Computational Modeling of Drug Resistance: Structural and Evolutionary Models

by

Maryah Safi

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Computational Modeling of Drug Resistance: Structural and Evolutionary Models

Maryah Safi
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2014

Active site mutations that disrupt drug binding are an important mechanism of drug resistance. Such resistance causing mutations impair drug binding, thus reducing drug efficacy. Knowledge of potential resistance mutations, before they are clinically observed, would be useful in a number of ways. During the lead prioritization phase of drug development, this knowledge may direct the research team away from candidate drugs that are most likely to experience resistance. In the clinical setting, knowledge of potential resistance mutations could allow the development of treatment regimens, with drug cocktails likely to maximize efficacy. In this thesis I present a structure-based approach to predict resistance and its evolution. This method utilizes a two-pass search, which is based on a novel protein design algorithm, to identify mutations that impair drug binding while maintaining affinity for the native substrate. The approach is general and can be applied to any drug-target system where a structure of the target protein, its native substrate and the drug is available. Furthermore, it requires no training data for predictions and instead predicts resistance using structural principles. Finally, I use approximate force-field calculations from MMPBSA and simple assumptions about the relationship between binding energy and fitness to build fitness landscapes for a target protein under selective pressure from either a single drug or a drug cocktail. I use a Markov-chain based model to simulate evolution on this fitness landscape and to predict the likely evolutionary trajectories for resistance starting from a wild-type. The structure-based method was used to probe resistance in four

drug-target systems: isoniazid-enoyl-ACP reductase (tuberculosis), ritonavir-HIV protease (HIV), methotrexate-dihydrofolate reductase (breast cancer and leukemia), and gleevec-ABL kinase (leukemia). This method was validated using clinically known resistance mutations for all four test systems. In all cases, it correctly predicts the majority of known resistance mutations. Furthermore, exploiting the relationship between binding energy, drug resistance and fitness of a mutant, evolution was simulated on the HIV-protease fitness landscape. This hybrid evolutionary model further improves the resistance prediction. Finally, good agreement between these evolutionary simulations and observed evolution of drug resistance in patients was found.

Dedication

To Baba, Ma and the Three Stooges

When I am with you, we stay up all night.
When you're not here, I can't go to sleep.
Praise God for those two insomnias!
And the difference between them.

Rumi

It was the joys of the former and motivation from the second that helped me through.

Acknowledgements

I would like to express my gratitude to my supervisors Dr. Ryan Lilien and Dr. Alan Moses for their guidance and continuous support during my PhD. Their knowledge and patience helped me overcome many stumbling blocks during this journey.
I would also like to thank other members of my thesis committee Dr. Brendan Frey and Dr. Allan Borodin. Their insightful comments and questions helped guide my research.
My sincere thanks also goes to my fellow students and members of Lilien and Moses labs: Abraham Heifets, Izhar Wallach, Navdeep Jaitly, Nilgun Donmez, Gelila Tilahun, Louis-Francois Handfield, Alex Nguyen Ba, Gavin Douglas, Taraneh Zarin and Bob Strome, for stimulating discussions and their friendly assistance at all hours.
Last but certainly not the least, I would like to thank Ayesha Saeed, Bushra Mir, Salman Mohsin, Rabia Nasir and my family and friends for supporting me at times I doubted myself, and for always lending me a listening ear and a helping hand.

Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

1.1 Drug Resistance

Evolution of drug resistance is a leading cause of treatment failure, making it a serious and growing public health risk (Knobler et al., 2003; Baddeley et al., 2013; Holohan et al., 2013). Drug resistance is defined as the reduced ability of a drug to cure or suppress a disease. At a molecular level, the Red Queen hypothesis (van Valen, 1973) suggests that as a patient begins a particular treatment regimen, adaptations occur in the drug targets (bacteria, virus or tumor cells) that allows them to survive and function in presence of the drug. This evolution in target cell population for survival under adverse circumstances i.e. under drug pressure results in reduced efficacy of the drug, thereby conferring drug resistance. Drug resistance is prevalent in a vast majority of diseases ranging from bacterial and viral epidemics such as malaria, tuberculosis and AIDS, to tumor related diseases such as cancer (Le Bras and Durand, 2003; Baddeley et al., 2013; Holohan et al., 2013).

Since evolution of resistance is a primary cause of treatment failure, various strategies are used in clinical practice to combat it. One such clinical strategy is the change of treatment as resistance to the drug being administered arises. This is often the case for diseases such as tuberculosis and malaria, where after the failure of first-line (first administered) drugs, second-line and third-line drugs are used (Miller et al., 2013; Zumla et al., 2013). Unfortunately this recourse to second and third-line drugs, while effective in some patients, is far from optimal, since first line drugs are often selected as they are effective in a large section of population with minimal side effects, making them the ideal drug candidates in the absence of resistance. An alternate clinical strategy to combat drug resistance in patients involves the use of multiple drugs or of drug cocktails (Majori, 2004; Spanagel and

Vengeliene, 2013). These drug cocktails are expected to increase efficacy of treatment and make appearance of drug resistance less likely. Unfortunately, the use of drug cocktails can correspond to an increase in side effects in patients. Furthermore, resistance often evolves to drug cocktails as well, impairing their efficacy (Daniela et al., 2003). Thus drug resistance is often reminiscent of an *evolutionary arms race* between the drugs or treatment regimens and the drug target. As the treatment regimen is changed to inhibit the (resistant) targets, mutations in the drug target arise to enable an escape from therapy.

Despite efforts to overcome resistance in clinical practice, many drug targets develop resistance to multiple drugs, leading to the appearance of highly resistant strains. An example of this is the emergence of the highly resistant *Mycobacterium tuberculosis* strains known as the multi drug resistant or MDR TB. These MDR TB strains are resistant to both Isoniazid and Rifampicin, which are first-line drugs for TB, thereby necessitating the use of second line drugs for the treatment of MDR TB (Baddeley et al., 2013). Often these second-line drugs are more expensive compared to Isoniazid and Rifampicin and cause numerous side effects in the patients (Baddeley et al., 2013). Furthermore, populations that are immunocompromised e.g. patients suffering from HIV, are much more likely to acquire and retain these resistant strains of tuberculosis, often causing fatalities (Daley and Caminero, 2013). To further complicate matters, these MDR TB bacteria give rise to increasingly resistant strains that are resistant to some of the most effective second-line anti-TB drugs as well, leading to a form of TB known as extensively drug-resistant tuberculosis or XDR TB. Rise of these XDR TB strains has caused serious concerns about the treatability of a future TB epidemic (LoBue et al., 2009; Daley and Caminero, 2013).

Currently, attempts to control and overcome drug resistance are focused at the clinical level and are executed once resistance is observed. However, *a priori* knowledge of resistance, i.e. before it arises, could be beneficial in a number of ways including lead prioritization during drug development and drug target selection. As pathogens become resistant to the existing drugs, development of new and innovative drugs can help disease management. Unfortunately this option is associated with high costs and risk. The cost of bringing a new drug to the market (i.e. from lead discovery to clinical trials and final approval) ranges anywhere between $870 million to $1.8 billion (DiMasia et al., 2003; Adams and Brantner, 2006). If resistance to the new drug arises quickly, its efficacy is compromised making the return on investment of the new drug development questionable in this scenario. Under these circumstances, *a priori* knowledge of resistance can be used to prioritize drug leads (lead compounds) during development such that a lead compound against which it is difficult to

develop resistance is favoured compared to another lead compound. In addition to aiding lead prioritization, such *a priori* knowledge of resistance can also help in prioritizing drug targets. Drugs often aim to inhibit an essential functional protein of the disease cell (bacteria, virus or tumor), thereby suppressing its activity. Thus, in cases where more than one potential drug target proteins exist, a protein that cannot evolve resistance easily would be a preferred target for drug development. In addition to drug development, *a priori* knowledge of resistance can aid treatment design as well. Knowing what resistant forms are likely to evolve for a given patient and different drug regimes, we can potentially design better treatment strategies that minimize the evolution of resistance.

As previously noted, *a priori* knowledge of resistance can be beneficial in a number of ways. However, acquiring this knowledge entails significant and possibly inhibiting scale of wet lab effort. For instance, in order to determine the effect of the different mutational changes happening in a target bacterial cell on drug resistance, we would need to engineer these mutant bacteria in the wet lab, followed by assays to quantify and measure the effect of these mutations on drug efficacy. Since, an exponential number of such mutant pathogens exist, exhaustive wet lab experiments to predict resistance are not feasible. On the other hand, a computational method that can predict the effect of these mutant pathogen forms on drug resistance is a more efficient option. However, it is important to note, that the aim of such computational approaches is not to replace, but to supplement and guide the wet lab efforts. For instance, a computational approach can select a small set of highly likely resistance conferring mutant bacteria. Wet lab experiments can then be performed on this small set, as opposed to the exponential number of mutants, to confirm drug resistance.

The aim of this thesis is to develop computational models that allow us to predict drug resistance *a priori* i.e. before it arises. The methods presented in this thesis are not intended as a replacement for wet lab efforts. Instead, they are intended to guide and supplement such efforts. The rest of this chapter is organized as follows. First, a brief biological overview providing necessary biological background to understand drug resistance is provided. Second, the approach taken in this thesis and its chapter contents are described.

1.2 Biological Background

Cells of living organisms contain molecules known as DNA. These DNA molecules encode genetic information as a sequence of nucleotides (one of A, T, G or C). Thus, a simple representation of DNA can be the sequence of nucleotides that compose it. For instance, a

Table 1.1: **Amino Acid Codes.** Single and three letter codes for all 20 naturally occurring amino acids are provided.

Amino Acid	Three Letter Code	Single Letter Code
Alanine	ala	A
Cystein	cys	C
Aspartic Acid	asp	D
Glutamic Acid	glu	E
Phenylalanine	phe	F
Glycine	gly	G
Histidine	his	H
Isoleucine	ile	I
Lysine	lys	K
Leucine	leu	L
Methionine	met	M
Asparagine	asn	N
Proline	pro	P
Glutamine	gln	Q
Arginine	arg	R
Serine	ser	S
Threonine	thr	T
Valine	val	V
Tryptophan	trp	W
Tyrosine	tyr	Y

sequence such as ATGCAAATCG can represent part of a DNA molecule. The genetic information encoded by the DNA controls various cellular functions and determines phenotype via *genes*. These genes are stretches of DNA and its functional information carrying units.

Before its function can be performed in the cell, the gene sequence has to be translated into the protein encoded by that gene. Like DNA, a protein is also a large molecule. However, instead of nucleotides, proteins consist of a single chain of amino acids. This chain often folds into a distinct 3D structure in the cell to allow the protein to perform its various functions. Similar to the gene that encodes it, a protein can also be represented by its amino acid sequence. However, unlike a gene sequence that has a four-letter alphabet consisting of the nucleotides A,T,G or C, the protein sequence alphabet is twenty-letter, where each letter represents one of the twenty amino acids (see Table 1.1 for a list of amino acids and their single letter codes).

As mentioned previously, the information contained in the gene has to be converted into the protein it encodes in the cell. The process where a gene produces its encoded protein

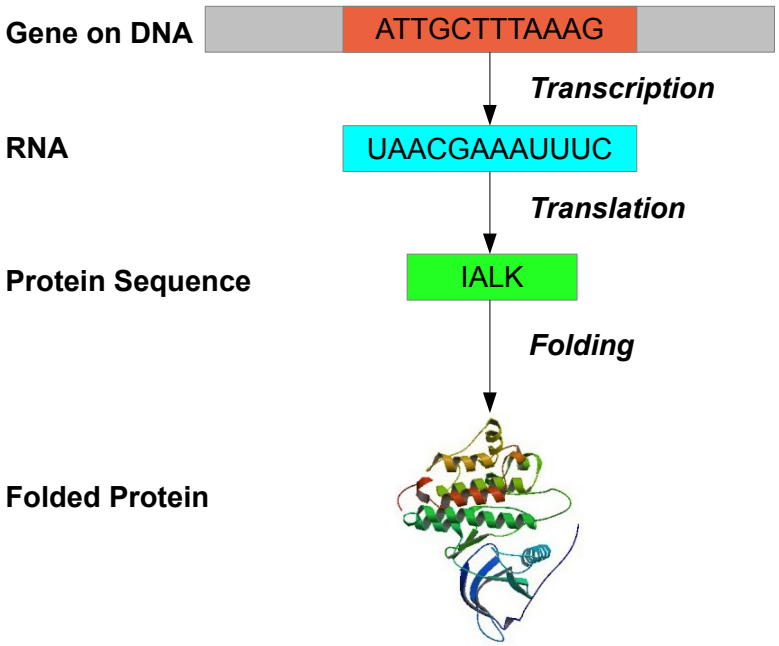

Figure 1.1: **Gene Expression** A gene is transcribed into RNA. This RNA is then translated into a protein sequence which folds into the 3D structure of the protein. The folded protein then performs a number of cellular functions.

is called *gene expression*. Gene expression involves two steps: transcription and translation. During transcription, the DNA sequence of the gene is copied into an intermediate RNA sequence by a molecule known as the RNA polymerase. During translation, this intermediate sequence is then decoded by the cell to produce the resulting protein's amino acid sequence (also known as a polypeptide). Finally, this protein sequence folds into its 3D structure. During gene expression, stretches of three nucleotide basis in the DNA sequence are translated to a single amino acid in the resulting protein sequence. (Thus the coding sequence of the gene is thrice the length of the protein sequence it encodes.) These stretches of three nucleotides are also known as codons. Since there are four nucleotides, there are $4^3 = 64$ distinct codons. As these 64 codons encode for 20 amino acids, there is redundancy in the genetic code implying that many codons encode for the same amino acid. The process of gene expression is illustrated in Figure 1.1.

Drug Mechanism The translated proteins perform a vast range of functions within the cell including DNA replication, catalysis of reactions, molecule transport, stimulus response etc. In order to perform these functions, the protein often interacts with other small molecules known as substrates or ligands. These ligands interact with the protein by binding it at

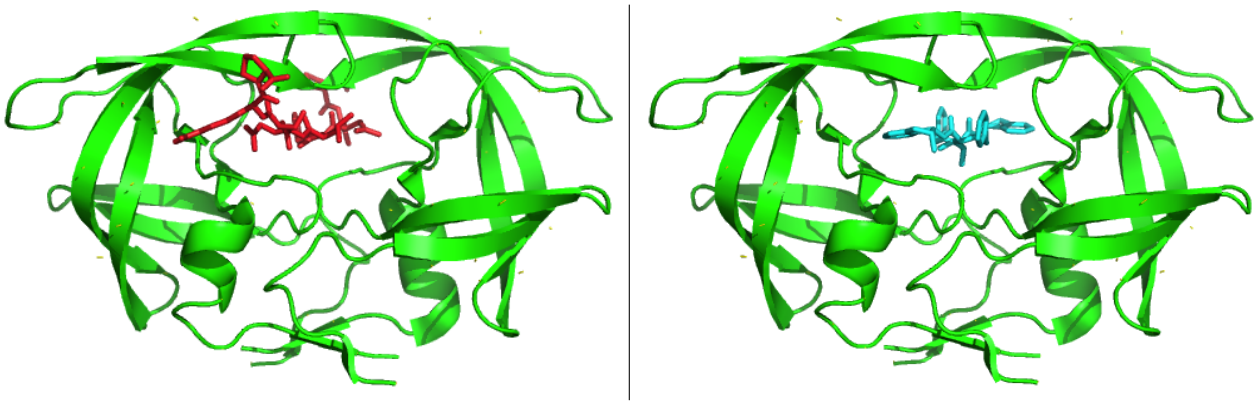

Figure 1.2: **HIV Protease** Left: HIV protease bound to its natural substrate (the peptide shown in magenta) (PDB ID:1KJF,(Prabhu-Jeyabalan et al., 2002)) Right: HIV protease bound to a protease inhibitor. (PDB ID: 1EBY, (Andersson et al., 2003)) Since the active site is occupied by the PI, HIV protease can no longer bind its natural substrate. Thus, the drug inhibits protease function.

a specific interaction location, also known as the *active site* (or alternately a *binding site*), and the resulting protein-ligand structure is known as a protein-ligand complex or simply a complex. The HIV protease protein is presented here as an example. The task of HIV protease in the virus is to cleave newly synthesized polypeptides at appropriate location so they can fold into their 3D structures. In order to perform this function, the HIV protease binds peptide ligands.

The task of a drug is often to suppress or kill the disease cell, e.g. by inhibiting a function necessary for its survival or replication. In order to achieve this, the drug is often designed to bind an essential protein and prevent it from binding its natural substrate, thereby inhibiting the natural function of the protein. An example of this is a class of drugs known as the protease inhibitors (PIs) aimed at inhibiting the function of HIV protease. These PIs bind the HIV protease at the peptide binding site. Since the binding site is occupied by the drug, the protease can no longer bind its natural substrate, i.e., the peptides. Thus, the polypeptide chains cannot be cleaved by the drug-bound protease. Since cleaving of the polypeptides by HIV protease is essential for the virus's ability to replicate and infect cells, its inhibition by the PIs renders the virus uninfectious (see Figure 1.2).

1.3 Mechanisms of Drug Resistance

In general, resistance to a drug can be acquired in one of four ways: mutations in non-target genes, reduced effective concentration of the drug in the cell, modification of the drug by the pathogen and finally, by point mutations in the target protein (Rosenzweig, 2012; Konig et al., 2013). These mechanisms are briefly described below.

Perhaps the most direct and prevalent method of drug resistance is through acquisition of point mutations in the drug target protein. A point mutation (or mutation) is a change in the DNA sequence of a gene encoding for the protein. These mutations in the gene's DNA sequence consequently alter the amino acid sequence of the translated protein, thereby altering its structure and impairing its binding with the drug. While these mutations can be present anywhere in the target protein, the most direct and well understood impact is caused by the mutations located within the drug binding site (Ode et al., 2006). On the other hand, mechanisms of the compensatory mutations which are located away from the binding site and do not directly alter drug binding are more ambiguous and varied, possibly including large scale structural changes or a change in protein dynamics (Ode et al., 2005; Rosenzweig, 2012). In addition to point mutations in the genes encoding for the drug target, mutations in other genes can cause drug resistance indirectly (de Vos et al., 2013). This is possible in a number of different ways. For instance, consider a scenario where a cellular function can be performed by two different proteins A and B. If protein A is the drug target, any change in the gene encoding protein B which increases its expression (indicating that more of B will be produced in the cell) will offset the fact that protein A has been inhibited by the drug. The cellular need for protein A is thereby compensated by an abundant protein B, causing drug resistance. Other than mutations, modification of the drug structure by a protein within the cell can inactivate the drug and cause drug resistance. An example of this mechanism is the penicillin resistance exhibited by some bacteria. These penicillin resistant bacteria produce a protein called beta-lactamase that alters penicillin's structure through hydrolysis. As a result of this hydrolysis, the drug undergoes a structural change that destroys its antibacterial properties, causing penicillin resistance (Drawz and Bonomo, 2010). Finally, another fascinating mechanism by which cells acquire drug resistance includes the use of efflux pumps. An efflux pump is a protein that is responsible for extrusion of toxic substances. Since a drug is a toxic substance for a pathogen cell, these efflux pumps remove the drug from the cell as well. In cases where the gene expression for efflux pumps is increased, indicating increased efflux activity in the cell, more of the drug is removed from

the cell. This results in reduced uptake and reduced concentration of the drug in the cell, thereby causing drug resistance (Konig et al., 2013). Figure 1.3 illustrates this process.

