Significant

Table 4.5: Experimentally determined functional interactions between single mutations in the extended-spectrum antibiotic resistance community network. Mutated residues (columns 1) and residue pairs (column 3) and their corresponding CTX resistance levels (columns 2 and 4, respectively) are compared to resistance levels when they occur together in the same sequence (column 5). The level of CTX resistance (an indicator of extended-spectrum antibiotic resistance) is shown in centimeters of linear growth on a 0.04 μg/ml CTX gradient. The difference between the combined effect (column 5) and the sum of the individual effects (column 2 + column 4), which represents non-additive functional interaction, is shown in column 6. In column 7 a significantly synergistic or antagonistic interaction are reported when the difference in column 6 the margin of standard error for a given number of replicates (Table B.1)

Central network paths and role of sequence context

The observed disconnect between co-occurrence and the CTX resistance phenotype of pairs of mutations included in the network suggests that the adaptive value of a given mutation or mutation pair is highly dependent on sequence context. Thus, an accurate assessment of the contribution of a given mutation to adaptation involves testing the effect of the mutation in the presence of different additional mutations, i.e. in a range of sequence contexts. Table 4.6 shows the impact of 14 of 16 mutations identified as of likely significance for extended-spectrum β-lactamase resistance based on shortest path betweenness centrality. Both the average effect (column 5) and the range of effects (in centimeters of continuous growth; column 6), obtained in a variety of sequence contexts, are shown. The number of sequence contexts tested (7

on average) is listed in column 4. The sequences tested and their measurements are listed in Table B.1. The main observations are as follows.

- The experimental results show a relationship between average phenotypic effect and representation in my database, with frequent mutations ($n > 4$) having a clear average positive effect (≥ 1 cm).

- The average effect of infrequent mutations ($n < 5$), is negative (-1.3 ± 1.6 cm), questioning the relevance of these mutations for extended spectrum resistance. The large negative effects that some of these mutations, L51P (-14.95 cm), K215E (-5.57 cm); R120S (-2.39 cm) have in specific contexts suggests that they are functionally important but that their effects are highly context-dependent. The two strongest antagonistic effects I detected for infrequent mutations, those of L51P and K215E, are shown in Figure 4.4(B).

- In agreement with the non-additive functional interactions analysis presented in Tables 4.4 and 4.5, most mutant positions exhibit a wide range of functional interactions, including synergistic, antagonistic, and neutral effects. The effect of the R164H mutation on CTX resistance for example ranges from -5.5 cm to $+14.04cm$, that of H153R, from -8.19 cm to 7.14 cm. This is a clear example of the role of sequence context in determining the potential functional impact of a mutation.

- R164H and L51P, two mutations with a known effect on resistance phenotype, had

large negative impacts in some sequence contexts: 5.5 and 14.95cm, respectively. These observations imply that a strong antagonistic effects may be as indicative of functional interactions as are synergistic effects. Therefore, the large antagonistic effects K215E (5.57cm) and L40W (3.4cm) suggest an important role for these residue positions that is only revealed in specific sequence contexts, although this remains to be experimentally confirmed.

- The network analysis identifies three positions whose phenotypic impact on extended spectrum resistance had not been previously identified: 265 (average 1.9 cm, up to 2.6), 153 (average 1.0 cm, up to 7.1), and 120 (average 0.4 cm, up to 2.9). The effect of 153 is strikingly sequence-context dependent, with values ranging from 8.19 to +7.14cm, which may explain why the role of this mutation has been hard to experimentally demonstrate.

Mutant position	Database count contexts	Mutation tested [cm]	Number of tested sequence [cm]	Average effect	Interval (min, max)
164	48	R164H	13	4.18	(- 5.50, 14.04)
104	48	E104K	15	4.04	(- 0.28, 9.54)
238	38	G238S	11	8.03	(0.23, 14.63)
240	31	E240K	8	3.96	(0.07, 10.53)
182	27	M182T	6	3.92	(0.62, 8.44)
265	20	T265M	2	1.90	(1.23, 2.57)
153	9	H153R	8	0.95	(- 8.19, 7.14)
173	5	I173V	8	3.82	(- 0.11, 8.64)
224	3	A224V	4	0.33	(- 0.30, 0.89)
120	3	R120S	4	0.43	(- 2.39, 2.86)
215	2	K215E	4	-2.89	(- 5.57, 0.19)
51	2	L51P	6	-3.69	(-14.95, 0.34)
268	2	N/A	N/A	NA	NA
40	1	L40W	3	-1.39	(- 3.36, 0.49)
39	1	Q39R	6	-0.56	(- 2.45, 0.58)

Table 4.6: Context dependence of extended spectrum mutations. Critical triple mutant trajectories (Table 4.2) contain only 16 unique individual residue positions (column 1). The number of sequences in experimental and clinical isolates that have this residue position mutated is shown in column 2. For each residue position, I tested the most frequent amino acid substitution in these sequences, with two exceptions: *K215E has equal frequency to K215R and K215Q in the extended-spectrum phenotype sequence database; **L40W and L40V have equal frequencies (column 3). Cefotaxime resistance of each mutation (centimeters of linear growth on a 0.04 μg/ml CTX gradient) was tested in a variety of sequence contexts. Each context consists of the relevant mutation plus different additional mutations, all of which are found in the critical triple mutant evolutionary trajectories. The number of sequence contexts tested is shown in column 4 and the different mutant combinations comprising each sequence context are shown in Table B.1. Averaging the effect of each mutation across all its sequence contexts yields a measure of its global contribution to extended-spectrum antibiotic resistance (column 5). In general, the effects are highly dependent on sequence context, as shown by the wide range of outcomes (column 6).

Representation of residues with known functional significance in central network paths.

The 48 triple mutant paths I identified as of special significance (Table 4.2) consist of different combinations of only 16 residue positions (listed in Table 4.6, column 1). These include 10 positions with a demonstrated effect on extended-spectrum β–lactamase resistance, out of 12 known to date [74]. The two false negatives (positions missing from analysis) are 175 and 179, mutations in each of which arises independently only once in my extended-spectrum sequence database. 175 is one of a number of positions in the Ω–loop (involved in active-site formation) that are known to play a role in extended-spectrum resistance [74, 105]. 179 was previously reported in a clinical isolate [61, 121] and in several experimental isolates [46] but appears to have a narrower substrate specificity than other mutations present in the extended-spectrum network community [74]. My analysis suggests that the remaining 6 mutant positions present in the nonzero betweenness triple mutant paths (40, 120, 153, 215, 224, and 265) should be considered as potentially important for adaptation.

4.5 Conclusions

Prevalence of antagonistic interactions in network

Given that the network is largely constructed with mutations that have experienced some degree of positive selection, and that mutant positions are linked when they

co-occur in the same sequence, I expected a predominance of positive interactions. To my surprise, I found that a large number of functional interactions within double and triple mutants were antagonistic. Because all the pairs of mutations tested co-occurred in at least one TEM sequence, I inferred that the interaction was positive in the original sequence, i.e. in the presence of additional mutations. From this, I conclude that examining the non-wild-type sequence context in which otherwise negatively interacting residues are allowed to occur together, would reveal the principles behind important complex compensatory mechanisms. Analysis of the network communities around negatively interacting pairs of residues could provide insights into potential compensatory contexts.

Significance of central network paths

By connecting individual nodes (representing mutated residue positions), paths through the extended spectrum network define potential evolutionary trajectories. Path centrality metrics allowed me to extend the trajectories beyond the pairs of co-occurring nodes used to build the network. I focused on combinations of three mutations, which are the most experimentally tractable ones. The basic hypothesis was that genetic adaptation necessitates a specific combination of functional milestones, where each amino acid mutation represents a potential milestone. According to this hypothesis, combinations of mutations that facilitate information flow through the network should contribute prominently to genetic adaptation. I used shortest

path betweenness centrality (a metric that can be interpreted as measuring a path's importance for information flow within the network) to identify trajectories of potential special significance for extended-spectrum β–lactamase resistance (Table 4.2). The following points support the special significance of triple mutant trajectories with nonzero betweenness centrality:

- They occur frequently in natural or experimental extended-spectrum β–lactamase evolution experiments (Table 4.2, column 4).
- The higher the betweenness centrality, the more likely they are to have been previously seen (Table 4.2).
- Presence of these mutations in reported (previously seen) sequences is associated with increased CTX resistance, an indicator of extended-spectrum activity (Table 4.5).

Limitations of the central network paths method

My method for identification of paths of special significance for adaptation assumes that each mutant position has a discrete effect on adaptation and that this effect is sufficiently unique that adaptation requires a composite solution. Therefore, global suppressors (such as mutations at position 182) or mutations with a large impact on their own (S130G, associated with inhibitor resistance, and G238S conferring extend-edspectrum resistance) will not not be adequately accounted for by the information flow metric.

Furthermore, the high fitness extended-spectrum triple mutant 104_238_182 in the list of nonzero betweenness centrality triplets (Table 4.2) is not present in the network. Amino acid substitutions at 104_238_182 were the most frequent combination obtained from TEM-1 libraries subjected to CTX selection [68]. The presence of a global suppressor (182) and of a mutation with a large impact on its own (G238S) likely explains why this triple mutant combination is not among the nonzero betweenness paths in Table 4.2. However, parallel, divergent evolutionary trajectories identified by this study are enriched for triple mutant trajectories with high betweenness centrality (detailed in Table S6 and Text S1 Results). Overall, triple mutant trajectories with nonzero betweenness centrality are frequently contained within mutational trajectories parallel to E104K_M182T_G238S. Thus, my method is able to identify paths of special significance for genetic adaptation, although with decreased sensitivity to mutations with a large impact on their own and to global suppressors.

The coevolution network model presented here is based on co-occurrence frequencies of pairs of mutated residues and does not consider the context of other mutations in which these pairs co-occur. If a given pair of mutations always co-occurs in the context of a third mutation, for example, it is possible that their combined functional effect is only positive in this context. Expanding the Jaccard index metric (Section 3.2.4) to incorporate higher-order (beyond pairwise) interactions could be important for ensuring that only positive interactions between mutations are present in the network representation.

Chapter 5

A phylogeny-based network of extended spectrum TEM evolution

5.1 Introduction

The alignment-based network model from Chapters 3 and 4 may present a desirable alternative to phylogeny when the timespan of the evolution of new resistance is short and there are no complex evolutionary relationships between sequences, i.e. most sequences evolve independently from a common ancestor. However rapid evolution under strong selection for function can give rise to complex evolutionary relationships between sequences. Some of the extant sequences can actually be ancestral to other currently extant sequences, yet both the ancestral and the extant sequence could be present in different populations. For example the E104K_G238S (TEM-15) mutant is

ancestral to E104K_M182T_G238S (TEM-52), but both have been found to exist in a clinical setting. When counting the occurrence of mutations in these two sequences, they should not be regarded as having independently evolved directly from TEM-1. In cases like this one, phylogeny can be incorporated into the network analysis to (1) better estimate how many times pairs of mutated residues arise *independently* and (2) potentially reveal the ordering of mutation events. The construction of a directed network of temporally ordered mutations could help improve prediction of adaptive trajectories including the preferred order in which mutations are selected.

Molecular phylogenetics use the extensive information encoded in molecular sequences. Given a set of sequences from different species a likely evolutionary tree is inferred based on the common ancestor assumption. The topology of a phylogenetic tree is the specific branching pattern of that tree. The branch lengths are related to the amount of evolutionary divergence [122]. A rooted phylogeny will also have a "root" which is the ancestor of all sequences considered in the tree, and the location of the root can be determined in various ways [123]. The path from the root node to any other node on the tree is unique and represents an evolutionary trajectory. Several methods for reconstructing phylogenetic trees molecular evolution including clustering by distance, parsimony, maximum likelohood, or Bayesian methods [122, 124].

This chapter describes all steps in the construction of the phylogeny-based network, whose performance is assessed in Chapter 6. The phylogeny is based on a codon alignment of all TEM mutant sequences for which both the coding DNA sequence and

corresponding protein sequences could be gathered (Section 5.2). Bayesian phylogenetic inference (Section 5.3) was then used to reconstruct the TEM β–lactamase evolutionary history.

5.2 Sequence database and alignment

I compiled 227 TEM and the closely related SHV sequences (and one PSE sequence to be used as an outgroup) from existing databases of naturally occurring β–lactamases [77]. SHV and PSE sequences are both in the same protein super-family as TEM, and are the most closely related sequences.

I started with 220 sequences from the Lahey clinic Class A β–lactamase database [77], containing resistant β–lactamases from pathogenic strains isolated in the clinic. I downloaded TEM/SHV nucleotide coding sequences from NCBI GenBank [125] using identifiers provided in [77].

Additionally, I queried human microbiome databases, containing β–lactamases that do not necessarily come from pathogenic strains. Because of the lack of defined selection, I expected these sequences to have intermediate levels of antibiotic resistance. I searched the Joint Genome Institute Integrated Microbial Genomes (JGI IMG) database [126,127] for all human microbiome genomes with available peptide sequences and obtained over 7 million peptide sequences. I also searched the Human Microbiome Project Reference Genome Database (HMRGD) [128] for all metagenomic samples

from all human tissues and obtained over 5 million peptide sequences. I queried a set of representative sequences from each major phylogenetic clade of the class A β–lactamases [129] to which TEM and SHV belong against these two peptide sequence databases.

BLAST+ [130] was used for the database search: Every query in the Class A β–lactamase group phylogeny [83, 129] was uniquely paired with the BLAST hits with which this query had the lowest E-value [130]. In this way, I ensured that orthologs, and no close paralogs were included. Specifically, if a given sequence was a hit both for TEM and SHV, but the E-value with the SHV sequence was much lower than the one for TEM, the sequence was considered a closer ortholog for SHV, rather than TEM and was not added to the list of TEM hits.

Finally, from the JGI-IMG + HMPRG database hits, I removed the sequences already present in the database of class A β–lactamase sequences reported in clinical samples [77]. The majority of β–lactamase sequences from these human microbiome databases overlapped with the known clinically relevant ones, possibly because there was no strong antibiotic resistance selection on the enzyme in the harmless strains in the microbiome. Furthermore, when β-lactamases are present in the non-pathogenic strains of the human microbiome, it can be assumed that they were acquired through lateral transfer [131]. As a result, the microbiome databases contributed with only two TEM and five SHV additional sequences.

The nucleotide sequences from the Lahey database were retrieved from Genbank

[125] by their identification numbers and were translated to protein sequences using the bacterial codon usage tables [132]. For the seven additional and non-redundant sequences from the IMG and HMP databases, I found the nucleotide and corresponding peptide sequences. Protein sequences were aligned by following the Ambler numbering scheme, which assigns amino acid positions based on optimized alignments of the class A β–lactamase superfamily [78]. The alignment of nucleotide sequences corresponding to the peptide alignment, was obtained by inserting a triple gap in the codon position corresponding to a given amino acid position.

5.3 Phylogeny reconstruction

Building a phylogeny using Bayesian MCMC

Phylogenetic inference uses similarities and differences among biological entities (species, genes, genomes) to reconstruct their evolutionary history. This inferred history is summarized in the form of a phylogenetic tree, typically a binary tree, for which nodes represent genetic sequences and links correspond to differentiation events. The underlying assumption behind the phylogenetic tree model is that the extant species have descended from a common ancestor. Methods for constructing phylogenetic trees can be grouped into distance-based, parsimony, and maximum likelohood classes.

Here, I use a Bayesian phylogeny inference method belonging to the maximum

likelohood phylogeny inference class. Classical maximum likelohood methods use a given model of evolution and search for the tree that maximizes the probability of observing the data (sequence alignment) given that tree. Bayesian methods on the other hand search for the tree that maximizes the (posterior) probability of this tree given the data and model of evolution. In MrBayes [123], the posterior probability of the ith phylogenetic tree (τ_i) conditional on the sequence alignment (X) can be calculated by Bayes theorem:

$$f(\tau_i|X) = \frac{f(X|\tau_i)f(\tau_i)}{\sum_{j=1}^{B(s)} f(X|\tau_j)f(\tau_j)} \tag{5.1}$$

where the summation is over all possible trees for s species, and the tree prior is uniform $f(\tau_i) = B(s)^{-1}$. The tree likelihood function, $f(X|\tau_i)$, is a multiple integral of tree parameters like branch lengths and molecular substitution rates. The summation and integrals cannot be calculated analytically and in MrBayes, so they are approximated using Metropolis-Coupled Markov Chain Monte Carlo $(MC)^3$ [133]. Beyond an initial burn-in time, the proportion of time a tree topology is visited during the Markov Chain is a valid approximation of its posterior probability [123]. For improved sampling of the space of possible trees, incrementally heated Markov chains can be run in parallel with a cold chain and are coupled via the Metropolis criterion. In MrBayes, the user can specify the number of chains and the frequency with which pairwise swaps of the current states are attempted between chains in order to optimize sampling. A swap of two tree states is accepted based on the Metropolis criterion,

and it allows escaping local minima in the posterior [134].