Measuring Drug Resistance In the wet lab, the degree of drug resistance conferred is measured by an IC50 value. This IC50 value denotes the concentration of the drug needed to reduce the activity of the drug target by 50%. The intuition is that as a mutant causes drug resistance, the amount of drug needed to reduce its target protein's activity increases, thereby increasing the IC50 for a resistant mutant.

Acquisition of drug resistance is a complicated biological process that is not understood fully. A number of confounding factors influence our understanding of its mechanism e.g. effects of the human immune system on drug resistance, effects of drug dosage and drug metabolism in humans etc. The methods described in this thesis do not address these aspects of drug resistance. Furthermore, our understanding of most mechanisms of drug resistance and of the biophysical and biochemical laws governing them is limited. For instance, our knowledge of gene interaction networks in different pathogens or humans is sparse. It will therefore be extremely challenging to accurately predict the effects of over expression of one mutant gene on another to acquire drug resistance. Similarly, most efflux pumps are not specific to drugs i.e. they are likely to extrude a number of potentially beneficial cellular products along with the drug as well. However, without complete knowledge of these extruded cellular products (and their impact/function in the cell), it is not possible to accurately quantify the effects of efflux pump over expression on drug resistance. However, despite our limited understanding of some of the more elusive mechanisms of drug resistance, one the most direct and common ways in which drug resistance is acquired is by point mutations in the drug target. Often, these point mutations are present within the active site of the target protein, thereby having direct influence on drug and substrate binding. Thus, to predict drug resistance, this particular mechanism of resistance acquisition is a potentially attractive starting point. However, computational methods that are capable of predicting such resistance *a priori* in a drug-target system are limited. In addition, these existing methods rely on the knowledge of previously observed resistance in a drug target to predict further resistance. Thus, these methods are disease-specific and are not applicable to drug targets where such knowledge is sparse such as in emerging diseases or new drug targets.

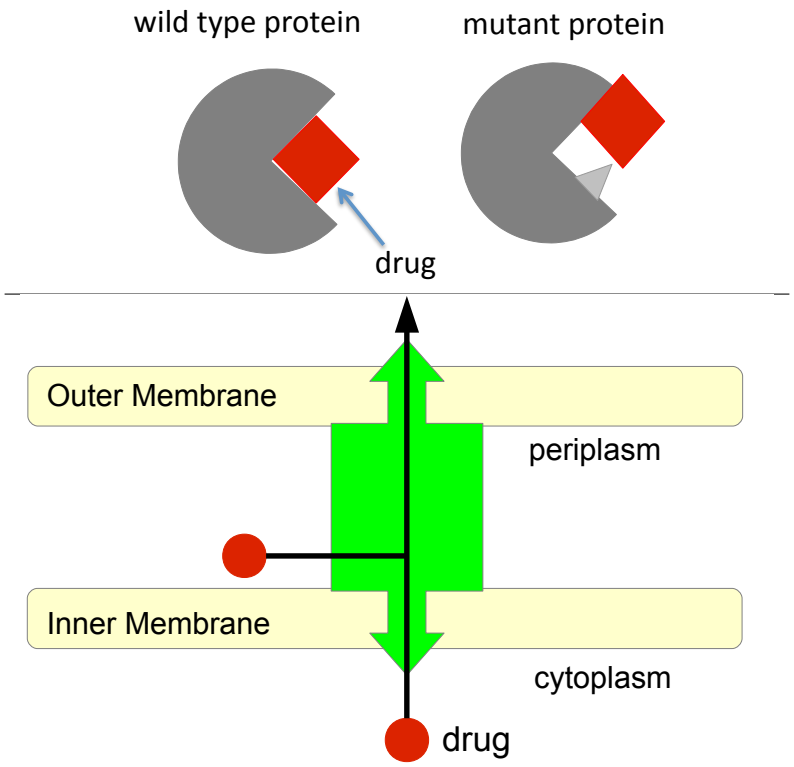

Figure 1.3: **Mechanisms of Resistance** Top: The drug (red) inhibits the wild type protein on the left by binding its active site. The altered local structure of the mutant protein on the right prevents this drug binding causing resistance. Bottom: An efflux pump (green) pumps out the drug from both the cytoplasm and periplasm, reducing drug concentration in the cell and causing drug resistance.

1.4 Thesis Outline

In this thesis, computational methods for predicting drug resistance *a priori* i.e. before it arises are presented. I present computational approaches that will model drug resistance arising due to point mutations in the drug target's active site. The chapter details for this thesis document follow.

The methods presented in this thesis borrow heavily from different areas in Computer Science. Chapter 2 contains sufficient background for the reader to understand the contents of this thesis. The chapter begins with a summary of previous computational approaches that attempt to predict resistance. Next, since the methods presented in this thesis are based on the search and score mechanisms from computational protein design, Chapter 2 also contains a review of these methods and related concepts. Finally, the chapter closes

with a description of Markov processes and computational models of evolution.

In this thesis, I aim to develop an algorithm that can predict resistance in any given drug target from structural principles and does not rely on existing data. To achieve this, resistance acquisition is modelled as a protein redesign problem. However, existing protein design approaches do not extend easily to the resistance problem. A resistance conferring mutant often has a small number of mutations relative to the wild type (Parikh et al., 1999; Arnold et al., 2005; Volpato et al., 2007a; Bang et al., 2011; Chen et al., 2013). Existing protein design algorithms are not well suited to search for these *restricted* mutants. To address this issue, Chapter 3 introduces restricted dead-end elimination (rDEE), which is a protein redesign algorithm tailored for restricted protein design problems such as those of resistance. In Chapter 3, the rDEE algorithm is explained and a number of increasingly efficient extensions are derived. The chapter closes with a proof of correctness for the algorithms presented.

Chapter 4 elaborates the algorithm to search and score resistance mutations in an arbitrary drug target using the rDEE algorithm from Chapter 3. A hierarchical methodology that employs rDEE, as the first of a two pass search-and-score algorithm followed by improved scoring using enhanced molecular modelling, is described. I use this method to predict resistance in four drug target systems including isoniazid-enoyl ACP Reductase (tuberculosis), Gleevec- ABL Kinase (leukaemia), Methotrexate-DHFR (chemotherapy) and Ritonavir-HIV Protease (AIDS). This chapter presents the predicted results and compares them to clinically determined resistance mutations for the four systems studied.

Chapters 3 and 4 focus on the development of computational methods to predict drug resistance from structural principles. In Chapter 5, resistance is modelled as an evolutionary process that takes a drug sensitive wild type protein to a drug resistant mutant. Chapter 5 uses the HIV protease gene from HIV-1 virus as an example system to study evolution of resistance to ritonavir. Highly likely evolutionary trajectories for resistance under ritonavir are predicted. Furthermore, the chapter includes an analysis of the evolution of drug resistance under drug cocktails and for varying levels of drug adherence by patients.

Finally, Chapter 6 concludes this thesis with a discussion of its overall contributions, results and limitations. Future extensions to this work and possible directions are also included.

Chapter 2

Background

The appearance of drug resistant mutants is a serious health issue today. A priori knowledge of resistance for drug targets can help combat drug resistance in various ways. To this effect, a number of computational studies has been aimed at predicting whether a mutant confers resistance. This chapter begins with a brief overview of these previous works. In addition, this thesis has been influenced by work in fields of computational protein design and computational evolution. This chapter also provides sufficient background for the reader in these areas.

2.1 Previous Computational Approaches to Predict Resistance

Existing works that aim to predict resistance in drug targets can be broadly categorized into two categories: sequence-based and structure-based. Of these, the sequence-based approaches work by exploiting either the gene or amino acid sequences of the drug target protein. Often, these sequence-based approaches are learning or inference techniques insofar that they rely on existing (albeit partial) data to train a computational model to predict resistance. On the other hand, structure based approaches rely on the 3D structure of the drug target protein to probe resistance. Previous works in both these categories are further described below.

2.1.1 Computational Sequence-Based Approaches

In general, given a query sequence and a drug, the goal of the sequence-based learning approaches is to classify the query sequence as either resistant or susceptible. In order to achieve this, these sequence-based approaches use existing data and train a learning model to learn the correlation between different mutations and their effects on drug resistance. Therefore, sequence-based approaches are dependent on the availability and abundance of training data that links mutant sequences to a resistance phenotype (i.e. if the mutant sequence is resistant or susceptible). Due to this limitation, attempts at predicting resistance using these approaches have been mostly limited to AIDS where a wealth of training data is available.

AIDS is a disease affecting the human immune system and is caused by the human immunodeficiency virus or HIV. The HIV *pol* gene encodes both the HIV protease and HIV reverse transcriptase proteins which are the molecular targets for both the protease and reverse transcriptase inhibitors. Drug resistance in HIV often arises as a result of mutations in this *pol* gene, making it a target for frequent genotypic resistance testing. A large set of such genotypic sequence data with corresponding phenotypic results is available through the Stanford HIV Drug Resistance Database (HIVDB) (Rhee et al., 2003; Shafer, 2006). The viral gene sequences represent the genotype whereas the phenotype is measured as the mutant's effect on resistance or its resistance factor (RF). This resistance factor is defined as the fractional change in the mutant virus's IC50 compared to the wild type virus IC50. (For a description of IC50, see Chapter 1).

$$RF = \frac{\text{mutantIC50}}{\text{wildtypeIC50}}, \tag{2.1}$$

Thus, using this HIVDB data linking the HIV genotype with resistance phenotype, the sequence-based resistance prediction algorithms for HIV attempt to predict the logarithm of RF values for a given query viral sequence. A few of these works pertaining to various drug targets and drugs for HIV are discussed briefly.

From a set of 650 matched genotype-phenotype pairs, (Beerenwinkel et al., 2003) constructed regression models for the prediction of phenotypic drug resistance from the HIV genotype. For each queried mutant sequence, their model provides a probability of membership in the resistant population. Another work using regression models for HIV drug resistance prediction comes from (Rhee et al., 2006). The authors used five statistical learning methods including least-squares regression, least angles regression, decision trees, neural

networks and support vector machines to relate the HIV reverse transcriptase and HIV protease mutations to drug susceptibility for 16 anti-HIV drugs. The authors found the regression methods, in particular least angle regression, to be the best learning model for the prediction of drug resistance in HIV. A more recent study employing regression for resistance prediction comes from (van der Borght et al., 2011). The authors used both linear and cross-validated stepwise regression to predict resistance to various HIV reverse transcriptase inhibitors. The study also identified novel mutants associated with resistance to the non-nucleoside reverse transcriptase inhibitors (NNRTI) drug class. These novel mutants were further confirmed to be resistance conferring in the wet-lab by generation of site-directed mutants and by determining the in vitro resistance levels. (Pasomsub et al., 2010) used Artificial Neural networks (ANNs) to predict HIV resistance phenotype from genotypic data. The study used 7500 pairs of HIV sequences with corresponding phenotypic fold change values for 14 drugs to train and test their model. A set of amino acid mutations known to be associated with resistance to HIV inhibitors was used as input to the training as well. Their results were comparable to other interpretation systems such as geno2pheno (Beerenwinkel et al., 2003). Similarly, (Heider et al., 2010) also used ANNs and random forests for the prediction of bevirimat resistance in HIV. In addition, they identified target mutation positions which are resistance hotspots. Another HIV study for protease inhibitor lopinavir was conducted in (Bembom et al., 2009). The authors used targeted maximum likelihood estimation to determine mutants which are significant contributors to lopinavir resistance. The learning methods described here often report high accuracies of prediction ranging from 90-96 percent on their test sets, often coupled with reasonable true positive rates (where reported). However, it is worth noting that high accuracy alone can be misleading in this context and should not be used to assess the difficulty of the resistance prediction problem. As only a small number of possible resistance mutants exist, a naive classifier can assign all mutants to be drug-sensitive and still achieve high accuracy. Thus, a challenge for resistance prediction methods is often to maintain a high true positive rate while minimizing the number of false positives.

Despite good predictive abilities, often the learning methods for HIV resistance prediction described above cannot explain the reasons why a particular query mutant is resistance conferring or susceptible. For instance, it is often not clear what mutations are enough to cause resistance or what interactions occur between mutations to cause either resistance or susceptibility in the mutant. A study by (Zhang et al., 2010) attempted to explain these interactions in the HIV protease. Using Bayesian statistics and probabilistic modelling, they

designed a statistical procedure to detect mutations associated with drug resistance. More importantly, their model predicts interaction patterns between different single point mutants. For instance, the authors reported that under their model, mutations at positions 46 and 54 of the HIV protease are conditionally independent given the amino acid at position 82.

The methods described previously use the gene or protein sequence of HIV virus to train the learning models and, to subsequently predict resistance. However, another class of sequence-based methods performs proteochemometric modelling of the HIV protein sequences and inhibitors to predict resistance. In these methods, each gene or protein sequence is converted into a set of descriptors that describe the chemical properties of the protein. A similar process is repeated for the ligands i.e. the drugs and the interactions between the ligand and protein are inferred based on complementarity of the chemical features. The model is then trained to correlate these chemical descriptors with the susceptibility data. (Junaid et al., 2010) and (Lapins et al., 2008; Lapins and Wikberg, 2009) have applied such proteochemometric modelling to HIV reverse transcriptase and HIV proteases respectively. The salient feature of the model is that since both the protein and inhibitor are somewhat generalized using their chemical descriptors, a model trained for inhibitor A can potentially be used to make predictions for a chemically similar inhibitor B for which the model has no training. However, in a case where novel inhibitors are found and similarity between drugs is minimal, the model accuracy will decline.

In addition to these computational learning approaches, rule-based systems that rely on expert knowledge of resistance are available for well studied systems such as HIV. These rule based systems rely on clinical data and human expert knowledge about drug resistance for the disease in question to determine rules that estimate the influence of mutations on single drugs or on drug combinations. At the time of writing, five such systems are publicly available for HIV and include: Agence Nationale de Recherche sur le SIDA (ANRS) (Meynard et al., 2002), HIV RT and Protease Sequence Database (HIVDB) (Shafer et al., 1999), Rega Institute (Rega) (van Laethem et al., 2002), Visible Genetics (VGI)(Reid et al., 2002) and HIV-GRADE (Obermeier et al., 2012).

2.1.2 Computational Structure-Based Approaches

Unlike sequence based approaches of the previous section, structure-based approaches use structure of the drug target protein and probe the binding interactions between a drug and this target protein. These methods often use available protein-inhibitor complex structures,

docking programs, molecular modelling programs and molecular dynamics. In general, starting with a wild type structure, structure-based methods introduce mutations in the target protein using modelling software. This step is followed by energy minimization and molecular dynamics of the target structure to incorporate any structural changes introduced by the mutation; binding energies between the drug and mutant are then calculated via docking algorithms or other molecular mechanics methods. Finally, a scoring function is used to evaluate the resistance conferring ability of these mutant proteins. These scoring functions often use the decreased drug binding with the mutant protein as an indication of resistance conferring ability.

A number of studies fall under this category. For instance, (Chen et al., 2001) used docking algorithms to study resistance in HIV protease, HIV reverse transcriptase and enoyl ACP reductase from *Mycobacterium tuberculosis*. They introduced specific single and double point resistance mutations in these drug target proteins using SYBYL (a molecular modelling program). Receptor-ligand binding was then studied using docking algorithms. They note that in majority of the cases, the energy shift in the binding energy of the mutant-protein-ligand was consistent with clinically known resistance data, indicating that resistance mutations disrupt drug binding, whereas those not conferring resistance might actually improve drug binding. Similarly, using molecular dynamics (Hou and Yu, 2007) studied the effect of the V82F/I84V double mutant of HIV protease on the protease inhibitor amprenavir and two novel inhibitors. The authors found that the double mutant distorts the binding site geometry, thereby weakening drug interactions between the mutant protein and protease inhibitors. Another resistance study for enoyl ACP reductase in *Mycobacterium tuberculosis* comes from (Wahab et al., 2009). The authors studied the S94A mutation that occurs in enoyl ACP reductase active site and confers resistance to isoniazid in tuberculosis patients. Molecular docking and molecular dynamics simulations were used to study and compare isoniazid binding to the wild type and the S94A mutant protein. The authors reported that the S94A mutation makes the mutant protein more mobile, impairing drug binding and causing isoniazid resistance. (Zhu et al., 2009) studied the mechanism of resistance for four single point mutants which make grass weed populations resistant to ACCase herbicides. A multidrug resistant mutant of HIV-1 protease was studied using molecular dynamics and NMR relaxation by (Cai et al., 2012). Their comparative analyses from both method shows that the enzyme dynamics are affected by the mutant. The authors hypothesized that these alterations in enzyme dynamics likely modulate the balance between protease substrate turnover and drug binding, and hence confer drug resistance. Finally, more recently, (Mittal et al.,

2013) carried out structural and binding thermodynamics to investigate the effects of mutations at residue 50 of HIV-1 protease on binding to atazanavir and amprenavir. The authors report a decrease in binding entropy, which is compensated by enhanced enthalpy for atazanavir binding to I50V variants, and amprenavir binding to I50L variants leading to hyper susceptibility.

The structural studies mentioned in this section manipulate and analyze the structural interactions of a drug and mutant protein to infer resistance in a drug target. Thus, the biochemical and structural causes of resistance through particular mutations can be explained using these structural methods. Furthermore, unlike the sequence-based methods from the previous section, these structure-based studies do not rely on extensive training data for the systems being studied. However, docking and molecular modelling is performed on specific, predefined mutants, and are not inherently search methodologies. Thus, these structural studies have been limited in modelling and analyzing a select number of predefined mutations, and do not perform an extensive search for possible resistant mutants for a drug target.