Choice of MrBayes model and constraints

Like other Markov Chain Monte Carlo methods, successful use of MrBayes requires careful choice of evolutionary model and constraints [135]. In order to allow increased variation in the evolutionary parameters in different parts of the input alignment (either sets of sequences or sets of positions along the sequences), MrBayes allows the input alignment to be partitioned. I determined the optimal partitioning scheme to be used as an input to the multiple sequence alignment and evolutionary model through PartitionFinder [135]. PartitionFinder is an algorithm that does automated model selection by sampling from different partition schemes and evolutionary models. For the TEM sequence alignment, the best partition of the nucleotide alignment was partitioning by codon position (1st, 2nd, or 3rd). The best evolutionary model was a subset of the generalized time reversible model of nucleotide substitution with gamma-distributed rates and a proportion of invariant sites (GTR+G+I). The number of gamma categories was increased from 4 to 6, in order to improve the smoothness of the gamma approximation and allow for the large variation of substitution rates across sites.

MrBayes runs

I ran the parallel (MPI) implementation of MrBayes, with 6 independent Metropolis-coupled MCMC runs. Each run had 8 Metropolis-Coupled chains (1 cold and 7 heated), which was to ensure that the cold chain does not become stuck in local minima of the posterior [136]. I ran for 30 million generations and discarded generations up to the 10 millionth one, to ensure that I was well beyond the burn–in period and into the equilibrated phase. Therefore, each of the 6 independent runs was thinned to every 25,000th generation, to remove most autocorrelations between phylogeny parameters.

MrBayes convergence monitoring

In theory, a Markov chain will eventually converge to a unique stationary distribution, in the limit of infinite number of steps [137]. However, there is no way to guarantee convergence has occurred in a *finite* number of steps. There are several ways to visually or statistically assess whether the Markov chain appears to have converged. The built-in convergence diagnostic for MrBayes is the split (tree branch bipartition) frequency standard deviation between the trees in different chains. The standard deviation of the number of bipartitions in the trees is expected to decrease throughout a run, and < 0.01 is a commonly accepted cutoff for stopping the runs. TRACER [138] and AWTY [139] are additional diagnostic tools for phylogeny MCMC sampling: Figure 5.1 shows the overall tree likelihood function from on of the 6 independent runs in MrBayes, and along all 30 million generations. The overall tree likelihood

includes both the tree topology and all the evolutionary parameters being optimized, including substitution rates and stationary nucleotide frequencies. I made sure that the distribution of the tree likelihood values is not multimodal, but the values fluctuate around a single average, corresponding to a single peak in the likelihood function.

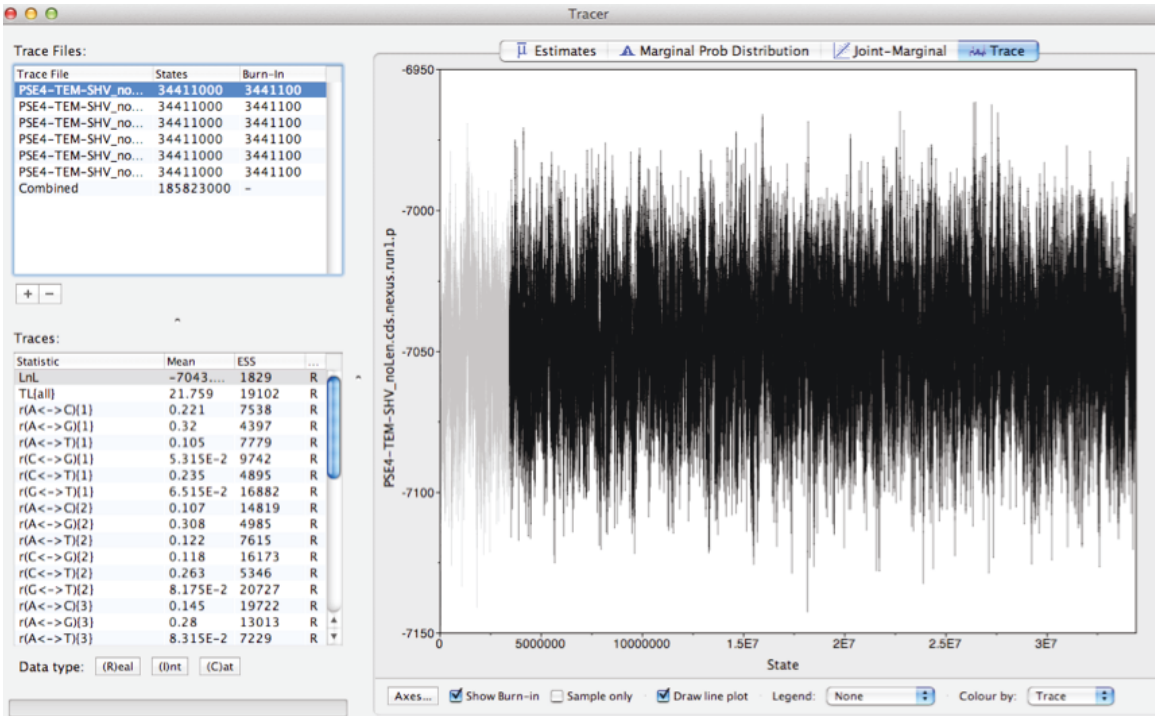

Figure 5.1: TRACER [138] output for tree likelihood function over 30 million generations of Mr Bayes' MCMC simulations.

MrBayes consensus tree

Figure 5.2 shows the consensus tree from the MCMC runs, based on mutually compatible clades that exist in at least 50 % of the trees. A set of clades (phylogenetic subtrees) sampled in the MrBayes run is considered mutually compatible if each leaf node (evolved TEM, SHV, or PSE sequence from the alignment) was assigned to

exactly one tree clade. The two major clades are for TEM and the closely related SHV β–lactamases. The outgroup is PSE-4, which is a third subgroup of the Class A (active site serine) β–lactamases. PSE-4 is closest in sequence to the TEM and SHV subgroups.

On this consensus tree, the TEM and SHV clades are each very shallow or comb–like, showing polytomies. This is because the consensus tree averages individual trees from the MCMC run, and therefore may not represent a realistic evolutionary tree [140]. Additionally, in the case of TEM β–lactamase, there were diverse topologies in the MrBayes run that were considered equally optimal in modeling the given sequence alignment, due to the low sequence divergence in the TEM family. Note that the TEM clade is deeper in individual trees from the MrBayes tree ensemble, than in the consensus (summary) tree. Therefore, rather than using the consensus tree for network construction the optimized trees from the post-burn-in tree ensemble were used in the analysis below. The statistics performed on these optimized trees from the MrBayes ensemble required knowing an optimized tree topology rather than a summary or consensus tree.

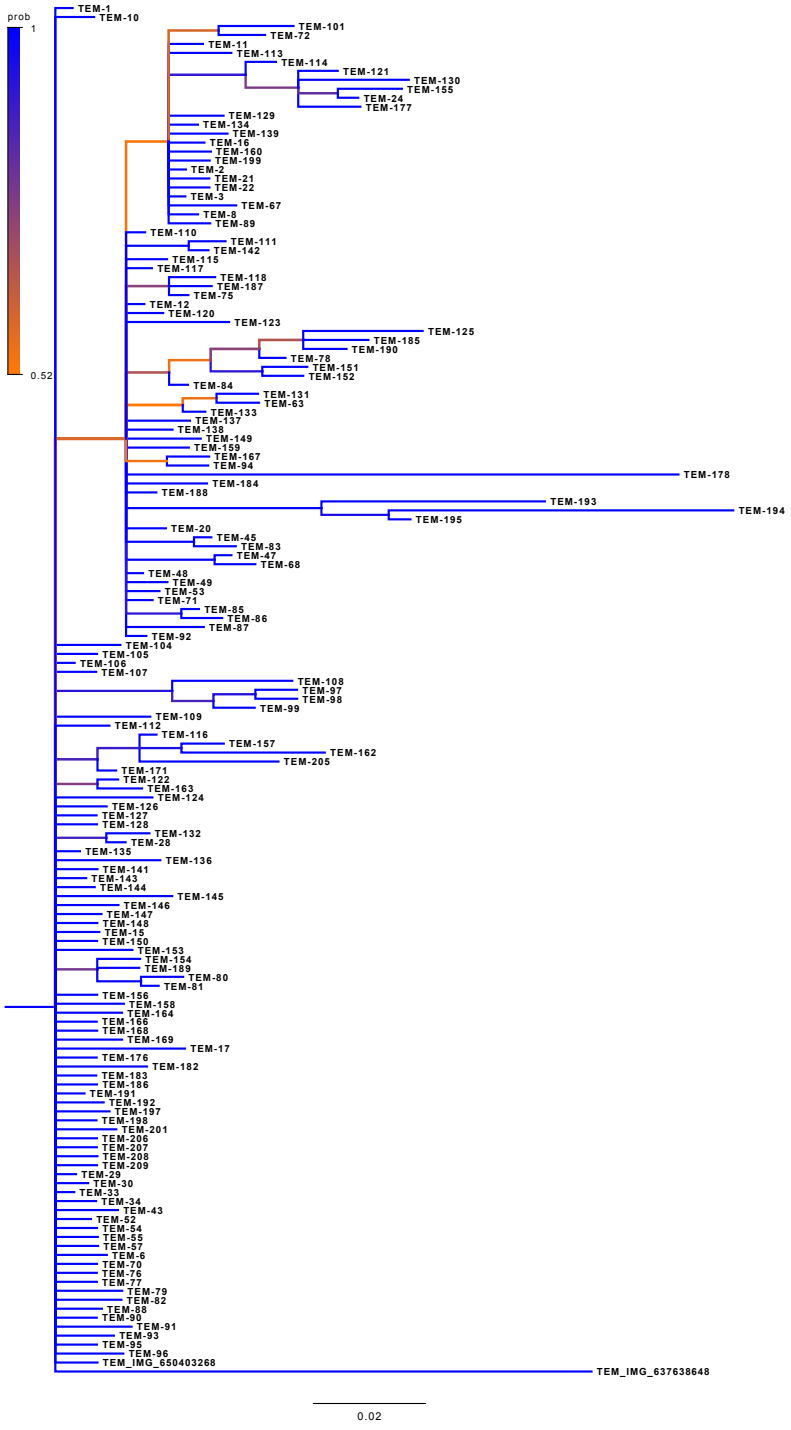

Figure 5.2: Consensus (summary) TEM subtree from the TEM, SHV, and PSE phylogeny. The consensus tree from the MrBayes run contains clades with at least 50 % support, i.e. which were present in at least 50% of the trees in the post burn in MrBayes ensemble.

5.4 Reconstructing ancestral states on the phylogeny

Ancestral reconstruction, a.k.a. character mapping, is an evolutionary biology method of identifying the phenotypic and genetic states of extinct ancestral organisms or genes [141, 142]. This occurs via extrapolation back in time to the common ancestor of pairs of sequences. When using maximum likelohood methods, the tree topology, branch lengths, and a substitution model can be used to infer the likelihood of a given phenotypic trait or the molecular sequence of an ancestral node.

Ancestral *sequence* state reconstruction

For all phylogeny-based network reconstructions, I needed to find pairs of mutations that arise in the same phylogenetic clade. Additionally information on the preferential order of mutations in time (earlier vs. later mutation in the same clade) was needed. Therefore, I reconstructed the sequences internal("extinct") nodes on the TEM β–lactamase phylogeny. The PyCogent Python package [143] was used: a maximum likelohood method was applied to find the likely nucleotide character distribution at every internal node and for each position in the gene. The same evolutionary model as in MrBayes was used: a generalized time-reversible substitution model with gamma-distributed rates. The resulting nucleotide internal node sequences were translated to amino acid sequences in the BioPython package [144]. I repeated this

ancestral reconstruction for every tree in the equilibrated phase ensemble from the MrBayes runs.

Ancestral *phenotypic trait* reconstruction

In addition to reconstructing the ancestral sequences at internal nodes, I needed to estimate how the resistance phenotype changes along branches of the tree. Since I was interested in mutation pairs contributing to adaptation, i.e. the acquisition of a given resistance phenotype, I only focused on pairs of mutations that acquired the new phenotype or maintained it. Using the known resistance phenotypes of TEM sequences in the alignment: broad-spectrum resistant, extended-spectrum resistant, inhibitor-resistant, and the combined extended-spectrum and inhibitor-resistant, I reconstructed the internal node phenotype at each node in each tree in the ensemble. The ACE function in the APE R package [145] was used for this reconstruction. ACE can estimate the phenotype of both continuous and discrete phenotype properties, or traits in the evolution of a gene/organism. For example, for TEM β–lactamase a continuous trait would be the amount of antibiotic resistance (MIC), and discrete phenotypic trait would be the main type of antibiotic resistance. Since the information on the amount of antibiotic resistance comes from different experimental assessments, it was not a consistent enough phenotypic trait to model. However, the type of resistance trait is available for most naturally occurring TEM mutant sequences. For discrete characters, ACE uses a maximum likelohood model [146]. There are three types of

transition rate matrices between the discrete states could be specified: an all-rates equal, a symmetric, and an all rates different transition matrix. I chose the all-rates different model, there are different levels of selection for extended-spectrum resistance vs. inhibitor resistance, for example. Additionally, the rates are not symmetric, since equal rates of transitioning to and away from a given phenotype are unlikely given a strong selection for that phenotype. As a result, ACE gives the likelihoods of each phenotypic trait on each internal node, and I picked the resistance phenotype with the maximum likelihood.

5.5 Building a phylogeny-based network of co-evolving positions

The phylogeny-based network of coevolving positions was constructed based on sampling each tree topology in the MrBayes equilibrated phase tree ensemble. A statistical test was performed for each pair of mutations in each of the trees in order to determine if the pair appeared functionally associated in the tree. The number of trees in which a pair of mutations passed the test of functional association was ultimately used to weight the mutation pair in the network.

Below, I describe the techniques used to walk through the each tree and find pairs of mutations with potential functional associations, the statistical test performed to determine association on the tree, and the method to aggregate the test results from

each tree in the ensemble.

Tree walking algorithm

Walks were performed along the phylogenetic tree, starting from the most recent common ancestral (MRCA) sequence of all TEM β-lactamases and ending at each leaf node in the tree. Only pairs of mutations occurring along a path from MRCA to leaf sequences that satisfied the constraints below were considered.

- The two mutations are no more than a *distance threshold* (chosen based on the tree depth) apart, assuming they both lie on a path from root to leaf nodes. This is equivalent to there being few (silent and non-silent) mutations occurring between them, along that common path. If there are many mutations between the two mutations, it becomes less clear whether the acquisition of the new function should be attributed to those mutations. Therefore, The mutations that are very far apart are not likely to have a strong functional association.

- Both mutations lead to a mutant that has the new resistance phenotype of interest. In other words, if the first mutation is associated with acquisition of extended-spectrum resistance but the second mutation is associated with inhibitor resistance, these two mutations arose under different selective pressures and therefore could not be associated when selecting for extended spectrum resistance only. To infer the resistance trait conferred by a mutation, I looked at the phylogeny nodes before and after the mutation (endpoints of the phylogenetic

edge along which the mutation occurred). The resistance phenotype traits of each of these nodes were reconstructed as in Section 5.4 with the ACE program [146]. If the node (TEM sequence) before the mutation was associated with the ancestral TEM-1 broad spectrum resistance trait, but the node after the mutation was associated with extended spectrum resistance trait, then I predicted that this mutation leads to extended spectrum resistance.

Tree distance filters for mutation pairs

Pairs of mutations that were spaced far away from each other in the tree were not included in the Fisher's exact test analysis for a given tree. My assumption was that mutations separated by many intermediate phylogeny branches (and hence other intermediate mutations) are not likely to functionally associated. This is because, if there are many mutations have been selected at different steps between the two mutations, it becomes less clear whether the acquisition of the new function should be attributed to the two mutations. A distance cutoff was therefore applied to any mutation pairs included in the contingency Tables for the Fisher's exact test. The cutoff was chosen based on the minimum tree depths (distances from the most recent common ancestor) observed in the MrBayes equilibrated phase tree ensemble. The distance cutoff of 0.5 was found to be less than the minimum tree depth in the ensemble, yet large enough that it would not significantly constrain the analysis and limit the number of mutation pairs identified.

5.5.1 Individual tree statistics for functionally associated pairs of mutations

I applied a statistic to each tree in the phylogeny ensemble, in order to find pairs of mutations with stronger functional associations than expected at random. For this, I applied Fisher's exact test to contingency tables representing co-occurrence or individual occurrence counts of mutations along the trees. Initially, I describe a method that is only based on functional associations, regardless of the temporal ordering of mutations. Then, I expand this method to consider the two possible orders of occurrence of mutations in a mutation pair. The first method results in an undirected network of functional associations, whereas the second method adds directionality (to be interpreted as temporal ordering) to the links.

Fisher's Exact Test for an undirected pair of mutations

I made the assumption that if two mutations tend to arise along the same path in the phylogenetic tree more frequently than would be expected from their individual frequencies, it is likely that their combination results in increased protein fitness. In other words, I expect that pairs of mutations that co-occur with significant frequencies to have been selected for *together* for increased function. Recalling that the nodes of these trees are populated by protein sequences and the edges by mutations that represent transitions between sequences, I compute, for each tree, the total size of the

clades in the tree (in number of tree edges) containing (1) both mutations; (2) one of the mutations exclusively; (3) the other mutation exclusively; (4) neither mutation. (Table 5.1). The counts are used to construct a contingency Table and Fisher's Exact Test is applied to assess whether the observed counts can be best explained by increased protein fitness, i.e. with a significant right-tail p-value from the test.

phylo edge pairs containing:	mutation 2	not mutation 2	total
mutation 1	n_{12}	$n_{1\bar{2}}$	n_1
not mutation 1	$n_{\bar{1}2}$	$n_{\bar{1}\bar{2}}$	$n_{\bar{1}}$
total	n_2	$n_{\bar{2}}$	N

Table 5.1: Contingency Table for co-occurring pairs of mutations. In a given tree, T, for two mutations, 1 and 2, n_{12} is the number of phylogenetic tree edges in tree clades containing both mutation 1 and mutation 2. $n_{1\bar{2}}$ is the number of phylogenetic tree edges in clades containing mutation 1 but not mutation 2, etc. $n_{\bar{1}}$ is the total number of phylogenetic tree edges in clades containing both mutation 1 and mutation 2.

Fisher's Exact Test for a directed pair of mutations

Unlike in the undirected mutation pair test, here I considered only directed pairs occurring in the same phylogenetic branches, i.e. mutation 1 then mutation 2. I performed Fishers exact test on each ordered pair, such that for two mutations 1 and 2, the test is performed twice: once on the ordered pairs with mutation 1 then mutation 2 (Table 5.2), and once with the reverse order of mutations. Note that the contingency Tables would be different depending on the order of mutations appear.

phylo edge pairs containing:	mutation 2 second	not mutation 2 second	total
mutation 1 first	n_{12}	$n_{1\overline{2}}$	n_1
not mutation 1 first	$n_{\overline{1}2}$	$n_{\overline{12}}$	$n_{\overline{1}}$
total	n_2	$n_{\overline{2}}$	N

Table 5.2: Contingency Table for directed pairs of mutations. Here, I work with directed paths on tree T, starting at the root and ending in a leaf. The co-occurrence analysis is similar to Table 5.1 except each direction, in this case, m1 then m2 is shown.

5.5.2 Network weights aggregated from individual tree statistics

The right-tailed p-value from Fisher's exact test for a given mutation pair was calculated for each tree in the equilibrated ensemble. Pairs passing a significance threshold (p-value < 0.01) for many of the equilibrated tree topologies were expected to have a higher functional association than pairs passing the threshold in fewer topologies. The weight of a link between a pair of mutations in the network was simply the number of trees in the equilibrated ensemble passing the significance threshold for that pair (Eq. 5.2).

$$w(M_i, M_j) = \sum_{\{T\}} I(p_{\text{FET}} \leq 0.01) \tag{5.2}$$

In this case, p_{FET} is the Fisher's Exact Test p-value, I is an indicator function with $I = 1$ when $p_{\text{F}} \leq 0.01$ and $I = 0$ otherwise.

5.6 Phylogeny-based network analysis

5.6.1 Network communities

Network communities for the undirected network and for the undirected rendition of the directed network were identified using the multilevel community algorithm [147] in the python implementation of the iGraph package [148]. The algorithm used here is different from the one in Section 3.4, but the same principle of finding the optimal network partition into densely connected sub-networks (communities) apply here. The multilevel community algorithm [147] goes through multiple rounds of network modularity optimization until convergence to the optimal network modularity. Network communities, when applied to the TEM β–lactamase coevolution networks, help identify positions that tend to be selected together in the evolution of a given function.

Figure B.1 shows the undirected version of the phylogeny-based network, with communities obtained using the multilevel community finding algorithm. As in Section 4.1, I find that the major adaptive mutations in residues G238 and R164 are split into different communities, representing different adaptive strategies and selective pressures in the context of mutations in these two residues [68].

5.6.2 Path betweenness centralities and evolutionary trajectories

In Chapter 4, I expanded the concept of shortest path betweenness of a single node to that of a path. I used central paths in the network to find fitness increasing evolutionary trajectories. However, adaptive trajectories do not necessarily only follow the paths leading to the quickest increase in fitness. Rather, at every step, a mutation is selected based on its effect on the latest sequence of accumulated mutations. I apply random walk centralities to better accommodate these local mutation dependencies, rather than global path optimization. I adapted the random walk algorithm from the single node case to the multiple node path case similarly to Section 4.4.

A path betweenness centrality based on random walks

I started with the k-path centrality algorithm which performs multiple (Markov chain) random walks on a network (of length up to k) [149]. Multiple short walks allow us to approximate an infinite length random walk in the case of an ergodic system. In order to approximate the betweenness centrality that would be obtained from a long walk, a minimum number of iterations (steps in the Markov chain random walk) is required. The number of required iterations is determined by network node size, k, and error tuning parameter [149]. This algorithm was developed for single node centralities, but I was able to expand it to a path of arbitrary length by counting how

many times a path was encountered in the random walks. This is the same definition as for path betweenness based on shortest paths as in Section 4.4, except here I count paths of all lengths. Therefore, the betweenness of a path P_{uv} between two nodes u and v is the number of random paths that pass through P_{uv} but do not start or end at u or v (Equation 5.3).

$$C_B(P_{uv}) = \sum_{\substack{\{s,t,u,v\}\in V \\ \{s,t\}\cap\{u,v\}=\varnothing}} \frac{\sigma_{st}(P_{uv})}{\sigma_{st}} \tag{5.3}$$

5.6.3 Frequency-based and alignment network predictors for comparison to the phylogeny-based predictors

My goal is to compare the performance of the phylogeny-based network to (1) a "naïve" method, in which functional interactions between mutations are not modeled, and (2) the TEM alignment-based methods presented in Chapters 3 and 4.

The "naïve" model assumes that each mutation impacts fitness on its own, and always by the same amount, regardless of other mutations present. This is a mutation frequency based metric which assumes that the more frequently sequences containing a mutation are selected, the greater this mutation's contribution is to TEM resistance. The frequency was defined as the fraction of sequences in the alignment containing that mutation. Furthermore, the independent selection assumption allows me to assign

a fitness prediction for multiple mutations based on a *sequence profile model* [23], in which the individual amino acid frequencies are multiplied (Equation 5.4).

$$F(m_1, m_2, .., m_L) = \prod_{i=1}^{L} f_i(m_i) \tag{5.4}$$

Comparing central paths of different lengths

In order to examine the effect of expanding a given evolutionary trajectory by adding mutations, I needed to compare central paths of the network of different length. In other words, I needed a predictor to compare the fitness (resistance) of the $M_1\_M_2$ double mutant to the original M_1 mutant. For example, I was interested in predicting whether adding mutation M_2 to an existing M_1 would lead to increased fitness. Centralities among networks are however not comparable because when the length of a path is increased,by appending a node, the number of possible paths containing the longer path is further constrained to pass through that node. Therefore, path centralities were normalized (Equation 5.5).

$$C_B^{\text{norm}}(P_{uv}) = \frac{\binom{n}{k(P_{uv})}}{\binom{n-k(P_{uv})}{2}} C_B(P_{uv}) \tag{5.5}$$

The combinatorial term on top accounts for the ways of choosing a path of length $k(P_{uv})$, where here the length is the unweighted length, specifically the number of nodes in the path, from the n nodes in the network. The bottom term is the maximum number of node pairs that remain in the network when the k nodes in the path are

removed. This would correspond to the maximum number of pairs of nodes with a path between them, in a fully connected network.

Comparing frequency-based predictors for different number of mutations

In the independent mutation, frequency-based model, individual mutation frequencies were multiplied to obtain an predictor of the fitness of each set of mutations (evolutionary trajectory). Therefore, here an evolutionary trajectory that gets extended by an additional node, will have a reduced score, because we will be multiplying by a number < 1 (frequency). Because the mutation frequency distribution is not normally distributed, but is skewed toward 0, I chose to use percentile ranks, rather than z-scores for comparing evolutionary trajectories of different lengths. If the percentile rank of the longer trajectory was lower than a constituent shorter trajectory, the fitness was predicted to decrease upon addition of mutations.

Chapter 6

Assessment of phylogeny-based network predictors

6.1 Introduction

In this chapter, I use the results from experimental tests of antibiotic resistance to examine functional interactions within predicted evolutionary trajectories. The analysis differs from Section 4.4, where I confirm that the most central paths tend to be enriched positive functional interactions. Here, I instead compare central paths of different lengths, specifically paths in which the shorter path is contained within the longer path. Comparing the centrality of mutation paths and paths contained in them, allows me to make predictions about whether fitness increases or decreases as mutations are added along evolutionary trajectories. The accuracy of the predicted fitness effect

can then be assessed, by testing the resistance level of the TEM mutants resulting from the predicted mutation trajectories. The experimental assessment described in this chapter is-based on dose response curves for (CTX) resistance (Section 6.2). It involves comparisons of single to double, single to triple, and double to triple TEM mutants. The phylogeny-based network central path predictor is compared to the naïve predictor assuming independent mutation effects and to the alignment network central path predictor (Chapter 5).

6.2 Analysis of dose response curves

Dose response curves were obtained by growing bacterial cells co-expressing the TEM β-lactamase mutant and GFP (in the same plasmid) in different concentrations of CTX [μg/ml]. The assumption is that GFP expression is directly proportional to plasmid-based β–lactamase expression. There were multiple replicates for a given TEM construct at a each CTX concentration, and the GFP fluorescence relative to no drug is plotted. The reasons for using relative fluorescence rather than optical density to measure growth were twofold. First, there was increased accuracy at very low and high bacterial growth, allowing for measurements of resistance over a larger dynamic range. Second, the normalization allows fluorescence due to baseline expression of the β-lactamase to be decoupled from the one due to the TEM mutant efficiency in breaking up the antibiotic. Specifically, bacterial survival under CTX selection is

largely dependent on two properties: the innate ability of the mutated β–lactamase to break down the new drug, and the baseline level at which the β–lactamase is being expressed in the cells. Co-expression with GFP and normalization by fluorescence in no drug (baseline expression) conditions makes it possible to isolate the first property, i.e. the innate enzymatic efficiency of TEM β–lactamase.

While dose response curves are ideally sigmoidal, often the concentration range that was measured, did not exhibit the full transition from complete to inhibited growth in every curve. Therefore, I could not identify a single CTX concentration, such as the minimum concentration at which growth is inhibited, and I could not compare constructs by this concentration. Rather, I chose the area-under-the-curve (AUC) measure, for each dose response curve. This measure incorporates measurements at all concentrations with non-zero relative fluorescence. I used the trapezoidal rule on a non-uniform grid (coinciding with the concentrations measured, on the x-axis) for obtaining the AUC. When there were multiple relative fluorescence measurements (y-values) for a given concentration, I took the arithmetic mean of the relative fluorescence at each concentration point,

$$\overline{\mathrm{AUC}} = \frac{1}{2}\sum_{k=1}^{N-1}\left(c_{k+1}-c_k\right)\left(\overline{f_{k+1}}+\overline{f_k}\right). \tag{6.1}$$

To assess the variability in the AUC used for comparison, I added up the contrbutions of the standard error to the AUC, also using the trapezoidal rule, but this time on the standard error in the fluorescence measurements,

$$\mathrm{SE}_{\overline{\mathrm{AUC}}} = \frac{1}{2}\sum_{k=1}^{N-1}(c_{k+1}-c_k)\left(\mathrm{SE}_{\overline{f_{k+1}}}+\mathrm{SE}_{\overline{f_k}}\right). \tag{6.2}$$

The standard error corresponds to the yellow areas above or below the solid red line (connecting the relative fluorescence means at each concentration) in Fig. 6.1.

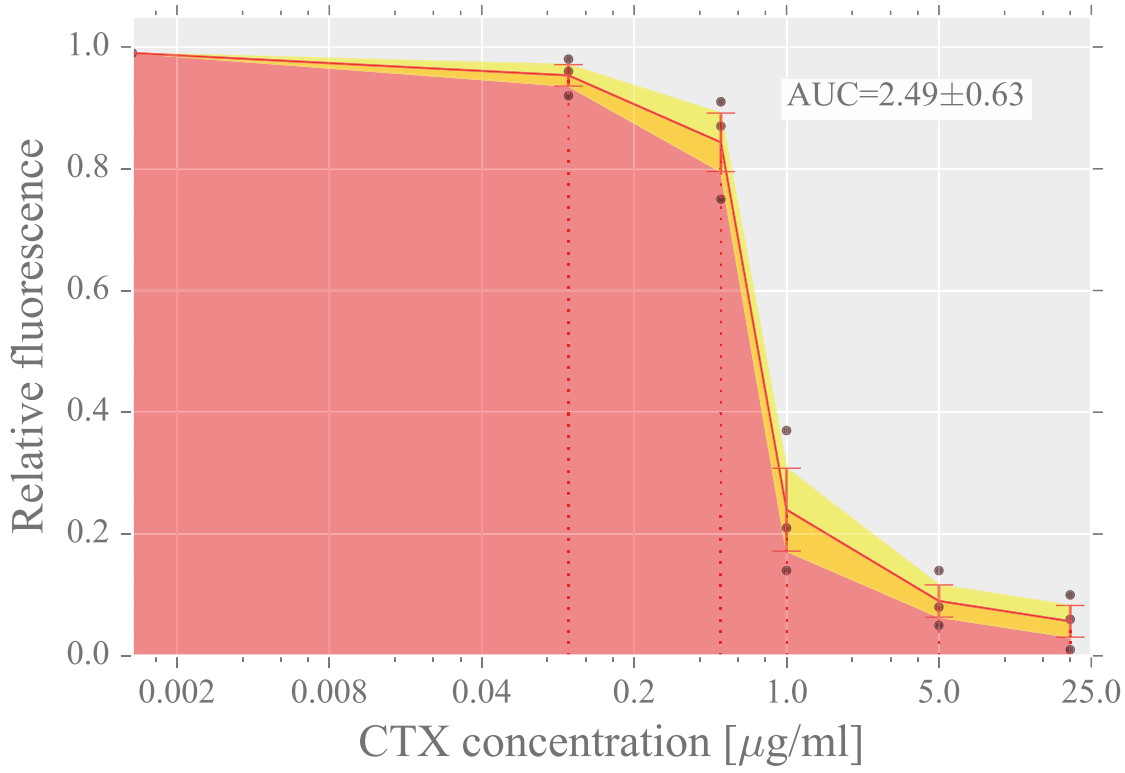

Figure 6.1: Toy example of a dose response curve. The grey circles represent the relative fluorescence measurements at each CTX concentration. The solid red line connects the arithmetic means of these measurements, and bounds the red area, which is the AUC,-based on the relative fluorescence means. The yellow area above (or below) the solid red line represents the standard error of the AUC calculated based on the means.