2.2 Introduction to Computational Protein Design

Computational protein design methods are used to engineer novel biological function in a target protein, to change its thermostability, or to change its binding profile (Ambroggio and Kuhlman, 2006; Ashworth et al., 2006; Chakrabarti et al., 2005; Desjarlais and Handel, 1995; Dwyer et al., 2004; Hu et al., 2008; Kraemer-Pecore et al., 2003; Kuhlman et al., 2003; Looger et al., 2003b; Marvin and Hellinga, 2001; Offredi et al., 2003; Shimaoka et al., 2000; Slovic et al., 2004). The design procedure generally starts with an input template or protein backbone structure which has been stripped of its amino acid side-chains. A computational protein design (CPD) algorithm then aims to find the amino acid sequence(s) that can best adopt the target 3D fold/structure specified by the input template and posses the desired change in function or binding profile. The fitness of a sequence is evaluated by the energy associated with it as it adopts the target structure. The goal of a CPD algorithm is then to search for the optimal solution or the Global Minimum Energy Conformation (GMEC). (For the rest of this section, the terms energy and fitness are used somewhat interchangeably.)

The search for the GMEC in the protein sequence-structure space is a daunting task which has been shown to be NP-hard (Pierce and Winfree, 2002). For a target protein of size n and the 20 naturally occurring amino acids, there are 20^n distinct amino acid sequence choices. Furthermore, in the structure of naturally occurring proteins, the amino acids side

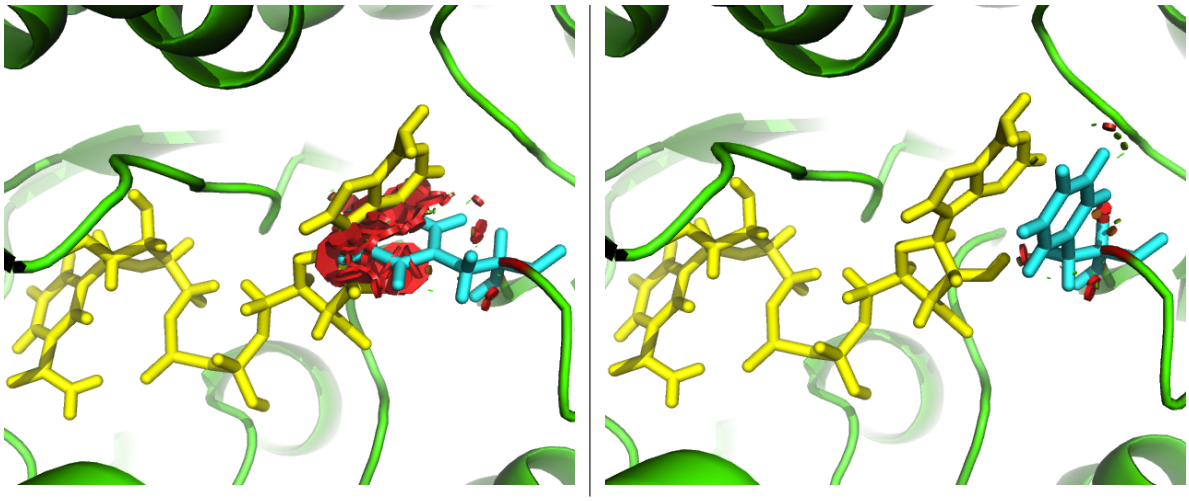

Figure 2.1: **Rotamers** Two rotameric conformations of phenylalanine side chain are displayed (blue). Left: Selection of the phenylalanine rotamer during protein redesign causes a steric clash with the ligand (yellow). Right: An alternate rotamer of phenylalanine avoids the steric clash with the ligand.

chains can adopt an orientation which is sampled from a continuous space. Therefore, each sequence corresponds to an infinite and continuous structure space. However, to approximate this behavior, rotamer libraries (Lovell et al., 2000a; Ponder and Richards, 1987) are used in protein design which discretize this continuous space by sampling the most probable side chain orientations or conformations for each amino acid (rotamer) (see Figure 2.1). Thus a degree of structural flexibility is incorporated in the protein design experiments, and each amino acid sequence corresponds to a finite, yet combinatorially large, discretized structural space. After the inclusion of rotamer libraries of size m (where m represents total number of rotamers for all amino acids), the total number of possible sequence-structure solutions or conformations in the search space becomes m^n which is prohibitively large. It is also worth northing that naturally occurring proteins are relatively long, containing on average 200 amino acids. Therefore, for many redesign experiments, only a small, specific part of the target protein is selected for redesign, such as a hydrophobic core containing 30 residues or an active site containing 10 residues (Looger et al., 2003b). However, even for modest design goals as these, the search space size prevents exhaustive enumeration, since for a rotamer library of size $m = 100$, there are 100^{30} and 100^{10} possible solutions for the aforementioned cases.

It is evident that the search space for even a modest protein design experiment is large enough to prevent exhaustive enumeration and evaluation of all solutions. Therefore, various

search and optimization methodologies are used for efficient solution of the protein design problem. These methods can be broadly categorized into stochastic and deterministic strategies. The former includes genetic algorithms (GA) and Simulated Annealing Monte Carlo (SA) methods, whereas the latter is composed of the Dead-End Elimination (DEE) based search strategies. The rest of this section details previous work in each of these categories.

2.2.1 Stochastic Computational Protein Design

The stochastic strategies include genetic algorithms (Holland, 1975; Holland, 1992; Fogel, 1998) and simulated annealing (Kirkpatrick et al., 1983; Cerny, 1985; Thomas et al., 2009). Brief details of both methods are described below.

Genetic Algorithms

In simple terms, genetic algorithms are search heuristics that draw on the concepts of natural selection and evolution to model optimization and search problems. A genetic algorithm is initialized with a parent generation. Each member of a generation is called a genome and encodes a valid solution. (It is worth noting that although the term genome seems to imply a sequence, the genomes are just possible solutions in a domain and often represent a candidate protein structure in protein design.) The parent generation is then evolved towards better solutions (as per a fitness or energy function) via the processes of mutation and crossover. The next generation of genomes then undergoes selection and forms the subsequent parent generation. This process continues until either the maximum number of generations specified has been examined, or the current population has members with an acceptable level of fitness, or the process converges to a single solution. The operators of mutation and crossover are predominantly used to create new generations by introducing small changes to or by randomly combining genomes from the parent generation. Each of these operations is described below:

Mutation: Mutations are small changes in the parents to ensure genetic diversity during the algorithm run. The classic example of a mutation operator involves a probability that an arbitrary bit in a parent genome will be changed from its original state. The mutation operator is analogous to the biological mutations.

CrossOver: New solutions that differ significantly from the parents can be generated by combining the parent genomes by cross over. Single or multiple cross over points are chosen. (The crossover points can be chosen randomly or by some pre-decided strategy.)

The crossover operator then generates genomes containing distinct parts that entirely belong to one parent or the other.

After new generations have been created by mutation and crossover, each new generation undergoes the process of selection which decides the genomes that will comprise the next parent generation for further breeding/evolution. Selection can be performed in many different ways. For instance, after the fitness function has been evaluated for each individual genome, the best genome out of a randomly selected subset can be chosen each time until the desired generation size is obtained (tournament selection).

Genetic algorithms can be easily extended to protein design and have thus been applied to protein design in various studies (Desjarlais and Handel, 1995; Jones, 1994; Tuffery et al., 1991; Voigt et al., 2000a; Xia and Levitt, 2004; Yang and Liu, 2006; Thomas et al., 2009). Each position in a genome can adopt m distinct values where m is generally the rotamer library size. Thus, each genome is a rotamer assignment for the target structure. The fitness of each solution can be calculated via an energy function which indicates the likelihood that the current rotameric and amino acid assignment folds into the target 3D structure. In the past, studies have also experimented with various simple fitness functions such as those only considering hydrophobicity or those only concerned with avoiding steric clashes in the target protein structure. Use of such simplistic fitness functions eases the computation but often presents mixed results such as suggested solutions where the amino acid composition does not conform to the naturally occurring proteins at all, instead predominantly using one amino acid such as Histidine. Use of sophisticated fitness functions, such as those better approximating the free energy of the target structure (by including terms like dihedrals, electro statics, vdW etc) will likely see better results.

Monte-Carlo Simulated Annealing

In the simulated annealing approach, random walks are created in the search space of interest. These walks are generated by creating successive candidate solutions by making small changes to the current candidate solution. The changes are accepted using an acceptance criterion such as the Metropolis Criterion (MC). Using MC, if the changes optimize the current solution i.e. if the energy of the new solution is lower (or fitness is higher) then the new solution is accepted. In order to prevent the algorithm from getting stuck in bad local minima prematurely, some potentially negative changes are also accepted where the new candidate solution is higher in energy (lower in fitness) as compared to the current solution. The simulated annealing process borrows heavily from statistical mechanics theory.

The process is described below in relation to its application to protein design (Godzik, 1995; Hayes et al., 2002; Hellinga and Richards, 1994; Jiang et al., 2000; Kuhlman and Baker, USA; Nilges and Brunger, 1991; Shakhnovic and Gutin, 1998). The redesign process begins from a random sequence. The energy of this sequence E_i (fitted into the target structure) is calculated. Random mutations (amino acid changes or just rotameric changes) are made to this sequence to generate a new candidate solution. The energy E_{i+1} of the new solution is then computed. If the difference of the energies $\Delta E = E_{i+1} - E_i$ is negative (indicating the new solution has a lower energy than the current one and hence better fitness), the mutations are accepted. On the other hand, if the energy difference is positive, indicating the new solution has a higher energy than the current solution, the new solution is accepted by evaluating the metropolis criterion (MC). The MC for any conformation c is evaluated according to the Boltzmann distribution and is accepted if and only if:

$$p < e^{-\Delta E_i / \kappa_\beta T_n} \tag{2.2}$$

where p is the acceptance probability, which is a random number between 0 and 1, T_n is the annealing temperature and κ_β is the Boltzmann constant. As can be seen from the MC, at higher temperatures, the probability that unfavorable changes made to a solution will be accepted is higher. This allows the algorithm to escape from local minima as the system is heated, by sampling a larger area of the search space. As the system cools down, the probability that unfavorable changes will be accepted becomes smaller, allowing the system to converge. The search procedure starts with the system at a higher annealing temperature which is slowly cooled down. At each temperature, a number of trials are carried out until the system equilibrates or converges.

Both simulated annealing and genetic algorithms are non-deterministic methods and selectively sample the search space. Thus, these stochastic approaches are often computationally efficient and can rapidly provide feasible solutions. On the other hand, due to selective sampling of the search space, these approaches have the potential to get stuck in local minima. Neither approach guarantees to find the minimum or best solution (the GMEC), and instead terminates when a good enough solution has been found or the maximum number of allowed iterations has expired. Like genetic algorithms, the results of simulated annealing Monte-Carlo methods are somewhat dependent on the starting conditions as well as the system parameters (mutation rate , crossover points, max. number of generations, annealing temperature etc). Depending on the target protein, these system parameters are determined

either randomly or by previous knowledge of the system. Often multiple runs of GA and SA are required to arrive at a set of acceptable solutions.

2.2.2 Deterministic Protein Design (Dead-End Elimination)

Both genetic algorithms and Monte-Carlo based approaches sample a subset of the allowed conformational space. Therefore, neither approach guarantees the GMEC as the final solution. Dead-End elimination (DEE) (Desmet et al., 1992) is a family of pruning techniques applied to the protein design problem which guarantee that the GMEC is not pruned. Combination of DEE along with best-first searches like A^* guarantees that the GMEC of the search space is returned.

Given a pairwise energy function, the total energy E_c of a conformation c can be specified as :

$$E_c = E_t + \sum_i E(i_r) + \sum_i \sum_{j>i} E(i_r, j_s) \tag{2.3}$$

This equation follows the notation established by the DEE community: E_t is the template self energy (i.e., the self energy of the backbone and other rigid parts of the protein), i_r denotes the presence of rotamer r at position i, $E(i_r)$ is the self energy for rotamer i_r (i.e., the sum of intra-residue and the residue-to-template energies), and $E(i_r; j_s)$ is the pairwise energy between rotamers i_r and j_s. The traditional DEE criterion of (Desmet et al., 1992) prunes a rotamer i_r if a second rotamer i_t is found such that the lowest (best) energy among the conformations including i_r is greater than the highest (worst) energy among all the conformations containing i_t (see Figure 2.2A). Intuitively, a satisfied DEE criterion implies that the energy of any conformation involving i_r can be improved by exchanging it for i_t; therefore, i_r can not be part of the GMEC. The traditional DEE pruning criterion is written as:

$$E(i_r) + \sum_j \min_s E(i_r, j_s) > E(i_t) + \sum_j \max_s E(i_t, j_s) + \Delta \tag{2.4}$$

where s is selected from the set of allowed rotamers at position j.

Similar to the genetic algorithms and simulated annealing approaches, DEE based algorithms search for the GMEC as defined by an energy function. State-of-the-art energy functions aim to approximate the experimental interaction energies. These approximations often come close to, but do not perfectly, predict the true interaction energies. Hence, in the traditional DEE criterion, a Δ term allows a window of low energy conformations to

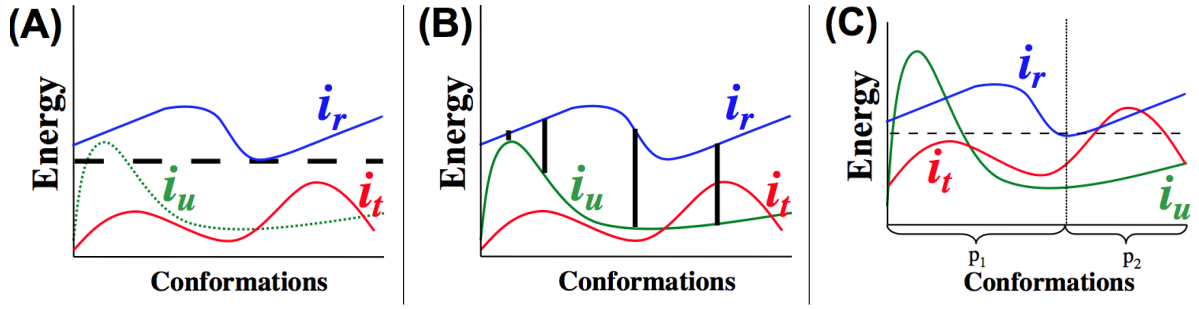

Figure 2.2: **Pruning by DEE** The abscissa represents all possible conformations of the protein (excluding residue i). In other words, each value on the abscissa corresponds to a partial protein conformation where a side-chain conformation is assigned for every position $j(j >= i)$. The curve for rotamer i_x represents the total energy of complete conformations containing rotamer x at position i and the remainder of the protein conformation dened by the position along the abscissa. (A) Using traditional DEE criterion, rotamer i_t can prune rotamer i_r because the best energy (dashed line) among the conformations including i_r is greater than the worst energy among all conformations containing it. Rotamer i_u is not able to prune i_r. (B) Goldstein DEE: Energy of conformations containing i_t is always lower than those containing i_r, hence i_t prunes i_r. (C) Split DEE: i_t eliminates i_r in first partition, whereas i_u eliminates it in the second. (This figure is simply illustrative; if one were actually to plot these energies, the curves would likely not be as smooth as presented here.)

be identified. When $\Delta > 0$, DEE will only prune rotamers that are not part of any conformation with energy within Δ of the GMEC. A number of these conformations would, then, be taken to the wet-lab for synthesis and testing. The conservative pruning condition of traditional DEE is often difficult to satisfy and hence provides limited pruning ability. However, a number of DEE extensions greatly enhance the methods power. Two of the more common extensions are the Goldstein (Goldstein, 1994) and split (Pierce et al., 2000) criteria. The Goldstein criterion results from an algebraic manipulation of traditional DEE criterion. Instead of requiring that the worst energy contribution of i_t be greater than the best energy contributions of i_r, Goldstein criterion requires that as long as lower energy can always be achieved by substituting i_t for i_r, i_r can be pruned. This translates to the following equation:

$$E(i_r) - E(i_t) + \sum_j \min_s (E(i_r, j_s) - E(i_t, j_s)) > 0 \tag{2.5}$$

The split-DEE criterion (Pierce et al., 2000) divides the conformational space into multiple partitions (see Figure 2.2C) and allows a different rotamer to prune i_r in each partition. The split-DEE criterion is satisfied if there is always some rotamer at position i which is energetically favorable to i_r. Therefore, split-DEE allows a set of alternate rotamers to prune

i_r. In the single split (or $s = 1$) criterion, we identify a single split position (or residue) p with q remaining (unpruned) rotamers. The method splits conformational space into q partitions, where in partition x, the rotamer at position p is set to $p_x(x \leq q)$. The Goldstein or traditional DEE criterion is then examined in each partition. If i_r is pruned in each partition, then i_r can be pruned from the entire conformational space. It is also possible to split on multiple residues (i.e., $s = 2$) and exchange a further expansion of conformation space for improved pruning ability. Splits beyond $s = 2$ have significant bookkeeping and implementation overhead and are rarely employed. Beyond the traditional, Goldstein, and split criteria, several powerful and provably correct extensions have been added to the DEE family of algorithms. These methods include the ability to handle rotamer pairs (doubles (Gordon and Mayo, 1998; Lasters and Desmet, 1993)) instead of single rotamers, the ability to allow rotamers with energy minimization (MinDEE (Georgiev et al., 2006b; Georgiev et al., 2008b)), limited backbone flexibility (BD (Georgiev and Donald, 2007)) and backbone backrub motions (BRDEE (Georgiev et al., 2008a)).

When pruning with DEE, dead-ending rotamers are identified by comparing all remaining rotamers against the specified DEE criterion. This process of evaluating the DEE criteria for all rotamers is defined as a DEE cycle. Highly efficient DEE cycles can be created by sequential execution of multiple different DEE criteria. For example, a single DEE cycle might consist of first applying the Goldstein criterion followed by a split-DEE criterion. The general contraction of conformation space in earlier cycles allows some rotamers to be pruned in later cycles, even if they were not initially identified as dead-ending in earlier cycles.

It is worth noting that DEE is a pruning criterion that only guarantees that no rotamers belonging to the GMEC will be pruned, however it does not guarantee that repeated DEE cycles will prune all but the GMEC. Therefore, a redesign algorithm will run multiple DEE cycles until either no more rotamers can be pruned or the GMEC has been identified (i.e., only one rotamer remains at each position). When multiple rotamers remain, the DEE cycles are followed by an enumeration stage generally based on the A^* branch-and-bound search algorithm. In (Leach and Lemon, 1998) an A^* best-first search is used to expand a conformation tree, such that the conformations are returned in order of increasing energy (i.e., best-first). Hence, the first conformation returned by the A^* search is the GMEC.