6.2.1 Pairs of mutants with significant difference in AUC

When a construct A had an AUC greater than a construct B, this difference in AUCs was considered significant when the following condition was satisfied:

$$\mathrm{AUC_A} > \mathrm{AUC_B} \quad \textbf{iff} \quad \overline{\mathrm{AUC_A}} - \overline{\mathrm{AUC_B}} > 3 * (\mathrm{SE}_{\overline{\mathrm{AUC_A}}} + \mathrm{SE}_{\overline{\mathrm{AUC_B}}}) \tag{6.3}$$

For pairs of mutants with significant differences in AUC, the following pairwise comparison function was defined (Equation 6.4):

$$\Theta^{\mathrm{Exp}}(A, B) = \begin{cases} 1, & \text{if } \mathrm{AUC_A} >> \mathrm{AUC_B}; \\ \text{-1}, & \text{if } \mathrm{AUC_A} << \mathrm{AUC_B}; \\ 0, & \text{otherwise.} \end{cases} \tag{6.4}$$

Rather than giving the amount of fitness change, this function qualitatively assesses whether construct Ahas greater or reduced resistance from construct B. All pairs of constructs for which the difference in AUCs was considered significant ($\Theta^{\mathrm{Exp}} \neq 0$) were ranked, and the ranking was used to assess the performance of the predictors derived in 6.

The number of triple mutants tested was 35, resulting in $35 \times 6 = 210$ possible pairwise comparisons between each triple mutant and constitutive double or single

mutants. Of these, the experiments were able to distinguish (beyond the margin of experimental error, $\Theta \neq 0$) 73 between triple vs. constitutive double, and between 38 triple vs. constitutive single mutants. The number of double mutants tested was 68, resulting in $68 \times 2 = 136$ possible comparisons between a double mutant and each constitutive single mutant. Of these 136 pairwise comparisons, 75 were found to be outside the margin of error from the experiments. Therefore, the 346 total possible comparisons between a mutant and its constituents for which the experiments were able to find 186 pairwise comparisons between a mutant containing two or three mutations and its possible constituents. For most of these comparisons (160/186), addition of one or two mutations was shown experimentally to increase the resistance of the resulting complex mutant. For a small number of pairwise comparisons (26/186), adding one or two mutations decreased fitness. This points to a negative functional interaction between the set of mutations being added and the set of mutations currently present in the mutant.

6.2.2 Correspondence between paths in the network and tested TEM mutants

A TEM mutant tested in the lab does not distinguish the order in which the mutations appeared. However, in the networks the three mutations in the mutant M1_M2_M3 could have appeared in six different ways for a directed network, i.e. all

possible orderings of three mutations. For an undirected network, there were three different ways in which this path could appear, depending on which of the three mutations was in the middle. When there were multiple arrangements of a path in a network, the centrality corresponding to the tested mutant was calculated by taking the average centrality over these arrangements. Then, to compare the (averaged) centralities of two different paths, a function similar to Θ^{Exp} was defined for each predictor (Equation 6.5). If the centralities of two paths, A and B differed by any amount, sign function of the difference Θ^{Pred} was nonzero.

$$\Theta^{\text{Pred}}(A,B) = \begin{cases} 1, & \text{if } \text{AUC}_\text{A} > \text{AUC}_\text{B}; \\ \text{-1}, & \text{if } \text{AUC}_\text{A} < \text{AUC}_\text{B}; \\ 0, & \text{if } \text{AUC}_\text{A} = \text{AUC}_\text{B}. \end{cases} \tag{6.5}$$

6.3 Assessment of predictors by experimental pairwise mutant comparisons

Using the AUC-based pairwise rankings, I assessed the accuracy and coverage of each mutation effects predictor. The *accuracy* of a predictor was the fraction of correctly predicted resistance changes in the comparisons between two mutant constructs. The resistance change was correctly predicted if, for a pairwise comparison between two mutants, its normalized network centrality and the measured fitness were

both increasing (or both decreasing). For example, for mutants A and B, this would mean $\Theta^{\text{Exp}}(A, B) = \Theta^{\text{Pred}}(A, B)$.

The *coverage* was defined as the fraction of experimentally obtained pairwise comparisons, which the computational predictors were also able to distinguish. For example, if the experimental comparison of mutants A and B found $\Theta^{\text{Exp}}(A, B) \neq 0$, and $\Theta^{\text{Pred}}(A, B) \neq 0$, for the predictor too.

Assessment of predictors by *all* experimental pairwise mutant rankings.

Table 6.1 shows the accuracy with which each model is able to correctly predict functional increase or decrease for each of the 186 experimental comparisons. The two best predictive methods were the ones based on coevolution, increased the overall accuracy by about 10%, when compared to the method based on independent functional effects, a.k.a. mutation frequency based method.

Both random walk and shortest path betweenness centralities were assessed, yet shortest path betweenness tends to have very low coverage (Table B.2) and were not included in this table. This low primarily results from the centrality of most paths being 0, because most paths do not lie on any of the shortest paths between pairs of nodes. For many pairs of paths, both had centrality 0 and $\Theta^{\text{Pred}}(A, B) = 0$, so predictions could not be made. The best performing centralities in terms of coverage were the random walk centralities, because there were fewer pairs of paths for which each path had a centrality 0.

The undirected phylogeny-based network had (because of the lack of directionality) a similar centrality for many different paths. As a result, differences in centralities were 0 for many paths, so predictions could not be made in these cases. As a result, the *directed* phylogeny-based network performed better (Table B.2).

Predictor type	Accuracy	Coverage
Frequency (independent)	0.64	0.99
Alignment-based network	0.74	0.65
Phylogeny-based (directed) network	0.72	0.76

Table 6.1: Performance of predictors based on all 186 pairwise comparisons in the experiments for which the two constructs in the pair were found to have different resistance ($\Theta \neq 0$). Comparisons are between double or triple mutants and their constitutive mutations or mutation pairs. Because each predictor can rank mutants containing the same number of mutations, the normalization/percentile ranking presented in 5.6.3 was used to compare constructs of different lengths. The frequency predictor assumes mutation effects are independent, and the alignment-based network was constructed similarly to 4 (see 5.6.3). The phylogeny predictor is based on the directed versions of the coevolution network described in 5.5.1.

Assessment of predictors for pairwise comparisons involving *negative* functional interactions.

In the pairwise comparisons resulting from dose response curve AUCs, I found that, for some mutant constructs, resistance decreased when an original single or double mutant acquired further mutations. The occurred even in cases when the added mutations had high frequency among TEM extended spectrum sequences. For example, starting with the TEM mutant E104K_S268G, and adding mutation E240K, led to decrease in resistance from the starting mutant. While E240K is the fourth most

frequent mutation in extended spectrum TEMs, the experiments consistently show that addition of E240K to a mutant containing mutation E104K leads to decreased resistance. This confirms previous findings of E104K and E240K having a negative functional interactions and generally occurring in different sequence contexts [28, 68].

The full table of negative interactions detected in the dose response assays can be found in Table B.3. Known negative pairwise functional interaction pairs [68] were confirmed in the experiments: E104K–E240K, G238S–R164H, and G238S–A237T. Additionally, the negative interactions between A237T–E104K/E240K and N175I–E104K/E240K frequently recur in Table B.3.

The accuracy on the set of 26 comparisons pointing to negative interactions decreases for all methods. The coevolving pairs predictor based on the TEM sequence alignment has the lowest accuracy. From all of the approaches, the frequency has the highest rate of predicting functional effects to be negative: 72 of all 186 pairwise resistance comparisons are predicted to be negative, 55 of 160 true increases in fitness were predicted as negative by this metric.

Predictor type	Accuracy	Coverage
Frequency (independent)	0.58	1.00
Alignment-based network	0.50	0.62
Phylogeny-based network	0.56	0.69

Table 6.2: Performance of predictors based on all 26 pairwise comparisons in the experiments for which the longer mutant was found to have lower resistance than the mutants contained in it (negative interactions). Comparisons are between double or triple mutants and their constitutive mutations or mutation pairs. Because each predictor can rank mutants containing the same number of mutations, the normalization/percentile ranking presented in Section 5.6.3 was used to compare constructs of different lengths. The frequency predictor assumes mutation effects are independent, and the alignment-based network was constructed similarly to Chapter 4 (see Section 5.6.3). The phylogeny predictor is based on the directed versions of the coevolution network described in Section 5.5.1.

The negative interactions that are missed by the mutation coevolution methods (both phylogeny and alignment based) contain mutations in residue A237T and E104K/E240K. While A237T mutations frequently co-occur with E104K/E240K, this typically happens in the context of mutations in residue R164.

As shown in Table 6.3, coevolution-based methods make fewer mistakes in predicting positive functional effects than the independent-effect, frequency approach. In contrast to the coevolution methods, the frequency-based, independent functional effect model, tends to overestimate the number of fitness decreasing combinations of mutations 6.3. The number of positive interactions incorrectly predicted to be negative is 24 and 32 (out of 160) for the alignment and coevolution network respectively, vs. 55 for the frequency based predictor. Frequency tends to mis-predict fitness increasing effects as fitness-decreasing when rarely occurring mutations (in the alignment) are

added to frequently observed mutations, along the evolutionary trajectories.

Predictor type	Negative: Experiment and predictor	Negative: Experiment not predictor	Positive: Experiment and predictor	Positive: Experiment not predictor
Frequency	15	11	103	55
Alignment	8	8	81	24
Phylogeny	10	8	92	32

Table 6.3: Breakdown of predictor performance by positive vs. negative interactions. Comparisons are between double or triple mutants and their constitutive mutations or mutation pairs. Because each predictor can rank mutants containing the same number of mutations, the normalization/percentile ranking presented in 5.6.3 was used to compare constructs of different lengths. The frequency predictor assumes mutation effects are independent, and the alignment-based network was constructed similarly to 4 (see 5.6.3). The phylogeny predictor is based on the directed versions of the coevolution network described in 5.5.1.

The improvement in the coevolution-based phylogeny/alignment predictors is likely due to the coevolution methods better modeling positive interactions between mutations. This is because, by construction, coevolution networks become enriched in positive interactions. In contrast to the coevolution methods, the naïve independent functional effect model, tends to overestimate the number of fitness decreasing combinations of mutations 6.3. However, as the dose-response comparisons show, some of these rare mutations may have strong positive functional interactions with the more frequent extended spectrum mutations. Examples of such rare mutations are I173V and S268G (first and third row 6.4), which were experimentally shown to have positive functional interactions with R164H and A237T respectively.

Construct	Naturally evolved TEM	Mean AUC	Standard error AUC	Improves fitness over n singles	Improves fitness over n doubles	Frequency	Alignment	Phylogeny
R164H_I173V_E240K	TEM_132	57.37	9.26	3	3	0.46	0.7	2.1
E104K_R164H_A224V		64.9	29.7	NA	3	0.73	0.2	1.5
A237T_E240K_S268G		2.24	0.03	3	3	0.46	0.4	1.4
E104K_R164H_I173V		73.2	27.3	NA	3	0.59	0.1	1.2
E104K_R164H_E240K		1.36	0.82	NA	1	0.96	0.4	0.9
R164H_A237T_E240K		123.	75.0	3	3	0.83	0.4	0.8
M182T_A237T_E240K		0.48	0.1	1	1	0.86	0.2	0.8
E104K_A237T_S268G		4.17	2.96	2	3	0.59	0.1	0.8
E104K_M182T_E240K		0.67	0.25	NA	0	0.99	0.6	0.6
E104K_R164H_M182T	TEM_43	186	43.4	3	3	0.96	0.2	0.6
E104K_R164H_A184V		50.62	28.6	3	3	0.80	0.3	0.4
E104K_M182T_A184V		6.06	2.09	NA	2	0.84		0.4

Table 6.4: Top-ranking (by phylogeny-based network centrality) triple mutants that were experimentally tested. The first column shows the mutations in the triple mutant, column 2 indicates the naturally occurring TEM (if any) consisting of these three mutations. The mean AUC and standard error (6.2) accross all experiments is shown in columns 3 and 4, respectively. The frequency predictor, the alignment-based network random walk betweenness centrality 4 (see 5.6.3), and the phylogeny random walk betweenness centrality are shown in columns 5-7.

Enrichment for fitness increasing triplets

Next, I examine the utility of network coevolution model in predicting new combinations of mutations that increase extended spectrum resistance. Table 6.4 shows the top 12 highest random walk betweenness paths in the directed phylogeny-based network, which correspond to triple mutants. This is specifically a subset of the highest centrality triples that were tested in the dose response assays. The mean AUC from all distinct experiments in which a triple mutant was tested is shown. Table 6.4 also contains the number of constitutive single mutants and double mutants over which the triple mutant improves fitness.

Two of the 12 top triple mutants that were assessed in the dose response curves have also been encountered in the natural evolution of resistance (TEM-43, and TEM-132). Both of these mutants are also shown to increase resistance from each of their constitutive double mutants, and have significantly high AUCs. New mutants with similar characteristics are E104K_R164H_I173V, R164H_A237T_E240K, and E104K_R164H_A184V, each of which represents fitness-increasing mutants not previously observed in natural evolution of resistance in the clinic. Two of these three fitness increasing triple mutants contained a relatively rare mutation (I173V in the first mutant above, and A184V in the third), such that the frequency-based metric did not give them a high score.

6.4 Conclusions

Here, I address the question of whether the two coevolution-based methods have utility in predicting the functional impact of multiple mutations on extended spectrum resistance in TEM β-lactamase. I compare these predictors to a naïve model that does not incorporate evolutionary/functional interactions between mutations. In this independent-effect model, each mutation acts on its own, and there are no functional dependencies on the context of other mutations.

The phylogeny-based network model assessed here aims to improve modeling of coevolution by reconstructing and incorporating the evolutionary history (phylogenetic tree) of the resistant TEM sequences. Using phylogenetic trees rather than alignments typically improves detection of coevolution between protein sites [150,151]. This occurs particularly when there are complex evolutionary relationships between sequences, resulting in poor correspondence between the number of times two mutations coevolved and their occurrence in the alignment, termed phylogenetic noise [151]. However, phylogenetic noise tends to be lower when all sequences in the alignment independently evolved from a single ancestor. This is possible in systems with high mutation rates and strong selection, as in the case of the TEM β-lactamase model system.

I find that, for this model system, both of the coevolution-based models (whether or not they incorporate the phylogenetic tree structure) improve prediction of the functional impact of multiple mutations when compared to the independent functional effects method. I reason that this improvement is due to the coevolution methods

better modeling positive interactions between mutations. One possible explanation is that, by modeling coevolution under positive selection, the coevolution networks become enriched in positive interactions. As can be seen in Table 6.4, the high ranking triple mutants by both coevolution metrics tend to improve resistance when the third mutation is added to any possible subsets of double mutants. This is true even when the overall resistance measured in the lab (dose response AUC) is not very high (e.g. the A237T_E240K_S268G mutant). This suggests that these methods may be useful for predicting the direction of functional change (increase/decrease) along an evolutionary trajectory, rather than the absolute fitness.

In contrast to the coevolution methods, the naïve independent functional effect model, tends to overestimate the number of fitness decreasing combinations of mutations Table 6.3. This tends to happen when rare mutations (in the alignment) are added to evolutionary trajectories currently containing frequently observed mutations. When a very rare mutation is added to an evolutionary trajectory, the frequency-based method tends to rank the resulting mutant lower. However, as the experimental measurements show, some of these rare mutations may have strong positive functional interactions with other mutations already present. Examples of such rare mutations are I173V and S268G (first and third row, Table 6.4), which were experimentally shown to have positive functional interactions with R164H and A237T respectively.

Chapter 7

Discussion

7.1 Utility of mutation coevolution networks in studying TEM β-lactamase evolution

Protein evolutionary trajectories can be modeled as gradual walks starting from an initial sequence and fitness and exploring new sequence-fitness combinations through the gradual acquisition of mutations [39, 52, 68]. At every step, mutations leading to fitter neighbor sequences are selected. The end result of these 'greedy' walks are locally or globally optimal protein sequences with higher fitness than their neighbors. The effect of mutations at every step is, however, dependent on the context of currently acquired mutations [152]. Therefore, models that incorporate the functional

relationships between mutations can be useful in better understanding observed evolutionary trajectories and predicting possible future evolutionary trajectories.

Here, I build networks of functional interactions in the evolution of specific function in an enzyme. I apply and expand standard graph-theoretical metrics to model complex interactions between mutations. Using the extensively studied TEM β–lactamase enzyme as a model system, I was able to systematically approach important properties of mutations in this model system, most salient being the following:

- Under multiple selective pressures (for different specialized phenotypic traits), mutations form densely connected clusters or communities in the network corresponding to such selective pressures (Sections 3.4 and 5.6.1). This principle could be expanded to other evolving proteins to identify diverse selective pressures and the types of adaptive mutations that tend to be selected.

- Central mutations in the network that have multiple connections to the mutation clusters associated with different selective pressures reflect a more general functional effect, not specific to the function being acquired. For exaple, mutations in TEM β–lactamase that thermodynamically stabilize the protein could be found in multiple communities and are associated with multiple selective pressures (Section 3.4).

- When a network is built from sequences under selection for a specific function, mutations seem to cluster based on different adaptive strategies that have been,

in independent analyses, shown to be initiated by distinct adaptive mutations. The mutations clustered with these adaptive mutations tend to be compensatory (in the case of pleiotropy), or modulate the effects of the original mutation (Section 3.4).

- In general, pairs of mutations with strong negative functional interactions tend to appear in different mutation clusters. Frequently, negative functional interactions occur between mutations representing alternative and incompatible ways of improving the protein's function (Section 4.2).

- Within a network of mutations under a specific selective pressure, central paths are enriched in mutation combinations (complex mutants) with increased fitness for the function being selected. This property is important for predicting adaptive evolutionary trajectories (Sections 4.4 and 6.3).

- Within a network of mutations under a specific selective pressure, central paths are also enriched in non-negative functional interactions. Triple mutant paths were experimentally shown to increase fitness from all or most of their constitutive single and double mutants (Sections 4.4 and 6.3).

- Mutations exhibiting negative functional interactions with strongly adaptive mutations can be used as starting points for novel evolutionary paths that explore new sequence space (Section 6.3).

7.2 Method applicability to protein evolution in other systems

Based on properties of the bacterial enzyme studied here, the coevolution network models and analysis could be applied to sequence databases for clinically relevant proteins undergoing high mutation rates and under selective pressure, whether from drug treatment or from the host immune response. Examples include surface proteins of pathogens (particularly of RNA viruses such as HIV [153]) or targets for chemotherapy in microbial pathogen or tumor cells. Notably, the methods described here are based on sequence data alone, and detailed tertiary structure information for the target gene is not necessary. For example, it was found that mutations in different proteins in the Zika virus could have increased the virus' geographic expansion over the past 50 years [154]. Most of the proteins in the virus have not yet been crystallized, yet many sequences from multiple continents have been isolated. A sequence-based approach, such as the one described here could help identify selective pressures, adaptive strategies, and (combinations of) adaptive mutations that could have contributed to its increased spread.

7.3 Future development

My most recent coevolution network model (Chapters 5 and 6) uses an inferred phylogeny to represent the correlations that arise from non-independent sequence evolution from a common ancestor. By further introducing a distance threshold, similar to [29], I was able to remove indirect interactions between mutations. A more formal analysis of the phylogenetic tree could be applied that uses Bayesian graphical models (BGMs) to, e.g., tease apart pairwise relationshionships that can be best explained by the presence of a third mutation [155]. BGM methods tend to be computationally expensive and tend to be applied to a consensus tree rather than being able to leverage the tree ensemble. Other Bayesian methods have been applied to a phylogeny ensemble and could lead to improvement in the identification of direct pairwise functional interactions [156].

More generally, both coevolution networks are based on pairwise co-occurrence counts (either in alignments, or in the phylogeny). By focusing on two mutations at a time, the assumption is that the pairwise interactions identified are not affected in the context of other mutations. As a result, I have observed mutation pairs with known negative interactions either directly linked or in the same central path in the network, even though the genetic context of original sequences always included additional mutations. One way to distinguish such *indirect* functional interactions from direct ones would be to combine covariation analysis with global inference analysis and message passing algorithms as in References [21, 23].

Coevolution networks tend to be enriched in positive functional interactions, since they are built on pairs that are frequently co-selected for a given function. However, negative interactions are also important to identify, as mutations exhibiting multiple negative interactions can ultimately lead to exploration of new sequence space. Fitter protein sequences than the ones previously observed can be found as a result [54]. Pairs of mutations with negative interactions typically have lower coevolution weights if they are found at all in the networks (Chapters 4 and 6). However, they are not directly represented in my predictive methods. Reversing the principle behind network weights, weighting pairs that co-occur less (rather than more) frequently than expected by could lead to a network of mostly negative functional interactions.

Finally, my coevolution networks model can predict novel *combinations* of mutations, but requires that these are already present in adaptive sequences. The reason for this is that, since the model is based on observed interactions, mutations that have not been previously observed cannot be included in the model. New mutations observed in laboratory evolution under defined selective pressures, or from mutagenesis experiments measuring the fitness of new mutants can be, however, incorporated into the network. By including sequences (with at least two mutations) obtained in the lab, the number and accuracy of predicted functional effects of evolutionary trajectories can be improved.

Appendix A

Glossary of Terms

adaptive In biology, adaptive traits, are traits with a current functional role in the life of an organism that is maintained and evolved by means of natural selection.. 5, 6, 17, 23, 24, 30, 31, 33, 38, 40–44, 58, 67, 85, 86, 107–109, 111

antagonistic Two or more mutations in a protein are antagonistic when these mutations individually improve a protein's fitness/function, but their combined effect is less than expected by addition of the individual effects.. 33, 53–56, 58–60, 63

autocorrelation In a simulation, autocorrelation is the cross–correlation of a simulated parameter with itself at different points in time, i.e. at different iterations.

Informally, it is the similarity between observations of the same parameter as a function of the time lag between them.. 73

betweenness A network node's betweenness is the number or fraction of (typically, shortest) paths from all vertices to all others that pass through that node.. ix, xi, 31, 38, 40–48, 50, 51, 56, 58, 62, 64, 65, 86, 87, 97, 102, 103

centrality A network node's centrality is a measure of that node's importance in the network. A local measure of centrality is the node's connectivity with its immediate neighbors, i.e. the node's degree. Global network centrality properties include node betweenness, closeness, or eigenvector centrality.. 38, 41

clade In a phylogenetic tree, a clade is a group of organisms that consists of a common ancestor and all its direct "lineal" descendants.. 5, 69, 74, 75

coevolution Within a protein, mutated residues exhibit coevolution, when they reciprocally affects each other's evolutionary characteristics, such as substitution rates. One way to identify coevolving residues is to look for patterns of covariation in the protein sequence.. 3, 5, 43

coevolution network Here: Refers to a network, in which mutated positions/ mutations are the nodes. Links represent covariation between two postions/mutations.. ii, 4–6, 37, 43, 44, 65

community In networks, communities of nodes, a.k.a. network clusters, are groups

of densely connected nodes, i.e. there are more and/or more highly-weighted links within the group of nodes than to nodes within other groups.. 22, 27, 30, 33, 34, 36–38, 41, 47, 48, 55, 58, 62, 85

covariation Covariation in amino acid sequences is a phenomenon whereby some pairs of residues appear to be altered more frequently than expected, typically within multiple sequence alignments.. 5, 110

CTX cefotaxime. 7, 31, 48–52, 55, 58, 59, 61, 64, 65, 91–93

E-value The BLAST E-value score is defined as the number of hits one can "expect" to see by chance when searching a BLAST database of a particular size, and it is a measure of the significance of the match.. 69

epistasis For mutations within the same gene/protein, the dependence in the effect of a mutation on protein function on other mutations present.. 9–11, 24

equilibrated phase For a simulation, such as a Markov Chain Monte Carlo simulation, the equilibrated phase, is the part of the simulation (the set of iterations) for which parameters simulated are fluctuating around a constant average, such that there are no drastic changes in this average.. 73, 78, 79, 81

ergodic A stochastic process, such as a Markov chain random walk is considered ergodic when its statistical properties can be deduced from a single, sufficiently long sample, or a collection of multiple, smaller random samples of the process..

86

extant When referring to a gene sequence, it represents a gene in a currently existing (not extinct) species.. 66, 70

extended-spectrum The extended-spectrum penicillins are a group of antibiotics that have the widest antibacterial spectrum of all penicillins. Extended-spectrum antibiotics affect additional types of bacteria beyond their precursor broad-spectrum antibiotics.. viii, xii, 9, 14, 15, 22, 23, 25, 26, 32–34, 38, 41, 42, 44, 45, 47, 48, 50–52, 54, 55, 58, 61, 62, 64, 65, 78–80

extinct When referring to a gene sequence, it represents a gene in a species without any currently living members. An extinct species may be ancestral to an extant species.. 77

maximum likelohood In statistics, maximum-likelihood (ML) methods estimate the parameters of a statistical model given the observed data. ML methods assume that a good estimate of the unknown parameters, would be the value of the parameters that maximizes the *likelihood* of the data, i.e. the probability of observing that particular set of data, given the chosen probability distribution model.. 67, 70, 71, 77, 78

MIC The minimum inhibitory concentration is the lowest concentration of an antimicrobial that will inhibit the visible growth of a microorganism after overnight incubation.. 10, 78

modularity The network modularity measures the strength of division of a network into modules (also referred to as clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in optimization methods for community structure detection in networks.. 22, 85

most recent common ancestor For an evolved set of entities, like genes or organisms, the most recent common ancestor (MRCA) is the entity from which all other entities in the group are descended.. 81

multimodal For a frequency curve or distribution, having several modes or maxima.. 74

ortholog Gene diverged from another gene by speciation.. 3, 69

paralog Gene diverged from another gene by gene duplication rather than speciation. 69

phenotype In general, an observable trait. For a protein, phenotype refers to a specific function performed by the protein. For example, extended-spectrum β–lactamases have the phenotype of conferring extended-spectrum antibiotic resistance in bacteria containing these enzymes.. 30, 78

pleiotropy The production by a genetic alteration of two or more apparently unrelated phenotypic effects.. 5, 10, 17, 24, 27, 108

polytomies Polytomy is a term for an internal node of a phylogenetic tree cladogram that has more than two immediate descendants (i.e, sister taxa). In contrast, any node that has only two immediate descendants is said to be resolved in the phylogeny.. 75

selective pressure The extent to which organisms possessing a given phenotypic trait are either eliminated or favored by environmental demands. It represents the intensity of natural selection experienced by an evolving population. Antibiotic resistance is an example of a selective pressure. When an antibiotic is used, bacteria that can resist that antibiotic have a greater chance of survival than those that are "susceptible." Susceptible bacteria are killed or inhibited by an antibiotic, resulting in a selective pressure for the survival of resistant strains of bacteria.. viii, 4–6, 9, 12, 13, 15, 22, 26, 29, 30, 36, 80, 85, 107–109, 111

synergistic Two or more mutations in a protein are synergistic when their combined effect on protein function is greater in magnitude than expected by addition of their individual effects.. 53–56, 58–60

trait A (phenotypic) trait is a specific characteristic of an organism's phenotype. A phenotype is comprised of multiple observable / measurable traits.. 77–81

Appendix B

Supplementary Figures and Tables

1	2	3	4	5	6	7	8	9	10
Mutant	n	Opt. conc	Min. growth	Max growth	Avge growth	Stand. dev.	Stand. err.	Stand. err.%	Norm. avge
Q39R	4	0.04	1.80	2.60	2.09	0.36	0.3	16.7	2.1
L40W	4	0.04	1.60	2.40	2.08	0.36	0.4	17.0	2.1
L51P	4	0.04	1.70	2.10	1.93	0.21	0.2	10.5	1.9
E104K	3	0.04	2.00	2.30	2.20	0.17	0.2	8.9	2.2
R120S	4	0.04	1.40	2.35	1.94	0.43	0.4	21.6	1.9
H153R	5	0.04	1.80	2.90	2.17	0.43	0.4	17.3	2.2
R164H	16	0.04	2.30	5.45	3.43	0.90	0.4	12.9	3.4
I173V	4	0.04	1.40	2.45	2.10	0.48	0.5	22.6	2.1
M182T	3	0.04	1.95	2.50	2.15	0.30	0.3	16.0	2.2
K215E	2	0.04	1.75	2.05	1.90	0.21	0.3	15.5	1.9
A224V	4	0.04	1.70	2.15	1.86	0.21	0.2	11.2	1.9
G238S	5	0.12	4.45	7.40	5.88	1.12	1.0	16.7	9.6
E240K	4	0.04	1.00	1.90	1.58	0.40	0.4	25.1	1.6
Q39R R164H	5	0.04	3.30	4.30	3.70	0.47	0.4	11.1	3.7
Q39R G238S	4	0.08	3.75	7.80	5.64	1.73	1.7	30.1	7.8
Q39R E240K	5	0.04	1.80	2.60	2.16	0.39	0.3	15.7	2.2
L40W R164H	4	0.04	1.90	2.80	2.13	0.45	0.4	20.8	2.1
L51P E104K	4	0.04	1.33	1.90	1.65	0.25	0.2	15.0	1.7
L51P R164H	4	0.04	1.70	2.00	1.90	0.14	0.1	7.3	1.9
E104K H153R	4	0.04	2.45	3.05	2.73	0.25	0.2	8.9	2.7
E104K R164H	20	0.12	0.60	8.70	4.69	2.44	1.1	22.8	8.4

Table B.1: Prediction of critical triple mutant evolutionary trajectories ...

1	2	3	4	5	6	7	8	9	10
Mutant	n	Opt. conc	Min. growth	Max growth	Avge growth	Stand. dev.	Stand. err.	Stand. err.%	Norm. avge
E104K I173V	2	0.30	2.70	4.00	3.35	0.92	1.3	38.0	10.8
E104K M182T	7	0.04	1.45	4.40	2.82	1.09	0.8	28.6	2.8
E104K K215E	4	0.04	2.00	2.85	2.39	0.45	0.4	18.5	2.4
E104K A224V	2	0.04	1.80	2.00	1.90	0.14	0.2	10.3	1.9
E104K G238S	6	2.00	2.30	4.10	3.23	0.59	0.5	14.6	16.8
R120S G238S	3	0.12	2.75	4.50	3.48	0.91	1.0	29.5	7.2
H153R G238S	4	0.30	3.15	4.60	4.01	0.68	0.7	16.5	11.5
R164H I173V	4	0.12	2.40	3.65	3.21	0.58	0.6	17.6	6.9
R164H A224V	6	0.04	2.90	4.70	3.90	0.65	0.5	13.3	3.9
R164H E240K	8	0.12	2.60	8.00	5.74	1.78	1.2	21.5	9.5
I173V E240K	2	0.08	1.40	1.60	1.50	0.14	0.2	13.1	3.6
M182T G238S	6	0.30	7.50	9.50	8.68	0.73	0.6	6.7	16.2
K215E G238S	12	0.08	1.65	7.90	4.22	2.28	1.3	30.6	6.3
G238S E240K	6	0.30	2.80	7.60	4.55	1.73	1.4	30.5	12.0
G238S T265M	10	0.30	1.70	5.10	3.36	1.17	0.7	21.5	10.8
Q39R G238S E240K	2	0.30	1.90	2.30	2.10	0.28	0.4	18.7	9.6
Q39R R164H E240K	6	0.12	2.40	7.00	5.37	1.92	1.5	28.6	9.1
L40W E104K R164H	4	0.08	1.30	4.50	2.94	1.33	1.3	44.5	5.1
L51I E104K G238S	4	0.04	1.60	2.00	1.88	0.19	0.2	9.9	1.9
L51P R164H I173V	4	0.04	1.65	2.00	1.79	0.15	0.1	8.2	1.8
L51I E104K M182T	6	0.04	1.43	4.55	2.54	1.31	1.0	41.2	2.5
E104K H153R I173V	3	0.04	2.10	3.20	2.65	0.55	0.6	23.5	2.7
L51I E104K M182T	6	0.04	1.43	4.55	2.54	1.31	1.0	41.2	2.5
E104K H153R I173V	3	0.04	2.10	3.20	2.65	0.55	0.6	23.5	2.7
E104K H153R R164H	3	0.12	6.90	8.30	7.77	0.76	0.9	11.0	11.5
E104K H153R G238S	5	2.00	2.95	6.20	4.06	1.31	1.1	28.2	17.7

Table B.1: Prediction of critical triple mutant evolutionary trajectories ...