If DEE was performed with $\Delta > 0$, conformations can be repeatedly enumerated by the A^* search until one is identified with an energy Δ above the GMEC. Conformation enumeration should stop at this point because there is no guarantee on the presence and proper ordering of conformations with energies which are Δ above the GMEC. The A^*

search eliminates the need to examine all remaining conformations and typically results in a combinatorial factor reduction in the search space. Enhanced efficient versions of A^* that work in conjunction with split-DEE (Georgiev et al., 2006b) have also been reported.

Stochastic and approximate methods such as genetic algorithms and Monte-Carlo simulations have been mostly replaced by the dead-end elimination based strategies since the latter are guaranteed to find the GMEC. It has previously been reported that unlike DEE, GA and SA based approaches almost always fail to find the GMEC (which is the best energy solution as defined by an energy function) (Voigt et al., 2000b). The emergence of stronger DEE criteria, coupled with faster enumeration stages have enabled DEE based computational designs of large proteins. For instance, using DEE, (Offredi et al., 2003) performed a successful design of a 216 amino acid long alpha/beta-barrel protein. The computational solution was tested in the wet-lab and exhibited a stable three-dimensional structure in solution.

2.3 Free Energy Calculations and Ligand Binding

The ability of proteins to bind with other proteins as well as small molecules in a highly specific manner is an important aspect of biological processes. In addition, specific binding of drugs to the drug target proteins is a crucial aspect of drug efficacy as well. Thus, characterization and calculation of energies involved in this ligand binding process is an important objective in computational biology and computational chemistry. A number of computational approaches have been developed towards this objective ranging from empirical or statistical based methods to those aimed at evaluating the actual biophysical energies involved in binding. Among the most time efficient methods for estimation of binding energies are the knowledge-based statistical or empirical approaches (Jain, 1996; Eldridge et al., 1997; Gohlke et al., 2000). A number of simplifying assumptions such as lack of explicit solvent as well as absence of conformational sampling contribute towards their time efficiency at the cost of accuracy. On the other hand, more rigorous and biophysically accurate methods of calculation tend to be more time consuming and are based on molecular force fields. These methods often involve gradual transformations between the states of interest using Monte-Carlo (MC) or Molecular Dynamics (MD) simulations.

2.3.1 Free Energy Perturbation and Thermodynamic Integration

Free energy calculations are generally formulated in terms of estimating relative free energies between two states A and B. In context of ligand binding, these states can represent the unbound protein and ligand and the bound complex. The free energy difference between two states can be obtained by using the Zwanzig equation (Zwanzig, 1954).

$$\Delta G = G_B - G_A = -\beta^{-}1 ln\langle exp(-\beta \Delta V)\rangle_A \tag{2.6}$$

where $\beta = 1/k_B T$, T is the temperature, k_B is the Boltzmann constant, and $\langle\rangle_A$ represents the ensemble average over state A, obtained by conformational sampling via MC or MD simulations. Free energy calculations converge when the difference between energies of two states A and B is small. When the energy differences are larger, the transformation between the two thermodynamic states A and B is replaced by a series of transformations between non-physical, intermediate states along a well-delineated pathway that connects A to B. This pathway is characterized by the general extent parameter λ, making the free energy a continuous function of λ between A and B. The total free energy can then be obtained by summing over these intermediate states along the λ variable.

$$\Delta G = G_B - G_A = -\beta^{-}1 \sum_{m=1}^{n-1} ln\langle exp(-\beta(V_{m+1} - V_m))\rangle_m \tag{2.7}$$

where m stands for the number of intermediate states used. This approach is generally referred to as the free energy perturbation (FEP) method.

An alternative to free energy perturbation method is the thermodynamic integration (TI) approach (Frenkel and Smit, 2001). The free energy in TI is given by:

$$\Delta G = \int_0^1 \langle \frac{\partial V(\lambda)}{\partial \lambda} \rangle_\lambda \partial \lambda \tag{2.8}$$

Equation 2.7 provides an exact formula for free energy calculations and, in principle, if the simulations are allowed to run long enough and adequate sampling is performed, FEP can provide highly accurate results. However, in practice, performing extensive sampling and simulations might not be feasible, and convergence of FEP can become an issue. On the other hand, TI avoids this problem. However, Equation 2.8 is essentially an integral that must in practice be computed by discretizing and computing the integrand at finite intervals,

thus affecting accuracy. Despite these possible pitfalls, both FEP and TI can provide highly accurate results at the cost of efficiency.

2.3.2 QM/MM

Computational approaches based on QM/MM combine the accuracy of quantum mechanics (QM) with relative efficiency of molecular mechanics (MM), and are employed to study chemical reactions including ligand binding (Warshel, 1976). These QM/MM based approaches are quite useful in describing phenomenon such as charge polarizations and electron transfer, which is not possible by the classical force field based methods. However, the accuracy in quantum mechanics is obtained at the cost of efficiency. A force field or MM based approach generally scales quadratically with the number of atoms N. This complexity is often further reduced by employing cutoff radii as well as by use of efficient particle mesh Ewald (PME) to calculate electrostatics (Darden et al., 1993). On the other hand, the computational time of quantum mechanics approaches generally ranges from N^3 to N^5 (Van der Vaart et al., 2000). Thus, in a QM/MM simulation only a small part of the system of interest, e.g. the binding site or the ligand, is treated using quantum mechanics, whereas the rest of the system is treated using classic molecular mechanics. A number of previous studies have employed QM/MM based protocols to investigate ligand binding and activity and report promising results. (Hayik et al., 2010) developed a QM/MM based approach to study binding energies in 23 metalloprotein-ligand complexes. More recently, (Rathore et al., 2013) applied and validated a QM/MM based method to a structurally diverse set of fructose 1,6- bisphosphatase (FBPase) inhibitors.

2.3.3 MM-PBSA and MM-GBSA

The Molecular Mechanics- Poisson Boltzmann Surface Area or MM-PBSA approach is based on analysis of MD trajectories using a continuum solvent approach (Srinivasan et al., 1998; Kollman et al., 2000). The free energy of a state by MM-PBSA is approximated as:

$$\langle G \rangle = \langle E_{MM} \rangle + \langle G_{PBSA} \rangle - T\langle S_{MM} \rangle \tag{2.9}$$

where $\langle E_{MM} \rangle$ is the average molecular mechanical energy including bond, angle, torsion, electrostatic and van der Waals terms as described by a force field. $\langle G_{PBSA} \rangle$ represents the solvation free energy as well as a surface area based estimate of the nonpolar free energy. As

the name indicates, the solvation free energies are calculated using the Poisson-Boltzmann equation. (Alternately, for MM-GBSA, the solvation free energy is obtained by using the generalized Born (GB) approximation to the Poisson Boltzmann equation). Both $\langle E_{MM} \rangle$ and $\langle G_{PBSA} \rangle$ are obtained by averaging over an MD trajectory of the system. Finally, $T\langle S_{MM} \rangle$ represents the solute entropy and can be calculated by using normal mode analysis (Srinivasan et al., 1998).

For binding energy calculations, Equation 2.9 can be evaluated independently for the ligand, receptor and complex; binding energy can subsequently be determined by:

$$\Delta G_{bind} = \langle G_{complex} \rangle - \langle G_{receptor} \rangle - \langle G_{ligand} \rangle \tag{2.10}$$

Alternately, each of the terms in Equation 2.10 can be estimated by using snapshots from the MD trajectory of the complex alone. In this scenario, the energy terms for the receptor and ligand alone are estimated by removal of one binding partner from the trajectory (ligand and receptor respectively). Consequently, calculation of binding energies using MM-PBSA using the complex trajectory alone assumes that the structure of the receptor as well as ligand does not change upon binding.

A number of factors influence the choice of a methodology for calculating binding energies including computational efficiency or running time and accuracy tradeoff. The quantum mechanics or free energy perturbation based models, in principle, avail high accuracies. However, significant computational resources are needed to perform these calculations, making these methods unsuitable for studies where large number of protein-ligand complexes need to be screened. Alternately, when a reasonable structure of the ligand and receptor is known, it is possible to limit the running time while offering modest declines in accuracy. A number of previous studies have employed MM-PBSA or MM-GBSA based schemes to evaluate binding energies in a time efficient manner (Ferrari et al., 2007; Guimaraes and Cardozo, 2008; Raju et al., 2010). However, it is worth noting that these efficient schemes make a number of simplifying assumptions and may not account for energy changes occurring from large scale conformational changes as well as large entropic contributions. Finally, none of these approaches are able to predict binding energies in absolute quantitative agreement with the experimental binding energies (Ferrari et al., 2007).

2.4 Computational Models of Molecular Evolution

Evolution is the process by which hereditary characteristics of organisms change over generations. In general, the term evolution suggests a change in visible or phenotypic characters. Examples of such evolution would be changes in human height, skin colour etc. However, whether or not evolution manifests itself at the phenotypic level, it is primarily a change in the underlying DNA of an organism by introduction of mutations. In this thesis, the term evolution specifically refers to the evolution of a single gene or its associated protein over time. Three concepts underlie the process of evolution: mutation, selection and fixation.

As genomes replicate and new cells are produced, mutations are seen in the resulting genomes. These mutations can potentially be one of three kinds: deleterious, neutral and beneficial. The acquisition of deleterious mutations makes an organism less *fit* than its parent, whereas the acquisition of beneficial mutations grants a benefit to the organism and increases its *fitness*. Unlike deleterious and beneficial mutations, the neutral mutations are those which have little to no effect on the fitness of an organism i.e. these mutations maintain the status quo. Thus, these neutral mutations can provide genetic variation in a population. The definition of whether a newly appeared mutant is deleterious, beneficial or neutral is largely dependent on the environment or external circumstances which the genome is subjected to. As this environment changes, formerly beneficial mutations might become deleterious or vice versa. The environment exerts a *selective* pressure on the genome that defines what mutations are beneficial, deleterious or neutral. In this context, the beneficial mutations confer a higher reproductive ability that allows these mutants to survive while most deleterious mutants are nearly eliminated. The mutants that survive in the offspring population are thus known to be *selected* by nature or by *natural selection*.

Consider the following simple synthetic example of evolution in a protein (see Figure 2.3). Two mutations m_1 and m_2 arise in the wild type protein, leading to three mutant proteins: first containing single mutation m_1, second containing single mutation m_2 and the third double mutant containing both m_1 and m_2. Also consider that the first mutant misfolds and has zero functional activity; the second mutant is more stable and shows higher functional activity than the wild type and finally, the third mutant has slightly decreased stability and functional activity close to the wild type. If protein function is assumed to affect fitness, in this example, m_1 is a highly deleterious mutation since the protein completely loses its function. On the other hand, m_2 is beneficial mutation since the second mutant has a higher functional activity compared to the wild type. And finally, the double mutant is nearly

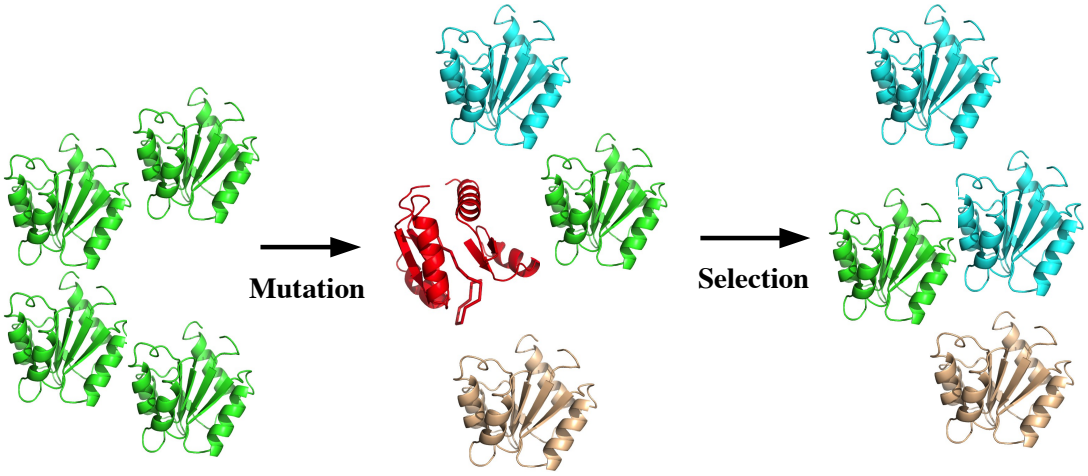

Figure 2.3: **Evolution In A Protein** Mutations in wild type protein (green) give rise to mutant types. First mutant mis-folds and is deleterious (red). Second mutant is slightly beneficial (cyan), whereas the third mutant is nearly neutral (wheat). Selection favours the beneficial mutant, increasing its frequency in the final evolved population.

neutral since it maintains an activity level compared to the wild type protein. Thus, over the course of evolution, the first highly deleterious mutant will be selected against and vanish from the population; whereas the second mutant will be favoured in a population. It is worth noting that while protein function might not always be a good indicator of fitness, in case of essential proteins, one can reasonably expect function to impair fitness.

2.4.1 Simulating Molecular Evolution

Evolution is often simulated as a computational process governed by laws of evolutionary theory. The Wright-Fischer model is a well-established evolutionary model (Hartl and Clark, 2007; McCandlish, 2011), where generations are assumed to be non overlapping. Thus, in the context of a Wright-Fisher process, evolution can be modelled as a Markov process where the current state of the population is only dependent on its previous state. As time goes on, mutations in the wild type protein sequence produce new mutant proteins. These evolutionary simulations are modelled to have the Markov property i.e. they are memory less. In the context of an evolutionary simulation, this entails that the probability that a mutation *mut* leads it from state i to state j is independent of the sequence of mutations that led from an initial state to state i.

Evolution as a Markov Process A Markov process is a stochastic process that satisfies

the Markov property. Having the Markov property implies that the future state of the process can be described completely based on its present state, without the knowledge of previous states. In this way, a Markov process is memoryless. A Markov process can be regarded as a directed graph where each vertex represents a *state* of the process. The edges between two states i and j have weights associated with them. In a continuous time Markov process, these weights correspond to the instantaneous *transition rates* q_{ij} between states (Norris, 1997; Yang, 2006). Thus, such a Markov process can be represented by its rate transition matrix Q of size $n \times n$ where n is the number of states. Each entry of the rate matrix represents a transition rate between two states. Finally, in a continuous time Markov process, the probability of transition between two states can be found using the probability transitions matrix $P(t)$ given by:

$$P(t) = e^{-Q\Delta t} \tag{2.11}$$

As mentioned earlier, evolution can be considered a Markov process (Yang, 2006). Thus to simulate evolution, we need to define the rate and probability transition matrices. The rate matrices for evolutionary simulations are often derived from models of nucleotide substitution. These substitution models define the rate of mutation between different nucleotides at a single DNA location.

Models of Nucleotide Substitution

During gene evolution, *mutation* samples new gene sequences. Various nucleotide substitution models that capture this process of mutation in evolution are used in evolutionary simulations. Since DNA is composed of four nucleotide bases: A, T, G and C, the mutational space for a single DNA site or *locus* can be represented as four discrete states. The substitution models represent the rate of mutation or the rate of change of one nucleotide base to the other. However, these models do not explicitly account for natural selection, which has to be accounted for separately in a simulation. These models of DNA sequence evolution are briefly discussed below:

Jukes and Cantor Models of DNA sequence evolution were first described by Jukes and Cantor in 1969 (Jukes and Cantor, 1969). In this model, the rate of substitution from one nucleotide to another is the same for all nucleotides. Thus, if we define q_{ij} to be the rate of substitution/mutation from nucleotide i to j, the rate matrix for an evolutionary Markov

chain described by the Jukes Cantor model is given by:

$$Q = \{q_{ij}\} = \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix}$$

where the nucleotides are ordered as A, T, G, C. Since Q represents a rate matrix for a continuous time Markov Chain, each row the matrix sums to 0. Intuitively, the product $q_{ij}\delta t$ gives the probability that the nucleotide i will change into nucleotide j given infinitesimally small time step δt. Under Jukes Cantor model, as time goes to infinity, the equilibrium frequencies for all substitutions are equal to 1/4, indicating that since rates of substitutions are the same, as times goes on, the identity of a nucleotide at position x for an evolving DNA sequence is random.

Kimura 1980 Unlike Jukes-Cantor, the substitution model proposed by Kimura (Kimura, 1980) makes a distinction in the substitution rates between different nucleotide bases. Two types of differences are accounted for: transitions and transversions. A substitution is called a transition if it occurs between bases which are chemically similar i.e. both purines (A or G) or both pyrimidines (C or T). A transversion is said to occur when the substitution changes a pyrimidine into a purine or vice versa. The rate matrix Q is given by:

$$Q = \{q_{ij}\} = \begin{pmatrix} -2\beta - \alpha & \beta & \alpha & \beta \\ \beta & -2\beta - \alpha & \beta & \alpha \\ \alpha & \beta & -2\beta - \alpha & \beta \\ \beta & \alpha & \beta & -2\beta - \alpha \end{pmatrix}$$

Felsenstein 1981 (Felsenstein, 1981) extends the models presented by (Jukes and Cantor, 1969) and (Kimura, 1980) by allowing distinct substitution rates between all base pairs. The rate matrix is given by:

$$Q = \{q_{ij}\} = \begin{pmatrix} -\pi_G - \pi_T - \pi_C & \pi_T & \pi_G & \pi_C \\ \pi_A & -\pi_A - \pi_G - \pi_C & \pi_G & \pi_C \\ \pi_A & \pi_T & -\pi_A - \pi_T - \pi_C & \pi_C \\ \pi_A & \pi_T & \pi_G & -\pi_A - \pi_T - \pi_G \end{pmatrix}$$

In addition to the foundational substitution models described here, a number of sophis-

ticated and organism and application specific rate matrices have been developed which are beyond the scope of this work (Hasegawa et al., 1985; Tamura and Nei, 1993; Tavare, 1986; Yang, 1994a; Yang, 1994b).