1	2	3	4	5	6	7	8	9	10
Mutant	n	Opt. conc	Min. growth	Max growth	Avge growth	Stand. dev.	Stand. err.	Stand. err.%	Norm. avge
E104K R164H A224V	5	0.12	2.70	8.00	5.58	2.42	2.1	38.1	9.3
E104K R164H I173V	4	0.30	8.80	9.20	9.00	0.20	0.2	2.2	16.5
E104K R164H M182T	6	0.60	5.20	7.90	6.94	0.99	0.8	11.4	16.9
E104K G238S T265M	3	0.60	9.40	9.55	9.48	0.08	0.1	0.9	19.4
E104K K215E G238S	4	0.30	3.10	4.80	3.78	0.73	0.7	18.9	11.3
R120S H153R G238S	3	0.30	5.30	7.70	6.87	1.36	1.5	22.4	14.4
R120S G238S E240K	2	0.60	2.90	3.10	3.00	0.14	0.2	6.5	12.9
H153R M182T G238S	6	0.60	6.70	8.90	8.03	0.80	0.6	8.0	17.9
H153R R164H G238S	3	0.12	1.60	3.40	2.27	0.99	1.1	49.3	6.0
I173V R164H E240K	9	0.60	4.50	9.35	7.56	1.59	1.0	13.8	17.5
WT	30	0.04	0.04	2.80	1.59	0.79	0.3	17.7	1.6
Δ	15	0.04	1.10	2.90	1.76	0.48	0.2	13.7	1.8
E104K R164S G267R	11	0.30	4.05	9.40	7.32	1.76	1.0	14.2	14.8
	9	0.60	1.60	4.35	2.66	0.80	0.5	19.7	12.6

Table B.1: Cefotaxime gradient measurements. All mutants and controls tested experimentally for cefotaxime resistance are listed in column 1. The total number of growth measurements (n) for each clone is provided in column 2. The concentration empirically found to produce adequate resolution (i.e. intermediate level of growth in the gradient) is listed in column 3. Individual measurements (in centimeters) of continuous growth at the optimized concentration were conducted and the limit of continuous growth at the optimized cefotaxime concentration was obtained (in centimeters). The minimum, maximum, and average of all the measurements for a given clone are shown in columns 4–6, with the corresponding standard deviation, standard error and % standard error in columns 7–9 respectively.

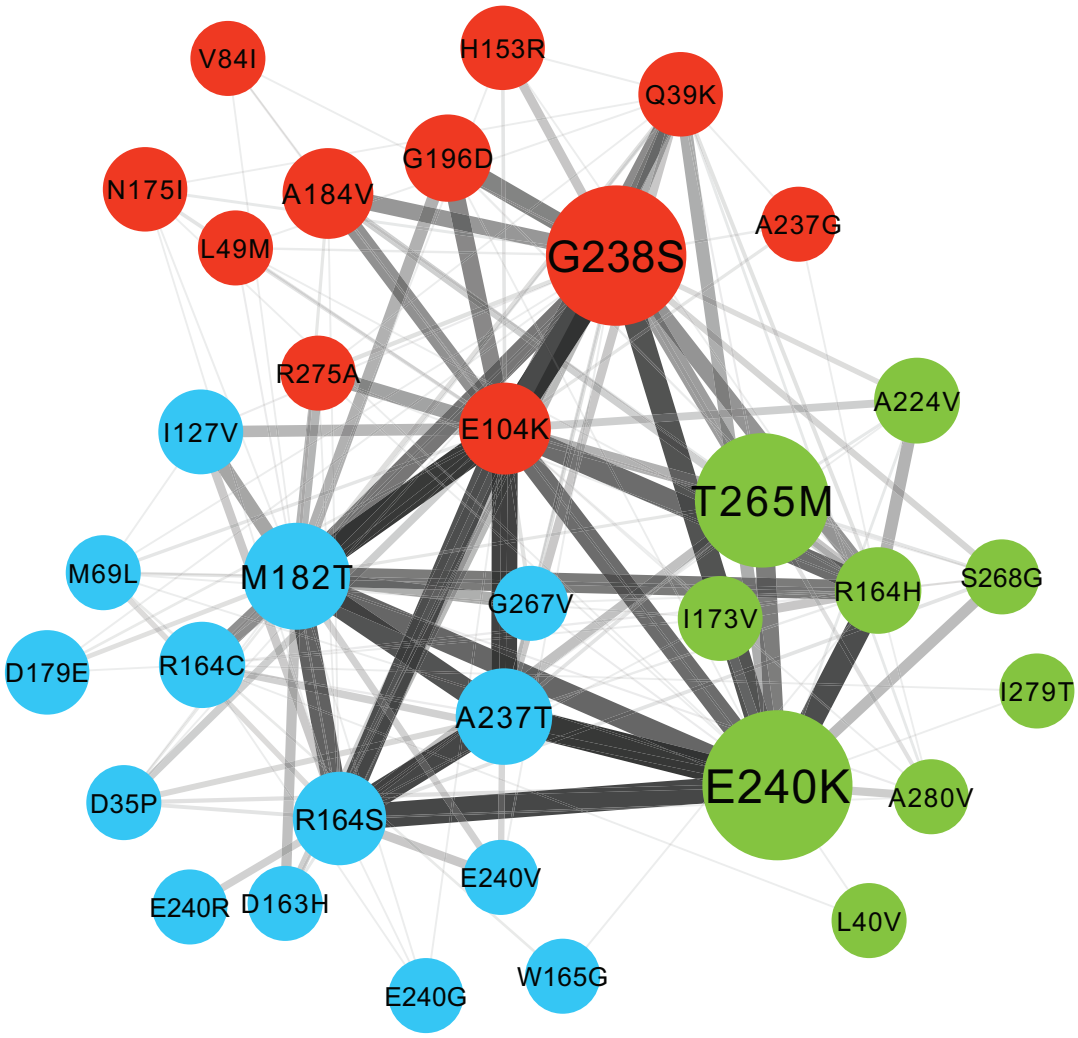

Figure B.1: The TEM coevolution network and its communities: The network was constructed based on frequencies of co-occurring mutated residue positions in the trees from the MrBayes phylogeny ensemble representing TEM β–lactamase evolution (Section 5.5). Node size is proportional to the k-path, random walk betweenness centrality (Section 5.6.2). Link thickness is proportional to the functional association weights from the phylogeny ensemble (Section 5.5). Node (residue) numbers are shown in Ambler notation. The multilevel community-finding algorithm [147] identified three major communities. As in the alignment-based network the frequent mutations in residues G238 and R164 are located in different communities. (red vs. green or blue).

Predictor	Accuracy	Coverage	Accuracy (negative)	Coverage (negative)
Frequency	0.64	0.99	0.58	1.00
Alignment network RW	0.74	0.65	0.5	0.62
Alignment network SP	0.77	0.17	0.5	0.08
Phylogeny network (undirected) RW	0.62	0.33	0.33	0.23
Phylogeny network (undirected) SP	0.65	0.11	1.00	0.04
Phylogeny network (directed) RW	0.72	0.76	0.56	0.69
Phylogeny network (directed) SP	0.59	0.12	0.5	0.15

Table B.2: Experimental assessment of frequency, and coevolution network predictors. Experimental assessment of independent (frequency model) predictor, and the alignment and phylogeny-based coevolution networks. Both the undirected and the directed version of the phylogeny network is shown. All network predictors are based on the paths' betweenness centralities. RW: random walk (adapted from the k-path algorithm, [149]) betweenness centrality, SP: shortest path betweenness centrality.

Higher resistance TEM mutant	Lower resistance TEM mutant	Frequency (Indep.)	Alignment Undir. Net	Phylogeny Dir. Net
A224V	**I173V** A224V	−		
A237G	A237G **S268G**	−		
A237T	A237T **E240K**	+	+	+
E104K	E104K **N175I**	−	−	−
E104K	E104K **N175I A184V**	−		
E104K	E104K **N175I E240K**	−	−	−
E104K A184V	E104K **N175I** A184V	−		
E104K M182T	E104K M182T **E240K**	+	−	−
E104K R164H	E104K R164H **E240K**	+	+	+
E104K S268G	E104K **E240K** S268G	+		
E240K	**A237T** E240K	−	+	+
E240K S268G	**E104K** E240K S268G	+	−	−
G238S	**A237G** G238S	−	−	−
G238S	**A237T** G238S	−		−
G238S	G238S **S268G**	−	−	−
G238S	**R164H** G238S	−	−	−
I173V A237T	**E104K** I173V A237T	+		
I173V E240K	**E104K** I173V E240K	+	−	−
M182T A237T	M182T **A184V** A237T	−		−
N175I	**E104K** N175I	+	+	+
N175I	N175I **E240K**	+		
Q39K E104K	Q39K E104K **A237T**	−	+	+
Q39K E240K	Q39K **A237T** E240K	−	+	+
R164H E240K	**E104K** R164H E240K	+	+	+
S268G	**A237G** S268G	−		
S268G	**G238S** S268G	+	+	+

Table B.3: Negative functional interactions identified in dose response curves. Column 1 shows the starting single/double mutant which was found to have a higher resistance compared to the mutant shown in column 2. The mutant in column 2 has 1 or 2 additional mutations from the one in column 1. These additional mutations are highlighted in boldface. The predicted change in resistance (+ and − for positive and negative) are shown in columns 3 (independent mutation, frequency-based model), 4 (alignment-based coevolution network), and 5 (directed phylogeny-based coevolution network).

Bibliography

[1] C. B. Anfinsen, "Studies on the principles that govern the folding of protein chains," *Nobel Lectures in Molecular Biology: 1933-1975*, p. 401, 1977.

[2] D. M. Taverna and R. A. Goldstein, "Why are proteins so robust to site mutations?" *J Mol Biol*, vol. 315, no. 3, pp. 479–484, 2002.

[3] M. Soskine and D. S. Tawfik, "Mutational effects and the evolution of new protein functions." *Nat Rev Genet*, vol. 11, no. 8, pp. 572–582, 2010.

[4] P. D. Stenson, M. Mort, E. V. Ball, K. Shaw, A. D. Phillips, and D. N. Cooper, "The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine," *Human genetics*, vol. 133, no. 1, pp. 1–9, 2014.

[5] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks." *Proc Natl Acad Sci U S A*, vol. 89, no. 22, pp. 10 915–10 919, 1992.

BIBLIOGRAPHY

[6] P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function," *Annu Rev Genomics Hum Genet*, vol. 7, pp. 61–80, 2006.

[7] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families," *Journal of molecular biology*, vol. 257, no. 2, pp. 342–358, 1996.

[8] A. I. Shulman, C. Larson, D. J. Mangelsdorf, and R. Ranganathan, "Structural determinants of allosteric ligand activation in RXr heterodimers." *Cell*, vol. 116, no. 3, pp. 417–429, 2004.

[9] B.-C. Lee, K. Park, and D. Kim, "Analysis of the residue-residue coevolution network and the functionally important residues in proteins." *Proteins*, vol. 72, no. 3, pp. 863–872, 2008.

[10] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe III, A. S. Kondrashov, and P. Bork, "Prediction of deleterious human alleles," *Human molecular genetics*, vol. 10, no. 6, pp. 591–597, 2001.

[11] C. Berezin, F. Glaser, J. Rosenberg, I. Paz, T. Pupko, P. Fariselli, R. Casadio, and N. Ben-Tal, "Conseq: the identification of functionally and structurally important residues in protein sequences," *Bioinformatics*, vol. 20, no. 8, pp. 1322–1324, 2004.

[12] M. E. Glasner, J. A. Gerlt, and P. C. Babbitt, "Evolution of enzyme super-

families," *Current opinion in chemical biology*, vol. 10, no. 5, pp. 492–497, 2006.

[13] J. C. Fay and C.-I. Wu, "Sequence divergence, functional constraint, and selection in protein evolution," *Annual review of genomics and human genetics*, vol. 4, no. 1, pp. 213–235, 2003.

[14] D. La, B. Sutch, and D. R. Livesay, "Predicting protein functional sites with phylogenetic motifs," *Proteins: Structure, Function, and Bioinformatics*, vol. 58, no. 2, pp. 309–320, 2005.

[15] A. S. Kondrashov, S. Sunyaev, and F. A. Kondrashov, "Dobzhansky–muller incompatibilities in protein evolution," *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 14 878–14 883, 2002.

[16] D. M. Jordan, S. G. Frangakis, C. Golzio, C. A. Cassa, J. Kurtzberg, T. F. f. N. G. , E. E. Davis, S. R. Sunyaev, and N. Katsanis, "Identification of cis-suppression of human disease mutations by comparative genomics." *Nature*, vol. 524, no. 7564, pp. 225–229, Aug 2015.

[17] J. Liao, M. K. Warmuth, S. Govindarajan, J. E. Ness, R. P. Wang, C. Gustafsson, and J. Minshull, "Engineering proteinase k using machine learning and synthetic genes," p. 16, Jan. 2007.

[18] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov,

"Epistasis as the primary factor in molecular evolution," *Nature*, vol. 490, no. 7421, pp. 535–538, 2012.

[19] D. M. Weinreich, "High-throughput identification of genetic interactions in hiv-1," *Nat Genet*, vol. 43, no. 5, pp. 398–400, 2011.

[20] R. E. Bellman and S. E. Dreyfus, *Applied dynamic programming*. Princeton university press, 2015.

[21] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 1, pp. 67–72, 2009.

[22] S. Chakrabarti and A. R. Panchenko, "Structural and functional roles of coevolved sites in proteins." *PLoS One*, vol. 5, no. 1, p. e8591, 2010.

[23] M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, and M. Weigt, "Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1," *Molecular biology and evolution*, vol. 33, no. 1, pp. 268–280, 2016.

[24] J. L. Martínez, F. Baquero, and D. I. Andersson, "Predicting antibiotic resistance," *Nature Reviews Microbiology*, vol. 5, no. 12, pp. 958–965, 2007.

[25] Â. Novais, I. Comas, F. Baquero, R. Cantón, T. M. Coque, A. Moya, F. González-Candelas, and J.-C. Galán, "Evolutionary trajectories of beta-lactamase ctx-m-1

cluster enzymes: predicting antibiotic resistance," *PLoS Pathog*, vol. 6, no. 1, p. e1000735, 2010.

[26] G. Jansen, C. Barbosa, and H. Schulenburg, "Experimental evolution as an efficient tool to dissect adaptive paths to antibiotic resistance," *Drug Resistance Updates*, vol. 16, no. 6, pp. 96–107, 2013.

[27] A. Kowarsch, A. Fuchs, D. Frishman, and P. Pagel, "Correlated mutations: a hallmark of phenotypic amino acid substitutions," *PLoS Comput Biol*, vol. 6, no. 9, p. e1000923, 2010.

[28] V. B. Guthrie, J. Allen, M. Camps, and R. Karchin, "Network models of tem beta-lactamase mutations coevolving under antibiotic selection show modular structure and anticipate evolutionary trajectories," *PLoS Comput Biol*, vol. 7, no. 9, p. e1002184, 2011.

[29] S. Kryazhimskiy, J. Dushoff, G. A. Bazykin, and J. B. Plotkin, "Prevalence of epistasis in the evolution of influenza a surface proteins," *PLoS Genet*, vol. 7, no. 2, p. e1001301, 2011.

[30] G. G. Perron, R. F. Inglis, P. S. Pennings, and S. Cobey, "Fighting microbial drug resistance: a primer on the role of evolutionary biology in public health," *Evolutionary applications*, vol. 8, no. 3, pp. 211–222, 2015.

[31] K. Bush, P. Courvalin, G. Dantas, J. Davies, B. Eisenstein, P. Huovinen, G. A.

Jacoby, R. Kishony, B. N. Kreiswirth, E. Kutter *et al.*, "Tackling antibiotic resistance," *Nature Reviews Microbiology*, vol. 9, no. 12, pp. 894–896, 2011.

[32] C. T. Kåhrström, "Entering a post-antibiotic era?" *Nature Reviews Microbiology*, vol. 11, no. 3, pp. 146–146, 2013.

[33] C. Walsh, *Antibiotics*. American Society of Microbiology, 2003.

[34] K. Bush, G. A. Jacoby, and A. A. Medeiros, "A functional classification scheme for beta-lactamases and its correlation with molecular structure." *Antimicrobial agents and chemotherapy*, vol. 39, no. 6, p. 1211, 1995.

[35] K. B. Holten, "Appropriate prescribing of oral beta-lactam antibiotics." *American family physician*, vol. 62, no. 3, 2000.