Evolution Models Incorporating Structure And Function The substitution models described in the previous section operate at the gene or DNA sequence level and assume site independence such that a mutation at one position is independent of the rate of mutation at another positions. However, the coding sequences of genes have to be translated into protein sequence which then fold into a specific structure to perform a specific function which is what often defines fitness in a cell. This process indicates that there are subtleties in the underlying biological process that are not nearly well captured by the substitution models described above. First, due to the redundancy in the genetic code, multiple codons (nucleotide triplets) translate to the same amino acid. Thus, many distinct DNA sequences will produce identical protein sequences. Thus, it stands to reason that the synonymous and non synonymous rates of substitutions in genes are different, since their effects on proteins and subsequent fitness are likely to be different. In addition, different structural locations in a protein differ in how tolerant they are to mutations. For instance, it is often that the surface of a protein tolerates mutations well, whereas most mutations in the hydrophobic core of the protein are deleterious due to large impacts on the folding and structure of the resulting protein. A number of substitution models that take such differences in structure into account have also been developed. In general, these models map the positions in protein sequences to a phenotypic property such as thermostability, contribution to protein function or protein structure etc, and the substitution rates are determined as a function of the modelled property. For instance, (Parisi and Echave, 2001) developed a substitution model where mutant protein sequences are selected against departure from a reference structure. This implies that in their model, any mutation that causes a structural breakdown or change compared to a reference structure is selected against. Despite progress in this field, general models of substitution that take into account the target protein structure, function etc are approximations that are far from perfect. Finally, even though the computational protein design algorithms described in previous sections do not define explicit substitution models, they also address the structural aspect of sequence evolution by searching for protein sequences which are compatible with a target structure. For a detailed review of evolutionary models and their structural aspects, see (Liberles et al., 2012).

Evolution In A Population The previous sections treat evolution as a single trajectory where new sequences or proteins are generated from old ones and selection acts as an agent

that either eliminates or selects for the new mutant. In reality, a gene or protein is a single entity in a population whereas selection acts at the protein/gene population level. The parameters of this population affect how selection and evolution manifest in a protein. As an example, one can consider two extreme populations: one consisting of two individuals and another one consisting of 1 billion individuals. The chance that a beneficial mutation will be selected for and appear in every individual of the first population is much higher than that for the second population. Thus, this underlying population size is an important parameter for evolution and is known as the effective population size N_e. This effective population size is the size of an idealized sample population that harbours the same genetic diversity as an actual population. Thus, the N_e individuals from an idealized population are a good representation of the actual population. As time goes on, some of the mutants selected by evolution will dominate others in the population. A mutation is said to reach fixation in a population if it sweeps through the population i.e. no other genetic variant remains in the population. The probability of such an event is known as the probability of fixation and is given by the following formula: (Fisher, 1930; Kimura, 1962; Wright, 1931):

$$p_{fix}(a \to b) = \frac{1 - e^{-2s_{a\to b}}}{1 - e^{-2N_e s_{a\to b}}} \tag{2.12}$$

where $s_{a\to b}$ represents the selective advantage (or disadvantage) of mutant b over mutant a. The formula assumes that a mutation reaches fixation before another mutation arises. Evolution under different population conditions will have different outcomes.

For a comprehensive review of evolutionary models of nucleotide substitution, physical models for evolution and influence of populations on evolution, see (Liberles et al., 2012).

Chapter 3

Restricted Dead-End Elimination

As previously described in Chapter 1, this thesis deals with resistance as a special protein design problem. The objective of this resistance redesign problem can be specified as the acquisition of decreased affinity for the drug by the mutant protein while maintaining its affinity for the native substrate. To this effect, this chapter focuses on the development of an efficient and provably correct protein redesign algorithm. The contents of this chapter have been previously published as (Safi and Lilien, 2010).

3.1 Introduction

Novel biological function can be engineered into a target protein with the use of algorithms for protein redesign. The task of computational protein redesign generally starts with a template structure, an initial sequence, and a set of assumptions. In this context, the protein redesign problem differs from de novo protein design where no initial sequence is provided and the goal is to determine a sequence that folds into the target structure. On the other hand, in protein redesign, changes are introduced in the known initial sequence of a protein to achieve desired results (novel function, improved thermostability etc). Most redesigns assume that a large subset of the protein is rigid, including the backbone and residues located sufficiently far away from the region of interest (typically the protein's active site). The side-chains of the non-rigid residues are considered *flexible*; they are allowed to change conformation during the redesign process, typically switching among a discrete set of low-energy conformations (*i.e.*, rotamers) (Lovell et al., 2000a). Flexible residues can be *immutable*, in which case they are restricted to assume rotamers from the wildtype amino acid, or they can be modeled as *mutable* in which case they are allowed to assume rotamers from a number of different amino

acid types. The goal of the redesign algorithm is to identify the protein sequence containing the Global Minimum Energy Conformation (GMEC). Under this set of conditions, protein design is NP-Hard (Chazelle et al., 2004; Pierce and Winfree, 2002).

Existing protein design strategies can be broadly categorized into deterministic and non-deterministic classes. The non-deterministic / heuristic approaches of random sampling, neural networks, and genetic algorithms are unable to make guarantees on the quality of their computational search. These heuristic methods have all but yielded to a family of deterministic techniques dominated by the pruning algorithms of Dead-End Elimination (DEE) (Desmet et al., 1992; Georgiev et al., 2006b; Goldstein, 1994; Gordon and Mayo, 1998; Lasters and Desmet, 1993; Pierce et al., 2000). DEE-based approaches are generally fast, are guaranteed to identify the GMEC, and have been used successfully in a number of protein redesigns (Filikov et al., 2002; Gielens et al., 2007; Looger et al., 2003a; Maglia et al., 2008; Novoa de Armas et al., 2007). For a detailed description of computational protein design, see Chapter 2.

3.1.1 Restricted Redesign

Given a protein with n mutable residues, the redesign problem is said to be κ-restricted if there exists a limit κ, $(\kappa < n)$ on the maximum number of allowed mutations. The κ-restriction is motivated by a number of different causes. First, a limited number of mutations is ideal due to wet-lab experimental constraints where, in practice, redesigns containing a large number of mutated residues are often problematic. Despite significant advances over the past thirty years, the assumptions made in molecular modelling still induce a number of fundamental limitations to its predictive accuracy. In the context of protein redesign, the risk of major unpredicted and undesirable conformational changes occurring during wet-lab experimentation increases with the number of mutations introduced in the redesigns. These risks can be partially mitigated by restricting the number of allowed mutations. Furthermore, the introduction of a large number of point mutations via site-directed mutagenesis is time consuming and expensive. To this end, most of the published work in protein redesign reports a small number of mutations in the redesigns tested in the wet-lab (Bae et al., 2003; Chen et al., 2009b; Ito et al., 2008; Jouaux et al., 2009; Lilien et al., 2004; Stachelhaus et al., 1999; Stevens et al., 2006). Therefore, the κ-restriction's upper bound on the number of allowed mutations can aid wet-lab experimentation. As, only a small number of mutations are often *essential* for a change in function, such an upper bound can help introduce a minimal number of mutations essential for the desired change in functionality. And finally,

the notion of restricted redesign can provide insights into protein evolution, for instance, during the acquisition of drug resistance where a small number of mutations are critical for the acquisition of resistance.

In this thesis, such restricted redesign will be applied to explore the effect of a small number of mutations in a wild-type protein on its binding profile. Unfortunately, the traditional Dead-End Elimination (DEE) criteria do not address the restricted redesign problem directly. Hence, solving a restricted redesign problem with traditional DEE approaches requires a combinatorial number of DEE runs and can be quite inefficient.

3.1.2 Dead-End Elimination

The DEE criterion described in this section addresses the *unrestricted* ($\kappa = n$) redesign problem and facilitates conformational search by identifying and pruning rotamers which are provably not part of the GMEC. Given a pairwise energy function, the total energy E_c of a conformation can be specified as:

$$E_c = E_t + \sum_i E(i_r) + \sum_i \sum_{j>i} E(i_r, j_s) \tag{3.1}$$

This equation follows the notation established by the DEE community; E_t is the template self energy (*i.e.*, the self energy of the backbone and other rigid parts of the protein), i_r denotes the presence of rotamer r at position i, $E(i_r)$ is the self energy for rotamer i_r (*i.e.*, the sum of intra-residue and the residue-to-template energies), and $E(i_r, j_s)$ is the pairwise energy between rotamers i_r and j_s. The traditional DEE criterion (Desmet et al., 1992) prunes a rotamer i_r, if a second rotamer i_t is found such that the lowest (best) energy among the conformations including i_r is greater than the highest (worst) energy among all the conformations containing i_t (Figure 3.1A). Intuitively, a satisfied DEE criterion implies that the energy of any conformation involving i_r can be improved by exchanging i_t for i_r; therefore, i_r can not be part of the GMEC. The *traditional DEE* pruning criterion is written as:

$$E(i_r) + \sum_j \min_s E(i_r, j_s) > E(i_t) + \sum_j \max_s E(i_t, j_s) + \Delta \tag{3.2}$$

where s is selected from the set of allowed rotamers at position j.

DEE pruning criteria search for the GMEC as defined by an energy function. State-of-the-art energy functions aim to approximate the experimental interaction energies. These approximations often come close to, but do not perfectly, predict the true interaction en-

ergies. Hence, in the traditional DEE criterion, a Δ term allows a window of low energy conformations to be identified. When $\Delta > 0$, DEE will only prune rotamers that are not part of any conformation with energy within Δ of the GMEC. A biologist would then test a number of these conformations to the wet-lab. For clarity, we will assume $\Delta = 0$ in the remainder of this document, however this is not a limitation of rDEE.

The conservative pruning condition of Eq. (3.2) is often difficult to satisfy and hence provides limited pruning ability. Fortunately, a number of DEE extensions greatly enhance the method's power. Two of the more common extensions are the Goldstein (Goldstein, 1994) and split (Pierce et al., 2000) criteria. The Goldstein criterion (Eq. (3.3)) results from a simple algebraic manipulation of Eq. (3.2),

$$E(i_r) - E(i_t) + \sum_j \min_s (E(i_r, j_s) - E(i_t, j_s)) > 0 \tag{3.3}$$

The split-DEE criterion (Pierce et al., 2000) divides the conformational space into multiple partitions (Figure 3.1C) and allows a different rotamer to prune i_r in each partition. The split-DEE criterion is satisfied if there is always *some* rotamer at position i which is energetically favorable to i_r. In the single split (or $s = 1$) criterion, we identify a single split position (or residue) p with q remaining (unpruned) rotamers. The method splits conformational space into q partitions, where in partition x, the rotamer at position p is set to p_x ($x \leq q$). The Goldstein or traditional DEE criterion is then examined in each partition. If i_r is pruned in each partition, then i_r can be pruned from the entire conformational space. It is also possible to split on multiple residues (*i.e.,* $s = 2$) and exchange a further expansion of conformation space for improved pruning ability.

Beyond the traditional, Goldstein, and split criteria (Figure 3.1) several powerful and provably correct extensions have been added to the DEE family of algorithms. These methods include the ability to handle rotamer pairs (doubles (Goldstein, 1994; Gordon and Mayo, 1998; Lasters and Desmet, 1993)), energy minimization (MinDEE (Georgiev et al., 2006b; Georgiev et al., 2008b)), limited backbone flexibility (BD (Georgiev and Donald, 2007)), and backbone backrub motions (BRDEE (Georgiev et al., 2008a)). Although we do not explicitly discuss restricted DEE in the context of these extensions, the rDEE technique can be combined with any previous DEE method to perform a restricted search.

When pruning with DEE, dead-ending rotamers are identified by comparing all remaining rotamers against the specified DEE criterion. This process of evaluating the DEE criteria for all rotamers is defined as a *DEE cycle*. Highly efficient DEE cycles can be created by

sequential execution of multiple different DEE criteria. For example, a single DEE cycle might consist of first applying the Goldstein and then the Split-DEE criteria. The general contraction of conformation space allows some rotamers to be pruned in later cycles, even if they were not initially identified as dead-ending in early cycles. Although DEE guarantees to not prune a rotamer that is part of the GMEC, it does not guarantee that repeated DEE cycles will prune all but the GMEC. Therefore, a redesign algorithm will run multiple DEE cycles until either no more rotamers can be pruned or the GMEC has been identified (*i.e.*, only one rotamer remains at each position). When multiple rotamers remain, the DEE cycles are followed by an enumeration stage generally based on the A* branch-and-bound search algorithm. In (Leach and Lemon, 1998), an A* search is used to expand a conformation tree, such that the conformations are returned in order of increasing energy (*i.e.*, best-first). Hence, the first conformation returned by the A* search is the GMEC. If DEE was performed with $\Delta > 0$, conformations can be repeatedly enumerated by the A* search until one is identified with an energy Δ above the GMEC. Conformation enumeration should stop at this point because there is no guarantee on the presence and proper ordering of conformations with energies Δ above the GMEC. The A* search eliminates the need to examine all remaining conformations and typically results in a combinatorial factor reduction in the search space.

3.1.3 Restricted Dead-End Elimination Solution

In this work, we present Restricted DEE (rDEE) as an efficient pruning technique capable of solving the restricted redesign problem. The DEE criteria of Section 3.1.2 (henceforth referred to as uDEE) addressed the *unrestricted redesign problem*. Redesign of a protein with n mutable residues using uDEE, can contain as many as n mutations. The target of a κ-restricted redesign is the κGMEC - the minimum energy conformation among those with at most κ mutations. Since the goal of an unrestricted redesign is identification of the GMEC, the uDEE criteria of Section 3.1.2 make no guarantees about preserving the rotamers of the κGMEC. If the κGMEC is different than the GMEC, it is likely to be pruned during a uDEE cycle (Figure 3.1D). Therefore, to solve a κ-restricted redesign using the uDEE criteria requires n-choose-κ separate runs. In each run an explicitly specified set of κ residues are allowed to mutate, uDEE cycles are repeated until convergence, and the local minimum energy conformation is identified using A*. Upon completion of all n-choose-κ runs, the resulting conformations are merged and the κGMEC is identified. For reasonable values of n and κ, this process can be quite inefficient. One example of this inefficiency

is the redundant computation of repeatedly pruning of the same rotamer; if a rotamer i_r can be pruned in multiple of the n-choose-κ runs, this fact must be rediscovered by each run independently (*i.e.*, it must be discovered a combinatorial number of times). rDEE removes the need for multiple runs to solve a κ-restricted redesign, efficiently searches only the restricted regions of conformational search space, and guarantees generation of a gap-free list of top-ranking conformations. In the redesign of three test proteins, our rDEE pruning criteria is over 10-times faster than previous approaches. The following contributions are made in this chapter:

1. *Restricted DEE* (rDEE): a restricted version of the traditional DEE criterion which allows specification of a maximum number of allowed mutations, and incorporates this constraint in the pruning process. rDEE is provably correct and guarantees not to prune rotamers that are part of the κGMEC.
2. *Goldstein Restricted DEE* (GrDEE): a restricted analog of the Goldstein DEE criterion with extended pruning capabilities as compared to traditional rDEE.
3. *Split Restricted DEE* (Split-rDEE): a conformational splitting based extension to traditional rDEE. Split-rDEE derives from the original Split-DEE criterion and is provably accurate in the context of a κ restricted redesign.
4. *Restricted* A* (rA*): a restricted version of the standard A* search that exploits the smaller conformation space inherent to a restricted redesign.
5. Application of rDEE and rA* in κGMEC-based redesigns of three protein systems, Gramicidin Synthetase A, Plastocyanin, and the β1 domain of protein-G.

3.2 Approach

In the κGMEC, mutations may occur at any of the n residues, but the total number of mutations must be at most κ. The rDEE pruning criteria will identify a rotamer i_r as dead-ending if the rotamer is provably not part of the κGMEC.

Intuitively, previous DEE-based approaches (Desmet et al., 1992; Georgiev et al., 2006b; Goldstein, 1994; Gordon and Mayo, 1998; Lasters and Desmet, 1993; Pierce et al., 2000) mark rotamer i_r as dead-ending if a second rotamer i_t can be identified such that the energy of any allowed protein conformation including i_r can be reduced by exchanging rotamer i_t for i_r. Restricted DEE is no different; however, the definition of *allowed* protein conformations is inherently more complicated. In this section, we first define a number of terms necessary

to describe rDEE; we next provide the intuition behind the three cases rDEE must handle; and finally we introduce the formal rDEE criterion along with several powerful extensions. In this description we first consider *traditional rDEE*, a pruning condition analogous to the *traditional DEE* criteria of (Desmet et al., 1992). We then extend rDEE to incorporate the more powerful pruning ideas of Goldstein (Goldstein, 1994) and Splitting (Pierce et al., 2000).

In the derivation and description of restricted DEE we consider pruning a rotamer at position i. Let the *neighborhood* of residue i be the set of residues $j, j \neq i$. Let the number of flexible residues in the protein be n. We define a *mutation-position-vector*, $m = (0,1)^n$ as an n-dimensional binary vector corresponding to the mutated state of each residue. In this notation, $m(j) = 1$ indicates that a mutation is allowed at position j whereas $m(j) = 0$ restricts the residue to rotamers of the wildtype amino acid. We define a *mutation subspace* $M^k_{/x}$ to be the set of mutation-position-vectors with k or fewer mutations at positions *other than* x:

$$M^k_{/x} = \left\{ m \;\middle|\; \textstyle\sum_{i \notin x} m(i) \leq k, \;\; \sum_{i \in x} m(i) = 0 \right\} \tag{3.4}$$

A mutation-position-vector or mutation subspace can therefore be used to define the allowed rotameric states for each residue. We define the set of allowed rotamers at position i under the mutation-position-vector m as $\mathcal{R}_m(i)$. Finally, we introduce the $\mathsf{type}(i_x)$ operator to probe rotamer i_x as being either wt (wildtype) or mut (mutant).