[36] N. Datta, P. Kontomichalou *et al.*, "Penicillinase synthesis controlled by infectious r factors in enterobacteriaceae." *Nature*, vol. 208, pp. 239–41, 1965.

[37] H. Ochman, J. G. Lawrence, and E. A. Groisman, "Lateral gene transfer and the nature of bacterial innovation," *Nature*, vol. 405, no. 6784, pp. 299–304, 2000.

[38] A. Hall and J. R. Knowles, "Directed selective pressure on a beta-lactamase to analyse molecular changes involved in development of enzyme function." *Nature*, vol. 264, no. 5588, pp. 803–804, 1976.

[39] M. Barlow and B. G. Hall, "Predicting evolutionary potential: in vitro evolution

accurately reproduces natural evolution of the tem beta-lactamase." *Genetics*, vol. 160, no. 3, pp. 823–832, 2002.

[40] A. A. Medeiros, "Evolution and dissemination of β-lactamases accelerated by generations of β-lactam antibiotics," *Clinical Infectious Diseases*, vol. 24, no. Supplement 1, pp. S19–S45, 1997.

[41] G. Jacoby and K. Bush.

[42] G. Dalbadie-McFarland, L. Cohen, A. Riggs, C. Morin, K. Itakura, and J. Richards, "Oligonucleotide-directed mutagenesis as a general and powerful method for studies of protein function," *Proceedings of the National Academy of Sciences*, vol. 79, no. 21, pp. 6409–6413, 1982.

[43] T. Palzkill and D. Botstein, "Probing β-lactamase structure and function using random replacement mutagenesis," *Proteins: Structure, Function, and Bioinformatics*, vol. 14, no. 1, pp. 29–44, 1992.

[44] W. P. Stemmer, "Rapid evolution of a protein in vitro by DNA shuffling." *Nature*, vol. 370, no. 6488, pp. 389–391, 1994.

[45] M. Zaccolo and E. Gherardi, "The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on tem-1 [beta]-lactamase," *J Mol Biol*, vol. 285, no. 2, pp. 775–783, 1999.

[46] J. Blazquez, M. I. Morosini, M. C. Negri, and F. Baquero, "Selection of naturally

occurring extended-spectrum tem beta-lactamase variants by fluctuating beta-lactam pressure." *Antimicrob Agents Chemother*, 2000.

[47] R. J. Hayes, J. Bentzien, M. L. Ary, M. Y. Hwang, J. M. Jacinto, J. Vielmetter, A. Kundu, and B. I. Dahiyat, "Combining computational and experimental screening for rapid optimization of protein properties," *Proceedings of the National Academy of Sciences*, vol. 99, no. 25, pp. 15 926–15 931, 2002.

[48] J. Osuna, A. Pérez-Blancas, and X. Soberón, "Improving a circularly permuted tem-1 beta-lactamase by directed evolution." *Protein Eng*, vol. 15, no. 6, pp. 463–470, Jun 2002.

[49] B. G. Hall, "Predicting evolution by in vitro evolution requires determining evolutionary pathways." *Antimicrob Agents Chemother*, vol. 46, no. 9, pp. 3035–3038, 2002.

[50] X. Wang, G. Minasov, and B. K. Shoichet, "Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs." *J Mol Biol*, vol. 320, no. 1, pp. 85–95, 2002.

[51] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik, "Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein." *Nature*, vol. 444, pp. 929–32, 2006.

[52] D. M. Weinreich, N. F. Delaney, M. A. Depristo, and D. L. Hartl, "Darwinian

evolution can follow only very few mutational paths to fitter proteins," *Science*, vol. 312, no. 5770, pp. 111–4, 2006.

[53] S. Bershtein, K. Goldin, and D. S. Tawfik, "Intense neutral drifts yield robust and evolvable consensus proteins." *J Mol Biol*, vol. 379, no. 5, pp. 1029–1044, 2008.

[54] B. Steinberg and M. Ostermeier, "Environmental changes bridge evolutionary valleys," *Science Advances*, vol. 2, no. 1, p. e1500921, 2016.

[55] J. Turnidge, "The pharmacodynamics of β-lactams," *Clinical infectious diseases*, vol. 27, no. 1, pp. 10–22, 1998.

[56] S. B. Vakulenko, B. Geryk, L. P. Kotra, S. Mobashery, and S. A. Lerner, "Selection and characterization of beta-lactam-beta-lactamase inactivator-resistant mutants following PCr mutagenesis of the tem-1 beta-lactamase gene." *Antimicrob Agents Chemother*, vol. 42, no. 7, pp. 1542–1548, 1998.

[57] J. Long-McGie, A. D. Liu, and V. Schellenberger, "Rapid in vivo evolution of a beta-lactamase using phagemids." *Biotechnol Bioeng*, vol. 68, no. 1, pp. 121–125, 2000.

[58] M. C. Orencia, J. Yoon, J. Ness, W. Stemmer, and R. Stevens, "Predicting the emergence of antibiotic resistance by directed evolution and structural analysis." *Nat Struct Biol*, vol. 8, no. 3, pp. 238–242, 2001.

BIBLIOGRAPHY

[59] M. Barlow and B. G. Hall, "Experimental prediction of the natural evolution of antibiotic resistance." *Genetics*, vol. 163, no. 4, pp. 1237–1241, 2003.

[60] M. Camps, J. Naukkarinen, B. P. Johnson, and L. A. Loeb, "Targeted gene evolution in escherichia coli using a highly error-prone DNA polymerase i," *Proc Natl Acad Sci U S A*, vol. 100, no. 17, pp. 9727–32, 2003.

[61] R. Fujii, M. Kitaoka, and K. Hayashi, "One-step random mutagenesis by error-prone rolling circle amplification." *Nucleic Acids Res*, vol. 32, no. 19, p. e145, 2004.

[62] ——, "Raise: a simple and novel method of generating random insertion and deletion mutations." *Nucleic Acids Res*, vol. 34, no. 4, p. e30, 2006.

[63] A. K. Holloway, T. Palzkill, and J. J. Bull, "Experimental evolution of gene duplicates in a bacterial plasmid model." *J Mol Evol*, vol. 64, no. 2, pp. 215–222, 2007.

[64] G. Kopsidas, R. K. Carman, E. L. Stutt, A. Raicevic, A. S. Roberts, M.-A. V. Siomos, N. Dobric, L. Pontes-Braz, and G. Coia, "Rna mutagenesis yields highly diverse mrna libraries for in vitro protein evolution." *BMC Biotechnol*, vol. 7, p. 18, 2007.

[65] S. Bershtein and D. S. Tawfik, "Ohno's model revisited: measuring the frequency

of potentially adaptive mutations under various mutational drifts." *Mol Biol Evol*, vol. 25, no. 11, pp. 2311–2318, 2008.

[66] M. Goldsmith and D. S. Tawfik, "Potential role of phenotypic mutations in the evolution of protein expression and stability." *Proc Natl Acad Sci U S A*, vol. 106, no. 15, pp. 6197–6202, 2009.

[67] M. A. DePristo, D. L. Hartl, and D. M. Weinreich, "Mutational reversions during adaptive protein evolution." *Mol Biol Evol*, vol. 24, no. 8, pp. 1608–1610, 2007.

[68] M. L. Salverda, E. Dellus, F. A. Gorter, A. J. Debets, J. van der Oost, R. F. Hoekstra, D. S. Tawfik, and J. A. de Visser, "Initial mutations direct alternative pathways of protein evolution," *PLoS Genet*, vol. 7, no. 3, p. e1001321, 2011.

[69] H. Jacquier, A. Birgy, H. Le Nagard, Y. Mechulam, E. Schmitt, J. Glodt, B. Bercot, E. Petit, J. Poulain, G. Barnaud *et al.*, "Capturing the mutational landscape of the beta-lactamase tem-1," *Proceedings of the National Academy of Sciences*, vol. 110, no. 32, pp. 13 067–13 072, 2013.

[70] M. F. Schenk, I. G. Szendro, M. L. Salverda, J. Krug, and J. A. G. de Visser, "Patterns of epistasis between beneficial mutations in an antibiotic resistance gene," *Molecular biology and evolution*, vol. 30, no. 8, pp. 1779–1787, 2013.

[71] E. Dellus-Gur, M. Elias, E. Caselli, F. Prati, M. L. Salverda, J. A. G. de Visser,

J. S. Fraser, and D. S. Tawfik, "Negative epistasis and evolvability in tem-1 β-lactamasethe thin line between an enzyme's conformational freedom and disorder," *Journal of molecular biology*, vol. 427, no. 14, pp. 2396–2409, 2015.

[72] M. Gniadkowski, "Evolution of extended-spectrum β-lactamases by mutation," *Clinical Microbiology and Infection*, vol. 14, no. s1, pp. 11–32, 2008.

[73] P. M. Hawkey and A. M. Jones, "The changing epidemiology of resistance," *Journal of Antimicrobial Chemotherapy*, vol. 64, no. suppl 1, pp. i3–i10, 2009.

[74] M. L. M. Salverda, J. A. G. M. de Visser, and M. Barlow, "Natural evolution of tem-1 beta-lactamase: experimental reconstruction and clinical relevance." *FEMS Microbiol Rev*, pp. 1–22, 2010.

[75] Z. Yang, "PAMl 4: phylogenetic analysis by maximum likelihood," *Mol Biol Evol*, vol. 24, no. 8, pp. 1586–91, 2007.

[76] A. Clauset, "Finding local community structure in networks," *Physical review E*, vol. 72, no. 2, p. 026132, 2005.

[77] K. Jacoby, G. Bush. (2011) Tem extended-spectrum and inhibitor resistant β-lactamases. [Online]. Available: http://www.lahey.org/Studies/temtable.asp

[78] R. P. Ambler, R. P. Ambler, A. F. Coulson, J. M. Frere, J. M. Ghuysen, B. Joris, M. Forsman, R. C. Levesque, G. Tiraby, and S. G. Waley, "A standard numbering

scheme for the class a beta-lactamases." *Philos Trans R Soc Lond B Biol Sci.*, vol. 276 (Pt 1), pp. 269–70, 1991.

[79] J. Felsenstein, "Phylip 3.5 (phylogeny inference package)," *Department of Genetics, University of Washington, Seattle*, 1993.

[80] P. Jaccard, "The distribution of the flora in the alpine zone," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[81] S. Drawz and R. A. Bonomo, "Three decades of beta-lactamase inhibitors," *Clin. Microbiol. Rev.*, vol. 23, no. 1, pp. 160–201, 2010.

[82] I. Wolfram Research, *Mathematica*. Champaign, Illinois: Wolfram Research, Inc., 2008.

[83] K. Bush and G. A. Jacoby, "Updated functional classification of beta-lactamases." *Antimicrob Agents Chemother*, vol. 54, no. 3, pp. 969–976, 2010.

[84] L. A. Abriata, M. L. Salverda, and P. E. Tomatis, "Sequence–function–stability relationships in proteins from datasets of functionally annotated variants: The case of tem β-lactamases," *FEBS letters*, vol. 586, no. 19, pp. 3330–3335, 2012.

[85] F. Walsh, *The multiple roles of antibiotics and antibiotic resistance in nature*. Frontiers Media SA, 2015.

[86] J.-M. Frère, "Beta-lactamases and bacterial resistance to antibiotics," *Molecular microbiology*, vol. 16, no. 3, pp. 385–395, 1995.

[87] J. Blazquez, M. I. Morosini, M. C. Negri, M. Gonzalez-Leiza, and F. Baquero, "Single amino acid replacements at positions altered in naturally occurring extended-spectrum tem beta-lactamases." *Antimicrob Agents Chemother*, 1995.

[88] E. B. Chaïbi, S. Farzaneh, J. Péduzzi, M. Barthélémy, and R. Labia, "An additional ionic bond suggested by molecular modelling of tem-2 might induce a slight discrepancy between catalytic properties of tem-1 and tem-2 β-lactamases," *FEMS microbiology letters*, vol. 143, no. 2-3, pp. 121–125, 1996.

[89] J. Blázquez, M.-C. Negri, M.-I. Morosini, J. Gómez-Gómez, and F. Baquero, "A237t as a modulating mutation in naturally occurring extended-spectrum tem-type β-lactamases," *Antimicrobial agents and chemotherapy*, vol. 42, no. 5, pp. 1042–1044, 1998.

[90] S. B. Vakulenko, P. Taibi-Tronche, M. Tóth, I. Massova, S. A. Lerner, and S. Mobashery, "Effects on substrate profile by mutational substitutions at positions 164 and 179 of the class a tempuc19 β-lactamase from escherichia coli," *Journal of Biological Chemistry*, vol. 274, no. 33, pp. 23 052–23 060, 1999.

[91] S. Vakulenko and D. Golemi, "Mutant tem beta-lactamase producing resistance to ceftazidime, ampicillins, and beta-lactamase inhibitors." *Antimicrob Agents Chemother*, vol. 46, no. 3, pp. 646–653, 2002.

[92] B. Caporale, N. Franceschini, M. Perilli, B. Segatore, G. M. Rossolini, and G. Amicosante, "Biochemical characterization of laboratory mutants of extended-

spectrum beta-lactamase tem-60," *Antimicrob Agents Chemother*, vol. 48, no. 9, pp. 3579–82–, 2004.

[93] K. L. Driffield, J. M. Bostock, K. Miller, A. J. O'neill, J. K. Hobbs, and I. Chopra, "Evolution of extended-spectrum beta-lactamases in a muts-deficient pseudomonas aeruginosa hypermutator." *J Antimicrob Chemother*, vol. 58, no. 4, pp. 905–907, 2006.

[94] W. Huang and T. Palzkill, "A natural polymorphism in beta-lactamase is a global suppressor." *Proc Natl Acad Sci U S A*, vol. 94, no. 16, pp. 8801–8806, 1997.

[95] I. Kather, R. P. Jakob, H. Dobbek, and F. X. Schmid, "Increased folding stability of tem-1 beta-lactamase by in vitro selection." *J Mol Biol*, 2008.

[96] M. Delaire, R. Labia, J. P. Samama, and J. M. Masson, "Site-directed mutagenesis at the active site of escherichia coli tem-1 beta-lactamase. suicide inhibitor-resistant mutants reveal the role of arginine 244 and methionine 69 in catalysis." *J Biol Chem*, 1992.

[97] U. Imtiaz, E. Billings, J. R. Knox, E. K. Manavathu, S. A. Lerner, and S. Mobashery, "Inactivation of class a .beta.-lactamases by clavulanic acid: the role of arginine-244 in a proposed nonconcerted sequence of events," *Journal of the American Chemical Society*, vol. 115, no. 11, pp. 4435–4442, 1993.

[98] I. Saves, O. Burlet-Schiltz, P. Swarén, F. Lefèvre, J. M. Masson, J. C. Promé, and J. P. Samama, "The asparagine to aspartic acid substitution at position 276 of tem-35 and tem-36 is involved in the beta-lactamase resistance to clavulanic acid." *J Biol Chem*, vol. 270, no. 31, pp. 18 240–18 245, Aug 1995.

[99] E. B. Chaibi, J. Péduzzi, S. Farzaneh, M. Barthélémy, D. Sirot, and R. Labia, "Clinical inhibitor-resistant mutants of the β-lactamase tem-1 at amino-acid position 69: kinetic analysis and molecular modelling," *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology*, vol. 1382, no. 1, pp. 38–46, 1998.

[100] E. Chaibi, D. Sirot, G. Paul, and R. Labia, "Inhibitor-resistant tem β-lactamases: phenotypic, genetic and biochemical characteristics," *Journal of Antimicrobial Chemotherapy*, vol. 43, no. 4, pp. 447–458, 1999.

[101] A. J. Baldwin, K. Busse, A. M. Simm, and D. D. Jones, "Expanded molecular diversity generation during directed evolution by trinucleotide exchange (trinex)." *Nucleic Acids Res*, vol. 36, no. 13, p. e77, 2008.

[102] J. Hecky and K. M. Mueller, "Structural perturbation and compensation by directed evolution at physiological temperature leads to thermostabilization of beta-lactamase." *Biochemistry*, 2005.

[103] D. C. Marciano, J. M. Pennington, X. Wang, J. Wang, Y. Chen, V. L. Thomas, B. K. Shoichet, and T. Palzkill, "Genetic and structural characterization of an

l201p global suppressor substitution in tem-1 β-lactamase," *Journal of molecular biology*, vol. 384, no. 1, pp. 151–164, 2008.

[104] Â. Novais, R. Cantón, T. M. Coque, A. Moya, F. Baquero, and J. C. Galán, "Mutational events in cefotaximase extended-spectrum β-lactamases of the ctx-m-1 cluster involved in ceftazidime resistance," *Antimicrobial agents and chemotherapy*, vol. 52, no. 7, pp. 2377–2382, 2008.

[105] C. Chouchani, R. Berlemont, A. Masmoudi, M. Galleni, J.-M. Frere, O. Belhadj, and K. Ben-Mahrez, "A novel extended-spectrum tem-type β-lactamase, tem-138, from salmonella enterica serovar infantis," *Antimicrobial agents and chemotherapy*, vol. 50, no. 9, pp. 3183–3185, 2006.

[106] E. C. Meng, E. F. Pettersen, G. S. Couch, C. C. Huang, and T. E. Ferrin, "Tools for integrated sequence-structure analysis with ucsf chimera," *BMC bioinformatics*, vol. 7, no. 1, p. 339, 2006.

[107] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74,, p. 47, 2002.

[108] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Res*, vol. 13, no. 11, pp. 2498–2504, 2003.

[109] C. Jelsch, L. Mourey, J. M. Masson, and J. P. Samama, "Crystal structure of escherichia coli tem1 beta-lactamase at 1.8 A resolution." *Proteins*, vol. 16, no. 4, pp. 364–383, 1993.

[110] B. Stec, K. M. Holtz, C. L. Wojciechowski, and E. R. Kantrowitz, "Structure of the wild-type tem-1 β-lactamase at 1.55 å and the mutant enzyme ser70ala at 2.1 å suggest the mode of noncovalent catalysis for the mutant enzyme," *Acta Crystallographica Section D: Biological Crystallography*, vol. 61, no. 8, pp. 1072–1079, 2005.

[111] X. Wang, G. Minasov, and B. K. Shoichet, "The structural bases of antibiotic resistance in the clinically derived mutant beta-lactamases tem-30, tem-32, and tem-34." *J Biol Chem*, vol. 277, no. 35, pp. 32 149–32 156, 2002.

[112] J. Knox, "Extended-spectrum and inhibitor-resistant tem-type beta-lactamases: mutations, specificity, and three-dimensional structure." *Antimicrobial Agents and Chemotherapy*, vol. 39, no. 12, pp. 2593–2601, 1995.

[113] C. Cantu and T. Palzkill, "The role of residue 238 of tem-1 β-lactamase in the hydrolysis of extended-spectrum antibiotics," *Journal of Biological Chemistry*, vol. 273, no. 41, pp. 26 603–26 609, 1998.

[114] M. F. Schenk, S. Witte, M. L. M. Salverda, B. Koopmanschap, J. Krug, and J. A. G. M. de Visser, "Role of pleiotropy during adaptation of tem-1 β-lactamase to two novel antibiotics." *Evol Appl*, vol. 8, no. 3, pp. 248–260, Mar 2015.

[115] M. E. J. Newman, "Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality," *Phys. Rev. E*, vol. 64, no. 1, pp. 016 132–, 2001.

[116] A. Petit, L. Maveyraud, F. Lenfant, J. P. Samama, R. Labia, and J. M. Masson, "Multiple substitutions at position 104 of beta-lactamase tem-1: assessing the role of this residue in substrate specificity." *Biochem J*, vol. 305 (Pt 1), pp. 33–40, 1995.

[117] W. Huang, Q.-Q. Le, M. LaRocco, and T. Palzkill, "Effect of threonine-to-methionine substitution at position 265 on structure and function of tem-1 beta-lactamase," *Antimicrobial Agents and Chemotherapy*, vol. 38, no. 10, pp. 2266–2269, 1994.

[118] K. V. Venkatachalam, W. Huang, M. LaRocco, and T. Palzkill, "Characterization of tem-1 beta-lactamase mutants from positions 238 to 241 with increased catalytic efficiency for ceftazidime." *J Biol Chem*, vol. 269, no. 38, pp. 23 444–23 450, 1994.

[119] R. Labia, A. Morand, K. Tiwari, J. Sirot, D. Sirot, and A. Petit, "Interactions of new plasmid-mediated beta-lactamases with third-generation cephalosporins." *Rev Infect Dis*, 1988.

[120] C. J. Troll, D. L. Alexander, J. M. Allen, and J. T. Marquette, "Mutagenesis and functional selection protocols for directed evolution of proteins in e. coli," *The Journal of Visualized Experiments*, vol. in press, pp. –, 2010.

BIBLIOGRAPHY

[121] J. Delmas, F. Robin, F. Bittar, C. Chanal, and R. Bonnet, "Unexpected enzyme tem-126: role of mutation asp179glu," *Antimicrobial agents and chemotherapy*, vol. 49, no. 10, pp. 4280–4287, 2005.

[122] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[123] J. P. Huelsenbeck and F. Ronquist, "Mrbayes: Bayesian inference of phylogenetic trees," *Bioinformatics*, vol. 17, no. 8, pp. 754–5, 2001.

[124] B. G. Hall, *Phylogenetic trees made easy: a how-to manual*. Sinauer Associates Sunderland, 2004, vol. 547.

[125] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler, "Database issue," *GenBank, Nucleic Acids Res*, vol. 37, no. 10, pp. 26–31, 2009.

[126] Joint genome institute integrated microbial genomes database. [Online]. Available: https://img.jgi.doe.gov/cgi-bin/m/main.cgi

[127] V. M. Markowitz, I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams *et al.*, "Img: the integrated microbial genomes database and comparative analysis system," *Nucleic acids research*, vol. 40, no. D1, pp. D115–D122, 2012.

BIBLIOGRAPHY

[128] (2013) The nih human microbiome project - reference genomes data. [Online]. Available: http://hmpdacc.org/HMRGD/

[129] D. Verma, D. J. Jacobs, and D. R. Livesay, "Variations within class-a beta-lactamase physiochemical properties reflect evolutionary and environmental patterns, but not antibiotic specificity," *PLoS Comput Biol*, vol. 9, no. 7, p. e1003155, 2013.

[130] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC Bioinformatics*, vol. 10, p. 421, 2009.

[131] V. K. Vaidya, "Horizontal transfer of antimicrobial resistance by extended-spectrum β lactamase-producing enterobacteriaceae." *J Lab Physicians*, vol. 3, no. 1, pp. 37–42, Jan 2011. [Online]. Available: http://dx.doi.org/10.4103/0974-2727.78563

[132] E. V. Koonin and A. S. Novozhilov, "Origin and evolution of the genetic code: the universal enigma." *IUBMB Life*, vol. 61, no. 2, pp. 99–111, Feb 2009. [Online]. Available: http://dx.doi.org/10.1002/iub.146

[133] L. Tierney, "Markov chains for exploring posterior distributions," *the Annals of Statistics*, pp. 1701–1728, 1994.

BIBLIOGRAPHY

[134] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, 1970.

[135] R. Lanfear, B. Calcott, S. Y. Ho, and S. Guindon, "Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses," *Mol Biol Evol*, vol. 29, no. 6, pp. 1695–701, 2012.

[136] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist, "Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference," *Bioinformatics*, vol. 20, no. 3, pp. 407–15, 2004.

[137] W. R. Gilks, *Markov chain monte carlo*. Wiley Online Library, 2005.

[138] A. Rambaut and A. Drummond, "Tracer: a program for analysing results from Bayesian MCMC programs such as beast & mrbayes," 2003.

[139] J. A. Nylander, J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford, "Awty (are we there yet?): a system for graphical exploration of mcmc convergence in bayesian phylogenetics," *Bioinformatics*, vol. 24, no. 4, pp. 581–583, 2008.

[140] M. Barrett, M. J. Donoghue, and E. Sober, "Against consensus," *Systematic Zoology*, vol. 40, no. 4, pp. 486–493, 1991.

[141] W. Cai, J. Pei, and N. V. Grishin, "Reconstruction of ancestral protein sequences and its applications," *BMC evolutionary biology*, vol. 4, no. 1, p. 1, 2004.

[142] J. P. Bollback, "Simmap: stochastic character mapping of discrete traits on phylogenies," *BMC bioinformatics*, vol. 7, no. 1, p. 88, 2006.

[143] R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, C. Lozupone, D. McDonald, M. Robeson, R. Sammut, S. Smit, M. J. Wakefield, J. Widmann, S. Wikman, S. Wilson, H. Ying, and G. A. Huttley, "Pycogent: a toolkit for making sense from sequence," *Genome Biol*, vol. 8, no. 8, p. R171, 2007.

[144] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski *et al.*, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[145] E. Paradis, J. Claude, and K. Strimmer, "Ape: analyses of phylogenetics and evolution in r language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.

[146] M. Pagel, "Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 255, no. 1342, pp. 37–45, 1994.

[147] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

BIBLIOGRAPHY

[148] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal, Complex Systems*, vol. 1695, no. 5, pp. 1–9, 2006.

[149] T. Alahakoon, R. Tripathi, N. Kourtellis, R. Simha, and A. Iamnitchi, "K-path centrality: A new centrality measure in social networks," in *Proceedings of the 4th Workshop on Social Network Systems*. ACM, 2011, p. 1.

[150] M. A. Fares and S. A. Travers, "A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses," *Genetics*, vol. 173, no. 1, pp. 9–23, 2006.

[151] J. Y. Dutheil, "Detecting coevolving positions in a molecule: why and how to account for phylogeny," *Briefings in bioinformatics*, vol. 13, no. 2, pp. 228–243, 2012.

[152] S. Kauffman and S. Levin, "Towards a general theory of adaptive walks on rugged landscapes," *Journal of theoretical Biology*, vol. 128, no. 1, pp. 11–45, 1987.

[153] S. Garimalla, T. Kieber-Emmons, and A. D. Pashov, "The patterns of coevolution in clade b hiv envelope's n-glycosylation sites," *PloS one*, vol. 10, no. 6, p. e0128664, 2015.

[154] L. Wang, S. G. Valderramos, A. Wu, S. Ouyang, C. Li, P. Brasil, M. Bonaldo, T. Coates, K. Nielsen-Saines, T. Jiang, R. Aliyari, and G. Cheng, "From

mosquitos to humans: Genetic evolution of zika virus," *Cell Host & Microbe*, 2016/04/17. [Online]. Available: http://dx.doi.org/10.1016/j.chom.2016.04.006

[155] A. F. Poon, F. I. Lewis, S. D. Frost, and S. L. K. Pond, "Spidermonkey: rapid detection of co-evolving sites using bayesian graphical models," *Bioinformatics*, vol. 24, no. 17, pp. 1949–1950, 2008.

[156] M. Pagel and A. Meade, "Bayestraits," *Computer program and documentation available at http://www. evolution. rdg. ac. uk/BayesTraits. html*, 2007.

Vita

Violeta Beleva Guthrie

Department of Biomedical Engineering
Institute for Computational Medicine
Johns Hopkins University, 220 Hackerman Hall
3400 N. Charles St. Baltimore, MD 21218

Educational History

Ph.D. in Biomedical Engineering, Johns Hopkins University	expected 2016
M.A. in Physics, Kent State University	2005
B.S. in Physics and Applied Mathematics, Kent State University	2004

Research Experience

Graduate Research Assistant, Johns Hopkins University (2009-2016)
Advisor: Prof. Rachel Karchin, Biomedical Engineering
Topic: Network models of protein evolution and metrics for predicting the functional impact of multiple mutations

Graduate Research Assistant, Johns Hopkins University (2007-2009)
Advisor: Dr. Andre Levchenko, Biomedical Engineering
Topic: Network centrality analysis of metabolic and regulatory pathways in E. coli

Undergraduate Research, Los Alamos National Laboratory (2004-2005)
Advisors: Dr. Angel Garcia and Dr. Kim Rasmussen
Topic: Atomistic molecular dynamics simulations of DNA denaturation

Peer-reviewed Publications

1. Niknafs N, **Guthrie VB**, Naiman DQ, Karchin R (2015) SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. PLoS Computational Biology 11(10):e1004416.

2. Chung C, **Guthrie VB**, Masica DL, Tokheim C, Kang H, Richmon H, Agrawal N, Fakhry C, Quon H, Subramaniam RM, Zuo Z, Seiwert T, Chalmers ZR, Frampton GM, Ali SM, Yelensky R, Stephens PJ, Miller VA, Karchin R, Bishop JA (2015) Genomic alterations in head and neck squamous cell carcinoma determined by cancer gene-targeted sequencing. Annals of Oncology 26(6):1216-1223.

3. Chen Y-C, Douville C, Wang C, Niknafs N, Yeo G, **Guthrie VB**, Carter H, Stenson PD, Cooper DN, Li B, Mooney S, Karchin R (2014) A probabilistic model to predict clinical phenotypic traits from genome sequencing. PLOS Computational Biology 10(9):e1003825.

4. Jiao Y, Pawlik TM, Anders RA, Selaru FM, Streppel MM, Lucas DJ, Niknafs N, **Guthrie VB**, Maitra A, Argani P, Offerhaus GJ, Roa JC, Roberts LR, Gores GJ, Popescu I, Alexandrescu ST, Dima S, Fassan M, Simbolo M, Mafficini A, Capelli P, Lawlor RT, Ruzzenente A, Guglielmi A, Tortora G, de Braud F, Scarpa A, Jarnagin W, Klimstra D, Karchin R, Velculescu VE, Hruban RH, Vogelstein B, Kinzler KW, Papadopoulos N, Wood LD (2013) Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. Nature Genetics 45(12):1470-1473.

5. **Guthrie VB**, Allen J, Camps M, Karchin R (2011) Network models of TEM β-lactamase mutations coevolving under antibiotic selection show modular structure and anticipate evolutionary trajectories. PLoS Computational Biology 7(9):e1002184.

6. Zou L, Wang J, **Beleva V**, Kooijman EE, Primak SV, Risse J, Weissflog W, Jkli A, Mann EK (2004) Langmuir monolayers of bent-core molecules. Langmuir 20(7):2772-2780.

VITA

Presentations, posters, and abstracts

1. **Guthrie VB**, Standley M, Camps M, Karchin R (2014) A Directed Mutation-centric Network Model of Protein Adaptation. Evolution Annual Meeting, Raleigh, NC June 20, 2014.

2. **Guthrie VB** (2011) The TEM -lactamase Protein Residue Coevolution Network. Johns Hopkins University Institute for Computational Medicine Annual Meeting, Annapolis, MD May 11 2011

3. **Beleva V**, Levchenko A (2008) Network centrality of the TCA cycle in the E. coli metabolism. Regulatory Genomics Systems Biology Meeting, MIT, Boston, MA, November 20, 2008.

4. **Beleva V**, Garcia A (2005) Modeling DNA Bubble Formation at the Atomic Scale. Biophysical Society 49th Annual Meeting, Long Beach, CA February 12, 2005.

Scholarships and fellowships

- **Graduate Scholarship** Johns Hopkins Institute for Multiscale Modeling of Biological Interactions (IMMBI), Johns Hopkins University (2005-2007)

- **Undergraduate Research Fellowship** Los Alamos National Laboratory (2004-2005)

- **Undergraduate Scholarship** Kent State University Honors College (20002004)

- **Undergraduate Research Fellowship** University of California San Diego (UCSD) Center for Theoretical Biophysics (Summer 2003)

- **Undergraduate Research Fellowship** National Institute of Standards and Technology (Summer 2002)

Academic and other honors

- Cummins and Harschbarger Awards in Mathematics, Kent State University (2002)

- John Wiley Academic Achievement Award in Physics, Kent State University (2002)

Teaching and mentoring

- **Teaching Assistant** Foundations of Computational Biology and Bioinformatics II, Johns Hopkins University (Spring 2013)
- **Research Mentor** for Grace Yeo Hui Ting, an undergraduate student in the Department of Biomedical Engineering. Johns Hopkins University, Baltimore, MD (2012-2013)
- **Teaching Assistant** Systems Bioengineering II, Johns Hopkins University
- **Research Mentor** for Allison Chlada, a student at the Baltimore Polytechnic Institute. Baltimore, MD (Spring 2012)
- **Research Mentor** for Andrea Corredor, an undergraduate student in the Department of Biomedical Engineering. Johns Hopkins University, Baltimore, MD (Spring 2012)