In traditional uDEE, rotamer i_r is dead-ending (and can be pruned) if a second rotamer i_t can satisfy the pruning condition (Eq. (3.2)). This equation is independent of $\mathsf{type}(i_r)$ and $\mathsf{type}(i_t)$. Regardless of whether i_r or i_t are wildtype or mutant, the neighborhood of i is drawn from $M^{n-1}_{/i}$. Therefore it is always allowable to swap rotamer i_t for i_r. In other words, swapping i_r for i_t will not violate a restriction on the number of allowed mutations because there is no restriction to violate. However, when performing a restricted redesign, one needs to consider $\mathsf{type}(i_r)$ and $\mathsf{type}(i_t)$. When $\mathsf{type}(i_r) \neq \mathsf{type}(i_t)$ the allowable mutation subspaces for the neighborhoods of i_r and i_t are not the same. Fundamentally there are three cases to consider (Fig. 3.2): (1) $\mathsf{type}(i_r) = \mathsf{type}(i_t)$, (2) $\mathsf{type}(i_r) = \mathsf{mut}$ and $\mathsf{type}(i_t) = \mathsf{wt}$, (3) $\mathsf{type}(i_r) = \mathsf{wt}$ and $\mathsf{type}(i_t) = \mathsf{mut}$. The following description of the three rDEE cases will build from the traditional DEE criterion of (Desmet et al., 1992) applied to rotamers i_r and i_t in subspace m:

$$\mathrm{DEE}_m(i_r, i_t) :=$$

$$E(i_r) + \sum_j \min_{s \in \mathcal{R}_m(j)} E(i_r, j_s) > E(i_t) + \sum_j \max_{s \in \mathcal{R}_m(j)} E(i_t, j_s) \tag{3.5}$$

3.2.1 Restricted DEE (rDEE)

During a restricted DEE pruning cycle, we prune rotamer i_r by identifying a rotamer i_t that satisfies the rDEE condition. The rDEE pruning criteria depend on $\mathsf{type}(i_r)$ and $\mathsf{type}(i_t)$ and fall into one of the three cases. Each criterion uses a set of mutation-position-vectors to *cover* all allowed conformation space. If the DEE criteria hold in all allowed regions, then rotamer i_r can be identified as dead-ending.

Case 1: $\mathsf{type}(i_r) = \mathsf{type}(i_t)$

When the type of both rotamers is the same, all conformations of the neighborhood of i_r are also valid for the neighborhood of i_t. By combining the conformation at position i with conformations of its neighborhood, we can specify an allowed set of conformations for the entire protein including i_r or i_t as $i_r \times M_{/i}^{\kappa}$ and $i_t \times M_{/i}^{\kappa}$ if i_r and i_t are both wildtype, or $i_r \times M_{/i}^{\kappa-1}$ and $i_t \times M_{/i}^{\kappa-1}$ if i_r and i_t are both mutant (where $\times$ represents the Cartesian product). The rDEE criterion is satisfied if the DEE criterion is satisfied for all appropriate values of m.

$$\begin{aligned} &\mathrm{rDEE}_{\kappa}(i_r, i_t) := \\ &\quad \begin{cases} \forall m \in M_{/i}^{\kappa} & \mathrm{DEE}_m(i_r, i_t) \quad \text{if } \mathsf{type}(i_r) = \mathsf{type}(i_t) = \mathsf{wt} \\ \forall m \in M_{/i}^{\kappa-1} & \mathrm{DEE}_m(i_r, i_t) \quad \text{if } \mathsf{type}(i_r) = \mathsf{type}(i_t) = \mathsf{mut} \end{cases} \end{aligned} \tag{3.6}$$

Case 2: $\mathsf{type}(i_r) = \mathsf{mut}, \mathsf{type}(i_t) = \mathsf{wt}$

When i_r is mutant and i_t is wildtype, the neighborhood of i_r is allowed to contain at most $\kappa - 1$ mutations. It therefore suffices to compare the protein conformations of $i_r \times M_{/i}^{\kappa-1}$ to $i_t \times M_{/i}^{\kappa-1}$. Although i_t can be part of conformations with κ mutations in its neighborhood, these neighborhoods containing κ mutations are not allowed for i_r and can therefore be ignored. The DEE criterion only requires that for each allowable conformation of i_r that we find a better conformation i_t. The rDEE criterion for Case 2 can be written,

$$\begin{aligned} &\mathrm{rDEE}_{\kappa}(i_r, i_t) := \\ &\quad \begin{cases} \forall m \in M_{/i}^{\kappa-1} & \mathrm{DEE}_m(i_r, i_t) \quad \text{if } \mathsf{type}(i_r) = \mathsf{mut}, \mathsf{type}(i_t) = \mathsf{wt} \end{cases} \end{aligned} \tag{3.7}$$

Case 3: $\mathsf{type}(i_r) = \mathsf{wt}, \mathsf{type}(i_t) = \mathsf{mut}$

If i_r is a wildtype and i_t is a mutant, then pruning inherently becomes a bit more involved. The neighborhood of i_r will consist of mutation-position-vectors in $M^{\kappa}_{/i}$. Unfortunately, the neighborhood conformations containing κ mutations can not be combined with rotamer i_t as they would result in a protein with $\kappa+1$ mutations, thereby violating the restricted redesign. We address this problem by dividing the allowable mutation space into two subsets (handled by Ψ_1 and Ψ_2 of Eq. (3.8), both of which must hold). The first subset consists of a mutation subspace with up to $\kappa - 1$ mutations. This subspace is valid for both i_r and i_t and can be handled by Eq. (3.9). The second subset consists of a constructed mutation space where residue i is paired with a second residue p, selected so that at most one of i and p is mutated. This mutation subspace is created by combining i and p with $M^{\kappa-1}_{/i,p}$. The intuition behind this maneuver is that for any allowable protein conformation including i_r and κ neighbor mutations, we can change i_r to i_t and convert any of the mutated neighbors, p, back to wildtype and achieve a lower energy. For this manipulation to work, the criterion must hold for all mutated residues p (a proof of correctness follows at the end of this section).

$$\begin{aligned}&\mathrm{rDEE}_{\kappa}(i_r, i_t) := \\ &\quad \left\{ \; \Psi_1 \wedge \Psi_2 \quad \text{if } \mathsf{type}(i_r) = \mathsf{wt}, \mathsf{type}(i_t) = \mathsf{mut} \right.\end{aligned} \tag{3.8}$$

$$\Psi_1 := \left\{ \; \forall m \in M^{\kappa-1}_{/i} \quad \mathrm{DEE}_m(i_r, i_t) \right. \tag{3.9}$$

$$\Psi_2 := \left\{ \; \forall p, p \neq i, \quad \forall m \in M^{\kappa-1}_{/i,p}, \quad \mathrm{DEE}^p_m(i_r, i_t) \right. \tag{3.10}$$

where

$$\begin{aligned}&\mathrm{DEE}^p_m(i_r, i_t) := \\ &\quad E(i_r) + \sum_{j, j \neq i,p} \min_{s \in \mathcal{R}_m(j)} E(i_r, j_s) + \min_{u \in [\mathsf{mut \text{ or } wt}]} \left[E(p_u) + \sum_{j, j \neq p} \min_{s \in \mathcal{R}_m(j)} E(p_u, j_s) \right] > \\ &\quad E(i_t) + \sum_{j, j \neq i,p} \max_{s \in \mathcal{R}_m(j)} E(i_t, j_s) + \max_{u \in \mathsf{wt}} \left[E(p_u) + \sum_{j, j \neq p} \max_{s \in \mathcal{R}_m(j)} E(p_u, j_s) \right]\end{aligned} \tag{3.11}$$

Equation (3.11) differs slightly from Eq. (3.5) in that Eq. (3.11) must include the pairwise energy terms involving residue p. On the left-hand side of Eq. (3.11), residue p is evaluated with wildtype rotamer i_r. Residue p is therefore allowed to assume either a mutant or wildtype conformation. On the right-hand side of Eq. (3.11), residue p is evaluated in the

context of i_t and is therefore restricted to assume only wildtype rotamers. Residues at position $j, j \neq i, p$ may assume rotamers as specified by m.

Algorithmic Complexity of rDEE: The DEE search procedure comprises of an energy matrix pre-computation and a DEE pruning cycle. For multiple DEE runs, the pruning and enumeration costs significantly dominate the energy pre-computation costs. We now calculate the algorithmic complexity of DEE pruning.

For a design problem with n mutable residues, and a total number of r rotamers at each position, evaluation of Eq. (3.3) is $O(nr)$. Hence, a single uDEE cycle over all positions and all rotamers, takes $O(n^2r^3)$. The rDEE criteria (Eqs. (3.6), (3.7), and (3.8)) distinguish between the wildtype and mutant rotamers, let w denote this number of wildtype rotamers. We now calculate the complexity for a single evaluation of the rDEE criteria of Eqs. (3.6) and (3.8).

Case 1/ Case 2: For given wildtype i_r, i_t and a subspace m, run time of Eq. (3.6) is order of $\kappa r + (n-\kappa)w$. Hence the evaluation of Eq. (3.6) over all subspaces is $O(\binom{n}{\kappa}\ (\kappa r+(n-\kappa)w)\ )$. Note that as $w \ll r$, the runtime of Eq. (3.6) for wildtype rotamers i_r and i_t is an upper bound for the algorithmic complexity of Eq. (3.7).

Case 3: For given i_r, i_t, a position p and a subspace m, complexity of Eq. (3.8) is also order of $\kappa r + (n-\kappa)w$. Hence the evaluation of Eq. (3.8) over all positions and subspaces is $O(\binom{n}{\kappa}\ n(\kappa r+(n-\kappa)w)\ )$. As $n \ll r$ for most redesign problems, the algorithmic complexity for any rDEE criterion becomes $O(\binom{n}{\kappa}\ (\kappa r + (n-\kappa)w))$. Therefore, the complexity of an entire rDEE cycle is $O(nr^2\ (\binom{n}{\kappa}\ (\kappa r + (n-\kappa)w))$.

Although the rDEE pruning criteria are inherently more involved than uDEE, rDEE is a more powerful pruning technique. To prune rotamer i_r we only need to consider conformations containing up to κ mutations. This contrasts with uDEE where to prune i_r the DEE condition must hold over all of mutation space. This superior pruning ability is manifested in significantly smaller runtimes for fifteen distinct redesigns (See Section 3.4). The pruning ability of the rDEE criterion (Eqs. (3.6), (3.7), and (3.8)) is further enhanced by inclusion of Goldstein and Split Restricted DEE conditions.

3.2.2 Goldstein Restricted DEE

Whereas the original DEE criterion compares the minimum pairwise energy between i_r and each residue j to the maximum pairwise energy between i_t and each residue j, the Goldstein criterion (Goldstein, 1994) examines the *difference* in pairwise energy between rotamers i_r

and i_t. This simple relaxing of the DEE criterion still guarantees to only prune residues that are provably not part of the GMEC. The rDEE criteria of section 3.2.1 can be extended to incorporate the Goldstein criteria by replacing the original $\text{DEE}_m(i_r, i_t)$ of Eqs. (3.6), (3.7), and (3.8) with the following Goldstein criteria defined over subspace m:

$$\text{GoldsteinDEE}_m(i_r, i_t) := \quad E(i_r) - E(i_t) + \sum_{j, j \neq i} \min_{s \in \mathcal{R}_m(j)} (E(i_r, j_s) - E(i_t, j_s)) > 0 \quad (3.12)$$

The traditional rDEE criteria can only prune a rotamer i_r if a single rotamer i_t is always more energetically favorable. Following the ideas of Split-DEE, we observe that the mutation subspaces m already induce a partitioning of conformation space. Therefore in Split Restricted DEE we remove the limitation that the same single rotamer i_t prune rotamer i_r in every mutation subspace.

3.2.3 Split Restricted DEE

The restricted DEE criteria can be extended to incorporate the Split-DEE criterion (Pierce et al., 2000) (see Section 3.1.2). In Split-DEE, conformation space is split into partitions using the rotamers at a residue $k, k \neq i$. A rotamer i_r is dead-ending if there exists some rotamer at position i capable of pruning i_r in each of the partitions defined by the rotamers v of position k. We can incorporate Split-DEE into rDEE by removing the explicit i_t and replacing the original $\text{DEE}_m(i_r, i_t)$ of Eqs. (3.6), (3.7), and (3.8) with the Split-DEE criteria defined over subspace m.

$$\begin{aligned} &\text{SplitDEE}_m(i_r, \cdot) := \exists\ k,\ (k \neq i) \text{ s.t. } \forall v \in \mathcal{R}_m(k)\ \exists\ i_t : \\ &\quad E(i_r) - E(i_t) + \textstyle\sum_{j, j \neq k, i} \left\{ \min_{u \in \mathcal{R}_m(j)} \left[E(i_r, j_u) - E(i_t, j_u) \right] \right\} \\ &\qquad\qquad + \left[E(i_r, k_v) - E(i_t, k_v) \right] > 0 \end{aligned} \quad (3.13)$$

Direct implementation of the criterion of Eq. (3.13) introduces a significant bookkeeping overhead which can empirically result in slower running times. Therefore, a slightly modified version of Eq. (3.13) was implemented. The pseudocode for the modified implementation is shown in Fig. 3.3.

Proof of Correctness A formal proof of correctness for the traditional rDEE criteria is provided below. Proofs of Goldstein and Split rDEE are nearly identical and are omitted to avoid redundancy.

We define $\hat{m}$ as the subspace containing the κGMEC. We indicate the rotamers of the

κGMEC with a subscript g (*i.e.*, the κGMEC contains rotamers $j_g, (j = 1 \ldots n)$). We define $\mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}})$ to be the energy of a conformation in $\hat{m}$ with κGMEC rotamers j_g at all positions $j \neq i$ and rotamer i_r at position i. skip

Case 1: $\text{type}(i_r) = \text{type}(i_t)$

To Prove: If i_t eliminates i_r by Eq. (3.6) then i_r cannot be part of κGMEC ($i_r \neq i_g$).

Proof by contradiction. Assume $i_r = i_g$.

Without loss of generality, assume $\mathsf{type}(i_r) = \mathsf{type}(i_t) = \mathsf{wt}$. Therefore the mutation subspace, $\hat{m}$, containing the κGMEC is a member of $M^{\kappa}_{/i}$.

Since $i_r = i_g$,

$$\mathsf{E}_{\kappa\mathsf{GMEC}} = \mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}}) = E_t + E(i_r) + \sum_j E(i_r, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g)$$

Because $\mathsf{type}(i_r) = \mathsf{type}(i_t)$, there exists a conformation in $\hat{m}$ with energy $\mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}})$.

$$\mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}}) = E_t + E(i_t) + \sum_j E(i_t, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g)$$

Then, because $\mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}}) = \mathsf{E}_{\kappa\mathsf{GMEC}} \leq \mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}})$, we can state $\mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}}) \leq \mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}})$ and

$$\begin{aligned} E_t + E(i_r) + \sum_j E(i_r, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g) \leq \\ E_t + E(i_t) + \sum_j E(i_t, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g)) \end{aligned}$$

Canceling identical terms we get,

$$E(i_r) + \sum_j E(i_r, j_g) \leq E(i_t) + \sum_j E(i_t, j_g) \tag{3.14}$$

We now note the following conservative bounds,

$$\sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_r, j_s) \leq \sum_j E(i_r, j_g) \quad \text{and} \quad \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_t, j_s) \geq \sum_j E(i_t, j_g)$$

Substituting these bounds into Eq. (3.14),

$$E(i_r) + \sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_r, j_s) \leq E(i_t) + \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_t, j_s) \tag{3.15}$$

As, Eq. (3.6) was used to eliminate i_r in $\hat{m} \in M^{\kappa}_{/i}$, Eq. (3.15) cannot be true. This contradiction

proves that $i_r \neq i_g$ and that i_r is not part of the κGMEC.

Case 2: $\mathsf{type}(i_r) = \mathsf{mut}, \mathsf{type}(i_t) = \mathsf{wt}$

To Prove: If i_t eliminates i_r by Eq. (3.7), then i_r cannot be part of κGMEC ($i_r \neq i_g$).

Proof by contradiction. Assume $i_r = i_g$.

Because i_r is a mutant, the mutation subspace, $\hat{m}$, containing the κGMEC is a member of $M_{/i}^{\kappa-1}$. Since $i_r = i_g$,

$$\mathsf{E}_{\kappa\mathsf{GMEC}} = \mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}}) = E_t + E(i_r) + \sum_j E(i_r, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g)$$

Then, as $\mathsf{type}(i_t) = \mathsf{wt}$, there exists a conformation with energy $\mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}})$:

$$\mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}}) = E_t + E(i_t) + \sum_j E(i_t, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g)$$

Then as $\mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}}) = \mathsf{E}_{\kappa\mathsf{GMEC}} \leq \mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}})$, we can state $\mathsf{E}_{\mathsf{i_r}}(\hat{\mathsf{m}}) \leq \mathsf{E}_{\mathsf{i_t}}(\hat{\mathsf{m}})$ and,

$$\begin{aligned} E_t + E(i_r) + \sum_j E(i_r, j_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g)) \leq \\ E_t + E(i_t) + \sum_j E(i_t, j_g) + \sum_j E(j_g) + \sum_j \sum_k E(j_g, k_g) \end{aligned}$$

Canceling identical terms we get,

$$E(i_r) + \sum_j E(i_r, j_g) \leq E(i_t) + \sum_j E(i_t, j_g) \tag{3.16}$$

We now note the following conservative bounds,

$$\sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_r, j_s) \leq \sum_j E(i_r, j_g) \quad \text{and} \quad \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_t, j_s) \geq \sum_j E(i_t, j_g)$$

Substituting these bounds into Eq. (3.16),

$$E(i_r) + \sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_r, j_s) \leq E(i_t) + \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_t, j_s) \tag{3.17}$$

As Eq. (3.7) was used to eliminate i_r in $\hat{m} \in M_{/i}^{\kappa-1}$, Eq. (3.17) cannot be true. This contradiction proves that $i_r \neq i_g$ and that i_r is not part of the κGMEC.

Case 3: $\mathsf{type}(i_r) = \mathsf{wt}, \mathsf{type}(i_t) = \mathsf{mut}$

To Prove: If i_t eliminates i_r by Eq. (3.8), then i_r cannot be part of κGMEC ($i_r \neq i_g$).

If κGMEC has less than κ mutations, i_r is eliminated by Eq. (3.9). The proof is thus similar to case 1 and case 2 above; the proof is not shown here.

We now address the case where the κGMEC has κ mutations and rDEE resorts to Eq. (3.10) for pruning.

Proof by contradiction. Assume $i_r = i_g$.

Then $\mathsf{E}_{\kappa\mathsf{GMEC}} = \mathsf{E}_{\mathsf{i_r},\acute{\mathsf{p}}_\mathsf{g}}(\hat{\mathsf{m}})$ for some arbitrary flexible residue position $\acute{p} \neq i$. The $\hat{m} \in M^{\kappa-1}_{/i,\acute{p}}$ and $\hat{m}$ can introduce up to $\kappa - 1$ mutations in positions other than i and $\acute{p}$. The rotamer $\acute{p}$ can be a mutant without violating the restricted redesign.

We write the energy of $\mathsf{E}_{\mathsf{i_r},\acute{\mathsf{p}}_\mathsf{g}}(\hat{\mathsf{m}})$ by separating energetic contributions of rotamer i_r and another rotamer $\acute{p}_g$:

$$\begin{aligned}\mathsf{E}_{\mathsf{i_r},\acute{\mathsf{p}}_\mathsf{g}}(\hat{\mathsf{m}}) &= E_t + E(i_r) + \sum_j E(i_r, j_g) + E(\acute{p}_g) + \sum_j E(\acute{p}_g, j_g) \\ &+ E(i_r, \acute{p}_g) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g) \qquad \text{where } (j, k \neq \acute{p}, i) \end{aligned} \tag{3.18}$$

There exists a conformation C identical to the κGMEC but with i_t instead of i_r and $\acute{p}_v$ instead of $\acute{p}_g$. This conformation has energy $\mathsf{E}_{\mathsf{i_t},\acute{\mathsf{p}}_\mathsf{v}}(\hat{\mathsf{m}})$:

$$\begin{aligned}\mathsf{E}_{\mathsf{i_t},\acute{\mathsf{p}}_\mathsf{v}}(\hat{\mathsf{m}}) &= E_t + E(i_t) + \sum_j E(i_t, j_g) + E(\acute{p}_v) + \sum_j E(\acute{p}_v, j_g) \\ &+ E(i_t, \acute{p}_v) + \sum_j E(j_g) + \sum_j \sum_{k,k>j} E(j_g, k_g) \qquad \text{where } (j, k \neq \acute{p}, i) \end{aligned} \tag{3.19}$$

Then, since $\mathsf{E}_{\mathsf{i_r},\acute{\mathsf{p}}}(\hat{\mathsf{m}}) = \mathsf{E}_{\kappa\mathsf{GMEC}} \leq \mathsf{E}_{\mathsf{i_t},\acute{\mathsf{p}}_\mathsf{v}}(\hat{\mathsf{m}})$, we can state $\mathsf{E}_{\mathsf{i_r},\acute{\mathsf{p}}_\mathsf{g}}(\hat{\mathsf{m}}) \leq \mathsf{E}_{\mathsf{i_t},\acute{\mathsf{p}}_\mathsf{v}}(\hat{\mathsf{m}})$ and substitute Eqs. (3.18) and (3.19). After simplification,

$$\begin{aligned} E(i_r) + \sum_j E(i_r, j_g) + E(\acute{p}_g) + \sum_j E(\acute{p}_g, j_g) + E(i_r, \acute{p}_g) \leq \\ E(i_t) + \sum_j E(i_t, j_g) + E(\acute{p}_v) + \sum_j E(\acute{p}_v, j_g) + E(i_t, \acute{p}_v) \end{aligned} \tag{3.20}$$

We now note the following conservative bounds,

$$\begin{aligned} \sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_r, j_s) &\leq \sum_{j, j \neq \acute{p}} E(i_r, j_g) + E(i_r, \acute{p}_g) \\ \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_t, j_s) &\geq \sum_{j, j \neq \acute{p}} E(i_t, j_g) + E(i_r, \acute{p}_v) \\ \min_{u \in \{\mathsf{mut},\mathsf{wt}\}} \left(E(\acute{p}_u) + \sum_{j, j \neq i} \min_s E(\acute{p}_u, j_s) \right) &\leq E(\acute{p}_g) + \sum_{j, j \neq i} E(\acute{p}_g, j_g) \end{aligned}$$

and for a wildtype rotamer $\acute{p}_v$

$$\max_{u \in \{\mathsf{wt}\}} \left(E(\acute{p}_u) + \sum_{j, j \neq i} \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(\acute{p}_u, j_s) \right) \geq E(\acute{p}_v) + \sum_{j, j \neq i} E(\acute{p}_v, j_g)$$

Substituting these four bounds into Eq. (3.20),

$$E(i_r) + \sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_r, j_s) + \min_{u \in \{\mathsf{mut}, \mathsf{wt}\}} \left(E(\acute{p}_u) + \sum_j \min_{s \in \mathcal{R}_{\hat{m}}(j)} E(\acute{p}_u, j_s) \right) \leq$$

$$E(i_t) + \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(i_t, j_s) + \max_{u \in \{\mathsf{wt}\}} \left(E(\acute{p}_u) + \sum_j \max_{s \in \mathcal{R}_{\hat{m}}(j)} E(\acute{p}_u, j_s) \right) \tag{3.21}$$

As Eq. (3.10) (using Eq. (3.11)) was used to eliminate i_r in $\hat{m} \in M^{\kappa-1}_{/i,p}$ for all mutant positions p other than i, Eq. (3.21) cannot be true. This contradiction proves that $i_r \neq i_g$ and that i_r is not part of the κGMEC.

3.2.4 Restricted A* Search

The restrictions of a κ-restricted redesign can be exploited during the enumeration stage of the redesign process. Following (Leach and Lemon, 1998), we introduce a restricted A* (or rA*) search for conformation enumeration in restricted redesign. An A* search is best-first and uses bounds to evaluate the goodness of a rotamer n_d at level d before adding it to the search queue. Specifically, before a rotamer n_d at depth d is selected, a score f^* is computed. The score f^* is a sum of two functions g^* and h^*. The function g^* is the energy of the partial conformation up to and including n_d (*i.e.*, rotamers at levels $[1 \ldots (d-1)]$); whereas h^* is a lower bound on the minimum energy required to complete the conformation ((Leach and Lemon, 1998) provides an excellent overview of A* in DEE),

$$h^* = \sum_{j=d+1}^{N-1} \min_s \left(E(j_s) + \sum_{i=0}^{d} E(i_r, j_s) + \sum_{k=j+1}^{N-1} \min_t E(k_t, j_s) \right) \tag{3.22}$$

The estimate h^* can be modified to exploit the κ-restriction. The modified estimate $\hat{h^*}$ approximates the minimum energy required to complete the conformation with at most κ mutations. We define $h^*(n_d, m)$ to be the minimum energy estimate required to complete the conformation which has

rotamer n_d at level d and lies in the mutation subspace m.

$$h^*_m(n_d) = \sum_{j=d+1}^{N-1} \min_{s \in \mathcal{R}_m(j)} \left(E(j_s) + \sum_{i=0}^{d} E(i_r, j_s) + \sum_{k=j+1}^{N-1} \min_{t \in \mathcal{R}_m(k)} E(k_t, j_s) \right) \tag{3.23}$$

where the rotameric identities used to compute Eq. (3.23) are determined by the mutation vector m. The overall minimum energy estimate for n_d can be obtained by taking the minimum value of $h^*_m(n_d)$ over all applicable $m \in M^{\kappa}_{/d}$.

$$\hat{h}^*(n_d) = \begin{cases} \min_{m \in M^{\kappa}_{/d}} & h^*_m(n_d) \quad \text{if } \mathsf{type}(n_d) = \mathsf{wt} \\ \min_{m \in M^{\kappa-1}_{/d}} & h^*_m(n_d) \quad \text{if } \mathsf{type}(n_d) = \mathsf{mut} \end{cases} \tag{3.24}$$

The A^* algorithm guarantees to generate the best solution that complies with the specified scoring functions g^* and h^*. Our definition of g^* is identical to the one in (Leach and Lemon, 1998). However, we restrict our bound h^* to estimate the minimum energy required to complete a conformation with at most κ mutations. Therefore, rA* only searches the κ-restricted space to identify the κGMEC.

We note that the uDEE criteria might prune rotamers that are part of the κGMEC. Therefore, restricted A* cannot guarantee to identify κGMEC using the results of a uDEE pruning cycle.

3.3 Methods

Sixty κ-restricted redesign experiments were performed (Table 3.1) to compare the performance of rDEE and restricted A* against their unrestricted counterparts. For each of the three protein systems (see Structural Models below), restricted redesigns were performed for $\kappa = 2, 3, 4$, and 5. Restricted DEE was evaluated using one of two rDEE cycles: (1) GrDEE (Section 3.2.2), or (2) GrDEE followed by (s=1) Split-rDEE (Section 3.2.3). Within a cycle, each rDEE criterion is evaluated until no additional rotamers can be pruned. Similarly, the entire rDEE cycle is repeated until a cycle produces no additional pruned rotamers. Following rDEE pruning, the κGMEC is identified using restricted A*. For comparison, we performed κ-restricted redesigns using the standard uDEE and unrestricted A* (uA*) algorithms implemented as n-choose-κ separate runs. The DEE cycles for these unrestricted criteria consisted of either (1) the Goldstein DEE criterion or (2) Goldstein DEE followed by ($s = 1$) Split-DEE. The κGMEC was identified as the lowest energy conformation among the n-choose-κ local minima generated by unrestricted A* enumerations. Performance of rDEE criteria was then compared against these benchmarks. All experiments were performed on a single processor.

Structural Models: Protein systems were selected based on our previous experience with the

system or its previous use as a benchmark in the DEE community. The selected systems are diverse and range from a small redesign of 9 active site residues (NRPS) to larger redesigns with 18 residues (plastocyanin). The first system is the NRPS enzyme GrsA-PheA (PDB: 1AMU) (Conti et al., 1997). Similar to (Lilien et al., 2005), nine active site residues were modeled as flexible (235,236,239,278,299,301,322,330,331) and were allowed to mutate to a set of hydrophobic amino acids (GAVLIFYWM). Our model also included the amino acid substrate, the AMP cofactor, and a steric shell consisting of all residues with at least one atom within 8 $\AA$ of the substrate. The second test system was the core of the β1 domain of protein-G (PDB: 1PGA) (Gallagher et al., 1994). Similar to (Georgiev et al., 2006a) and (Shah et al., 1999), we allowed 12 flexible core residues (3,5,7,9,20,26,30,34,39,41,52,54) to mutate to the hydrophobic set (GAVLIFYWM). Finally, the third test system was plastocyanin (PDB: 2PCY) (Garret et al., 1984). Similar to (Gordon et al., 2003), we model 18 core residues as flexible (5,14,21,27,29,31,37,38,39,41,72,74,80,82,84,92,96,98) allowing mutations to the hydrophobic set of residues (AVLIFYW). In all systems, side-chain flexibility was modeled using the Richardson's rotamer library (Lovell et al., 2000a). A validated implementation of the AMBER energy function (electrostatic, vdW, and dihedral energy terms) (Cornell et al., 1995a; Weiner et al., 1984a) was used to compute pairwise energies.

3.4 Results and Discussion

Restricted Redesign: The runtime results of sixty redesign experiments are summarized in Table 3.1. The results are consistent across all three test systems, (a) GrsA-PheA, (b) β1 domain of protein G, and (c) plastocyanin. The unrestricted DEE and A* based redesigns require up to 10 times as long to complete than their restricted counterparts. This trend holds for all tested values of κ. We expect this improvement in runtime to become even more pronounced for larger values of n and κ. The results also confirm the expected impact on pruning efficiency provided by the split criteria. A significant number of experiments did not complete within twelve hours. For example, none of the uDEE + uA* runs for plastocyanin completed within twelve hours. In contrast, all of the Split-rDEE runs completed within the allowable time window.

The use of GrDEE-only pruning, results in a runtime advantage for both systems for small values of κ. For $\kappa = 2$, restricted redesign finishes in less than half the time taken by the unrestricted benchmark. However, for GrsA-PheA, as κ increases, the pruning ability of GrDEE does not keep pace with the additional bookkeeping required for a restricted search. As a result, the GrDEE cycle is occasionally slower than the corresponding unrestricted redesign. The specific case of GrsA-PheA may be explained by a peculiarity of the protein system, which contains three alanines (each with a single rotamer) among the nine mutable residues. The presence of three alanines provides a significant pruning advantage to uDEE, where almost 90% of the n-choose-κ runs have at least

		uDEE and uA*		rDEE and rA*	
	κ	Goldstein	Split-DEE	GrDEE	Split-rDEE
GrsA-PheA	2	2.0	2.1	0.6	0.3
($n = 9$)	3	4.8	5.1	6.8	0.5
	4	7.4	7.2	63.0	1.4
	5	8.4	11.2	+	3.4
β1 domain protein-G	2	11.2	10.9	0.6	0.5
($n = 12$)	3	19.6	34.8	8.4	0.9
	4	65.2	62.6	32.3	3.0
	5	115.6	117.6	238.0	7.7
Plastocyanin	2	+	+	+	1.3
($n = 18$)	3	+	+	+	8.4
	4	+	+	+	44.4
	5	+	+	+	216.4

+ : did not complete within 12 hours

Table 3.1: **Runtimes (in minutes) for κ-Restricted Redesigns of GrsA-PheA ($n = 9$), core of β1 domain of protein-G ($n = 12$) and Plastocyanin ($n = 18$).** Runtimes include DEE pruning as well as A* enumeration. DEE criteria evaluated were Goldstein (Goldstein), restricted Goldstein (GrDEE), unrestricted split (s=1) (Split-DEE), and Split (s=1) Restricted (Split-rDEE) criteria. All restricted DEE criteria (rDEE) were followed by restricted A* (rA*) enumeration. All uDEE criteria were followed by uA* enumeration. Most rDEE based redesigns are 10 times faster than their uDEE counterparts. All experiments were performed on a single processor.

one position with a single rotamer (essentially reducing the number of flexible residues). However, despite this situation, our Split-rDEE runs significantly outperform their uDEE counterparts. This illustrates the importance of combining multiple different rDEE criteria.

The rDEE criteria are provably correct and therefore guaranteed not to prune any rotamer that is part of the κGMEC. As a result, we expect the same κGMEC to be identified by the restricted and unrestricted searches. This result was confirmed by comparing the set of top-ranking conformations returned by our rDEE based methods with that returned by their uDEE counterparts. An analysis of the specificity determining residues for the NRPS redesign with different ligands is provided in (Stachelhaus et al., 1999). We note that the top mutation sequences returned by rDEE based methods contain the specificity determining mutation A301G for the GrsA-PheA redesign, pointing to the biochemical feasibility of the solutions suggested by rDEE. We also note that certain aspects, such as sequence recovery, remain unchanged between uDEE and rDEE approaches as they are independent of the DEE criteria and are dependent on the energy function instead.

Restricted A* Evaluation: The performance advantage provided by restricted A* was further evaluated against its unrestricted counterpart. For two systems (GrsA-PheA and β1 domain of protein G) and $\kappa = 2, 3, 4$, and 5, rDEE pruning was followed either by an unrestricted uA* or a restricted rA* enumeration stage (Table 3.2). As κ increases, the number of irrelevant conformations (*i.e.*, conformations with more than κ mutations) in uA* grows quickly, leading to a significant loss in efficiency.

For GrsA-PheA, the number of possible conformations containing κ or fewer mutations (*i.e.* the entire restricted search space), is 2.4×10^8, 1.1×10^{10}, 2.7×10^{11}, and 4.2×10^{12} for $\kappa = 2, 3, 4$, and 5 respectively. The uA* search does not limit the number of allowed mutations, consequently it is forced to consider a significantly larger search space. For example, after pruning with Split Restricted DEE, the number of GrsA-PheA conformations considered by uA* is 9.1×10^{12}, 1.4×10^{12}, 6.8×10^{12}, and 1.5×10^9 (for $\kappa = 2, 3, 4$, and 5 respectively). Note that in GrsA-PheA (where $n = 9$), the number of unpruned conformations for $\kappa = 5$ (where over half the residues are mutable) is smaller than that for $\kappa = 4$. This reflects the fact that in a redesign with a large number of allowed mutations, a highly mutated, low energy solution may prune large number of rotamers compared to a smaller redesign where no such solution exists. Table 3.2 shows the number of partial solutions examined before the κGMEC was found in both unrestricted and restricted A* runs. For all runs, restricted A* evaluates far fewer partial solutions, a fact reflected in the runtimes of Table 3.1. For increasing values of κ, uA* evaluates between 10 and 100 times as many partial solutions as does restricted A*. These results provide insight not only into the pruning ability but also the memory requirements of the two approaches.

		Goldstein Restricted DEE		Split Restricted DEE	
	κ	uA*	rA*	uA*	rA*
GrsA-PheA	2	8.2×10^4	1.6×10^4	7.0×10^3	1.3×10^3
($n = 9$)	3	6.3×10^5	6.5×10^4	5.6×10^4	8.0×10^3
	4	+	2.4×10^5	1.5×10^5	2.4×10^4
	5	+	+	2.7×10^5	7.0×10^4
β1 domain protein-G	2	7.5×10^3	2.1×10^2	7.8×10^3	7.0×10^1
($n = 12$)	3	6.3×10^5	9.8×10^3	2.1×10^4	1.3×10^3
	4	1.7×10^6	1.2×10^5	6.1×10^4	5.7×10^3
	5	+	2.3×10^5	2.5×10^5	2.1×10^4

+ : did not complete within 12 hours

Table 3.2: **Comparison of the Number of Partial Solutions Evaluated by uA* and rA* for Redesigns of GrsA-PheA and core of β1 domain of protein-G**. For both Goldstein rDEE and Split-rDEE the restricted A* search evaluates an order of magnitude fewer conformations than the unrestricted A* search. In all cases, the number of conformations evaluated by A* is far fewer than the total number of allowed conformations (see text).

3.5 Conclusion

This chapter addresses the restricted protein redesign problem as a first step towards a structural method of predicting drug resistance. In a κ-restricted redesign, solutions may contain up to κ mutations among the protein's n mutable residues. First, rDEE, a novel version of DEE for use in restricted redesign problems, was introduced. Second, restricted A*, an enhancement to unrestricted A* search for enumerating conformations in restricted redesigns, was presented. The results presented in the chapter support rDEE as the method of choice when a protein redesign requires an upper limit on the number of allowed mutations.

Following the introduction of rDEE and restricted A*, several enhanced pruning criteria were also introduced. These extensions include the Goldstein Restricted and Split Restricted criteria. All rDEE pruning criteria are both deterministic and provably correct and therefore guarantee not to prune any rotamer that is part of the κGMEC. It was demonstrated that the rDEE criteria and restricted A* search reduce the runtime of restricted redesigns by an order of magnitude when compared to traditional methods. These results were illustrated on three test systems. As the values of n and κ increase, the execution of multiple unrestricted DEE runs becomes extremely time consuming. Therefore, we can conclude that the use of the presented methods is both a necessary and enabling part of any large restricted redesign.

The challenge for protein redesign has been and will continue to be the development of accurate and efficient models that simultaneously honor the underlying biophysics, are provably accurate

with respect to the scoring function, and are computationally efficient. Although our restricted redesigns were able to complete within a day, for larger protein systems it may be useful to extend rDEE to incorporate the advanced pruning techniques of split flags, $s = 2$ split-DEE, or dead-ending pairs. Additionally, we can incorporate a model of backbone flexibility, by combining rDEE with BD or BRDEE. We can also handle energy minimization by adapting rDEE to use Energy Minimized DEE (MinDEE). Incorporation of these criteria would result in more efficient searches thereby allowing κ-restricted redesigns for larger values of n and κ. We note that all of these extensions are feasible given our rDEE formulation.

In this chapter, the impact of a small number of mutations (redesigns) on a protein's ligand binding profile was explored. Next chapters of this thesis will employ rDEE to model the acquisition of drug resistance mutations in drug target proteins. Furthermore, the evolutionary life of a drug target protein will be probed by exploring the local mutational landscape surrounding a wild type sequence. Apart from its use in modelling drug resistance, this work can also facilitate structure-based drug discovery both in terms of lead optimization and lead prioritization. Finally, for a given disease, restricted mutational analysis may also allow prioritization among multiple candidate protein targets.

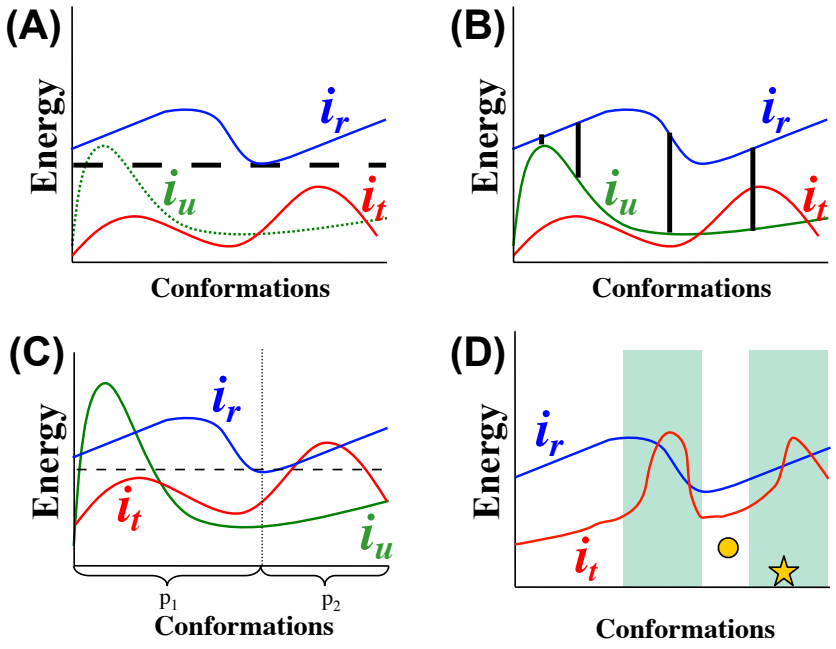

Figure 3.1: **Pruning by Traditional DEE, Goldstein DEE, Split-DEE and Restricted DEE.** Following the conformational plots of (Pierce et al., 2000) the abscissa represents all possible conformations of the protein (excluding the conformation of residue i) and the curve for rotamer i_x represents the total energy of conformations containing i_x at i. (a) Traditional DEE: rotamer i_t can prune rotamer i_r because the lowest (best) energy (dashed line) among the conformations including i_r is greater than the highest (worst) energy among all conformations containing i_t. Rotamer i_u is not able to prune i_r. (b) Using Goldstein DEE both rotamers i_u and i_t can prune i_r. The vertical lines represent the energy difference of Eq. (3.3). (c) Using Split-DEE, rotamer i_t and i_u together are able to prune i_r (i_t is a better alternative in partition p_1, i_u is a better alternative in p_2). (d) In the restricted redesign problem, conformations with greater than κ mutations are disallowed (shaded regions). In an unrestricted redesign, rotamer i_t would not be able to prune i_r; however i_t can prune i_r in a restricted redesign. In this example the GMEC (star) contains greater than κ mutations; the κGMEC is designated with a circle.

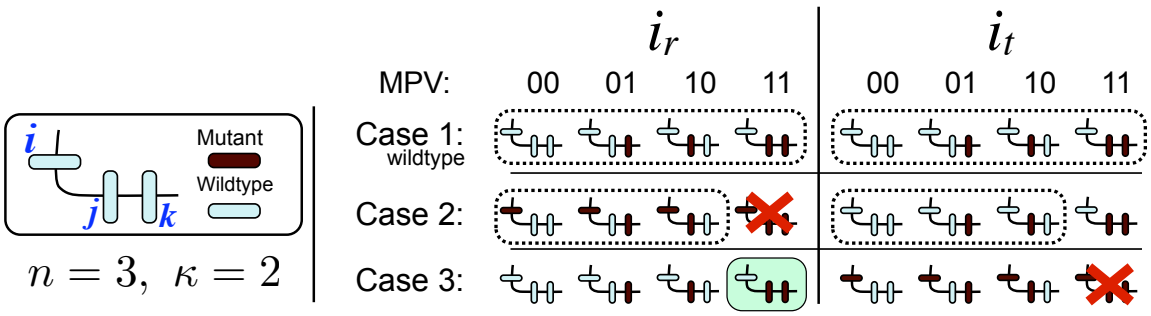

Figure 3.2: **Graphical Representation of the Mutation-Position-Vectors (MPV) for the Three Cases of Restricted DEE.** A small three residue protein is shown for a $\kappa = 2$ restricted redesign. A mutated residue is darkly shaded, a wildtype residue is light. For Case 1 (shown with wildtype i_r) and Case 2, every allowable conformation of the neighborhood of i_r (circled structures) is also an allowed conformation for the neighborhood of i_t. This is not true for Case 3, where an MPV of (11) is allowed for rotamer i_r but not for rotamer i_t.

canEliminate($\mathbf{i_r}$)

for each split position k
  for each partition k_v
    for each subspace m
      S_m = < $\exists i_t, i_t$ eliminates i_r in m >
    if $\forall m, S_m$
      prune partition k_v
  if all partitions are pruned
    return true
return false

canSplitEliminate($\mathbf{i_r}$)

for each split position k
  for each partition k_v
    S_{k_v} = < $\exists i_t, i_t$ eliminates i_r in k_v >
  if $\forall k_v, S_{k_v}$
    return true
return false

canSpaceEliminate($\mathbf{i_r}$)

for each valid subspace m
  S_m = < $\exists i_t, i_t$ eliminates i_r in m >
if $\forall m, S_m$
  return true
return false

Figure 3.3: **Pseudocode for Split Restricted DEE** (Left) The pseudocode for the Split-rDEE of Eq. (3.13). (Right) An alternative pseudocode for restricted Split-rDEE. Multiple runs of **canSplitEliminate** are performed until no more pruning can be achieved, followed by multiple runs of **canSpaceEliminate** until no more pruning can be achieved again. The results in Table 3.1 were obtained using the Right implementation of Split-rDEE. Partition k_v refers to the partition with splitting rotamer k_v.

Chapter 4

Efficient A Priori Identification of Drug Resistant Mutations using Dead-End Elimination and MM-PBSA.

Chapter 3 described restricted Dead-End Elimination (rDEE) as the protein design algorithm of choice when the objective is to limit the number of mutations in the redesign. When the objective of such a restricted redesign is a reduction in binding affinity of the mutant protein for the drug, rDEE can be used to predict resistance conferring mutations. This chapter describes a resistance prediction algorithm that utilises rDEE to predict drug resistance *a priori* in four drug-target systems. This chapter has been reprinted and adapted with permission from (Safi and Lilien, 2012).

4.1 Introduction

The emergence of drug resistance remains a significant and frustrating cause of treatment failure. Four molecular mechanisms underly the majority of drug resistance: point mutations in the drug target, alterations in non-target compensatory genes, increased drug metabolism, and reduction in intracellular concentration through reduced cellular uptake or upregulated small molecule efflux (Blanchard, 1996; Borst, 1991; Erickson and Burt, 1996) (see Chapter 2). In this chapter, we focus on modelling the most direct and prevalent ways in which resistance is conferred: the introduction of mutations within the drug binding site (Parikh et al., 1999; Arnold et al., 2005; Volpato

et al., 2007a; Bang et al., 2011; Chen et al., 2013). To cause resistance, these mutations must maintain near-native protein function, otherwise they are effectively inherently inhibited. Therefore, in order to confer resistance, a binding site mutation should reduce drug binding while maintaining native substrate binding at near original levels (see Figure 4.1). The resistance problem is therefore similar to that of modeling binding selectivity in that both problems evaluate the binding preferences of the protein receptor for one molecule over another (*i.e.*, substrate vs. drug) (Huggins et al., 2012; Noble et al., 2004; Ohtaka et al., 2002; Pastor and Cruciani, 1995).

Knowledge of potential resistance mutations, before they are clinically observed, would be very useful. During lead prioritization, this knowledge may direct the research team away from candidates that are most likely to confer resistance. Knowledge of resistance mutation hot-spots would allow pharmaceutical researchers to favor leads that avoid interactions with these problematic regions of the active site. Furthermore, this approach could complement the idea of respecting the substrate envelope in lead optimization (Altman et al., 2008; Nalam et al., 2010; Nalam et al., 2013; Shen et al., 2013) in guiding medicinal chemists toward modifications aimed at evading resistance. In the clinical setting, knowledge of potential resistance mutations could allow the development of treatment regimens, with drug cocktails likely to maximize treatment efficacy.

Methods for modeling and identifying known resistance mutations are slowly emerging. These methods can be divided into two categories: sequence-based methods and structure-based methods. Sequence-based methods currently include both computational and wetlab/clinical analyses. The bulk of sequence-based methods, both computational and experimental, are knowledge based and make use of existing sequence and phenotype data to generate and score potentially resistant sequences. Among the sequence-based methods are the genotypic resistance assays (GRTs) that are primarily used to identify resistant strains in a clinical setting (Eboumbou Moukoko et al., 2009; Operario et al., 2010; Van Laethem et al., 2005). GRTs predictions are based on previously identified molecular markers of resistance. Existing sequence-based computational approaches closely mimic the GRT resistance assays. These approaches employ machine learning and statistical methods such as neural networks and random forests (Buendia et al., 2009; Chen et al., 2009a; Fjell et al., 2009; Heider et al., 2010; Pasomsub et al., 2010; Zhang et al., 2010). Genetic features indicative of resistance are identified by analyzing known sensitive and resistant sequences. The presence or absence of these features are used to detect resistance in a candidate sequence. These methods are useful in identifying known or combinations of known patterns of resistance; however, these methods are not useful in identifying novel resistance mutations. For systems where knowledge of previously known mutations is small or non-existent, such as in emerging diseases and new drug targets, the utility of these knowledge-based methods is significantly reduced.

A second class of methods attempts to model the structural effects of known resistance mutations using molecular modeling and molecular dynamics simulations (Dixit et al., 2009; Frieboes

et al., 2009; Lapins and Wikberg, 2009; Pricl et al., 2005; Velazquez-Campoy et al., 2003; Wahab et al., 2009; Zhu et al., 2009). For example, Chen *et. al.* used a docking algorithm to study resistance in several drug targets (Chen et al., 2001). They introduced specific, known resistance mutations into the protein target and used docking algorithms to study the effects of these mutations on drug and substrate binding. These molecular modeling approaches have been useful in understanding the structural basis of known resistance mutations, but these algorithms are not designed to search the combinatorial number of potential mutations to identify novel resistant sequences.

Since resistance mutations typically involve a small number of amino acid changes to the active site, it may be possible to predict new resistance mutations in the drug targets before they arise. This involves searching through and ranking a large number of possible candidate solutions. An exciting new direction is the use of protein redesign algorithms to identify novel resistant sequences using first principles (*i.e.,* without the use of known resistance data). The techniques utilized in protein redesign are extremely efficient at searching exponentially large search spaces and generating a ranked list of candidate redesigns. This approach does not rely on clinical data and can be useful for systems where very little is known about possible resistance. For example, the Donald lab successfully used the K^* ensemble-based protein redesign algorithm (Lilien et al., 2004) to identify novel resistance mutations in dihydrofolate reductase (Frey et al., 2010). Their work demonstrates that computational methods can be useful when modeling *a priori* drug resistance. For a detailed description of previous works in the field, see Chapter 2.

In this chapter, a general framework to probe active site localized resistance mutations is described. Our approach uses the restricted Dead-End Elimination (rDEE) based protein design described in Chapter 3 (Safi and Lilien, 2010), coupled with a two-pass search and scoring method based on the more biophysically accurate MM-PBSA model. The use of DEE based methods makes this approach deterministic, fast, and guaranteed to identify the lowest energy solution from the specified search space. DEE was initially described by (Desmet et al., 1992) and has been successfully employed in several redesigns (Gielens et al., 2007; Looger et al., 2003a; Maglia et al., 2008; Novoa de Armas et al., 2007). In this chapter, the model's ability to predict resistance mutations in four diverse drug target systems will be tested. Hence, these four systems serve as a validation set for the model. For each experiment, the algorithm had no foreknowledge of known resistance mutations. In all cases, the model's predicted mutations have good agreement with the known mutations. Therefore, it can be concluded that the use of protein redesign methods, including rDEE, has significant potential to identify previously unseen resistance mutations in a range of drug targets.

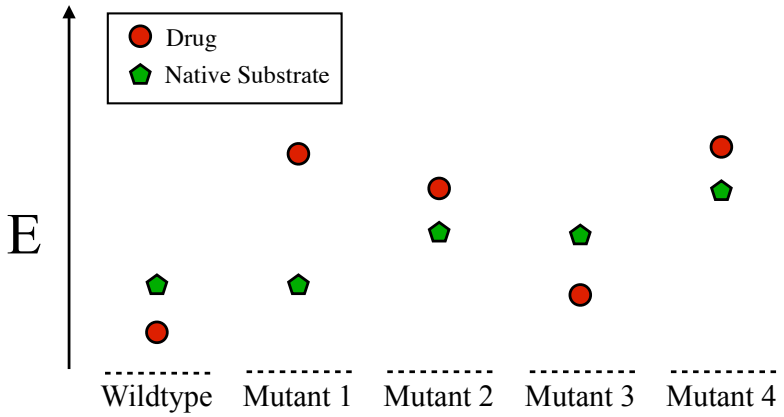

Figure 4.1: **Binding free energy shifts of mutant sequences.** Hypothetical binding profiles for the native substrate (green polygon) and the drug (red circle). The wild-type protein binds the drug more strongly than the native substrate and is therefore sensitive to the drug. Mutant 1 represents the ideal resistant case; the protein's interaction with the native substrate is no worse than that of the wild type yet binding of the drug is significantly impaired. Mutant 2 represents the more realistic resistant case where both native substrate and drug binding are affected. Mutant 3 preferentially binds the drug over the native substrate and therefore remains sensitive to the inhibitor. Mutant 4 prefers to bind the native substrate; however, the significant decrease in binding energy may result in impaired native function and thus a constitutively inhibited protein.

4.2 Methods

We define the term 'mutation sequence' to refer to the sequence of amino acids of a mutated gene variant. Our method consists of two stages. The first stage uses an efficient restricted Dead-End Elimination (rDEE) (Safi and Lilien, 2010) search of allowable mutation space to identify potential resistance mutations. In the second stage, the more accurate yet computationally expensive MM-PBSA scoring method is used to validate and rank the identified mutation sequences.

4.2.1 Stage 1: Efficient Dead End Elimination Based Search

In the first stage, we use a Dead-End Elimination based search to identify mutation sequences that disrupt drug binding while maintaining sufficient binding of the native substrate. To enforce this constraint, a two-pass search procedure was utilized: the *native substrate pass* and the *drug pass*. To improve computational efficiency, our method utilizes restricted Dead-End elimination (rDEE) in combination with restricted A^* enumeration. These methods are fully described elsewhere (Safi and Lilien, 2010). Protein conformations are scored using an efficient pairwise energy function that includes the dihedral, electrostatic and vdW terms of the AMBER energy function (Case et al., 2005; Pearlman et al., 1995; Cheatham and Young, 2001; Ponder and Case, 2003). A flowchart

depicting the approach appears in Figure 4.2.

Our use of DEE closely parallels its use in a typical DEE-based protein redesign. Each of the protein's residues is modeled as either rigid and not mutable, flexible and not mutable, or flexible and mutable. Side-chain flexibility is modeled using a discrete rotamer library of low-energy conformations (Lovell et al., 2000b). Rigid residues include those sufficiently far from the region of interest such that that the rigid approximation is not likely to significantly affect the model. A subset of the n active site residues are modeled as flexible and mutable while the remaining active site residues are modeled as flexible and not mutable. In the context of modeling drug resistance, it is extremely unlikely that a large number of active site residues will mutate simultaneously. In other words, for most drug-target systems, we expect a resistant sequence to contain k mutations where $k < n$ and typically $k \leq 4$. Since the number of mutations in a resistant phenotype is often much smaller than the number of residues in the active site and because traditional DEE algorithms do not have a mechanism for restricting the search to any of the $\binom{n}{k}$ allowed solutions, traditional DEE algorithms are not an efficient choice for resistance prediction. To date, restricted redesign has been accomplished either by running $\binom{n}{k}$ separate searches where in each search only k specific residues are allowed to mutate or by using a single run where solutions are enumerated until a sequence with only k mutations is generated. In the latter case, care must be taken to avoid pruning the desired solution. Recently, we described a restricted version of DEE (rDEE) and A^* specifically tailored to facilitate the search for redesigns with a limited number of allowed mutations (Safi and Lilien, 2010). An rDEE A^* search is therefore an ideal choice for predicting resistance mutations. It removes the need to perform $\binom{n}{k}$ separate searches and is faster than searching through the results of an unrestricted run.

Native Substrate Pass: The search for mutation sequences capable of maintaining native substrate binding is termed *positive* design. A gap free list of mutation sequences ranked on their predicted ability to bind the native substrate is identified using rDEE and restricted A^* enumeration. Mutation sequences whose predicted native substrate binding energies are no more than 1.5 kcal/mol worse than the wild type were identified. Mutation sequences whose predicted native substrate binding energies were worse than this threshold were eliminated from consideration. We emphasize that this threshold is a system parameter that the user can use to control the number of mutant sequences that are further evaluated. We chose a threshold of 1.5 kcal/mol for all reported experiments to allow a compromise between computational efficiency and percentage of the mutation search space evaluated beyond the native substrate pass. However, the user may modify this threshold when evaluating different protein-drug systems.

Drug Pass: The search for mutation sequences with reduced drug binding is termed *negative* design. The output of the native substrate pass is a set of sequences S predicted to bind the native substrate no worse than 1.5 kcal/mol of the wild type sequence. In the drug pass, all sequences in

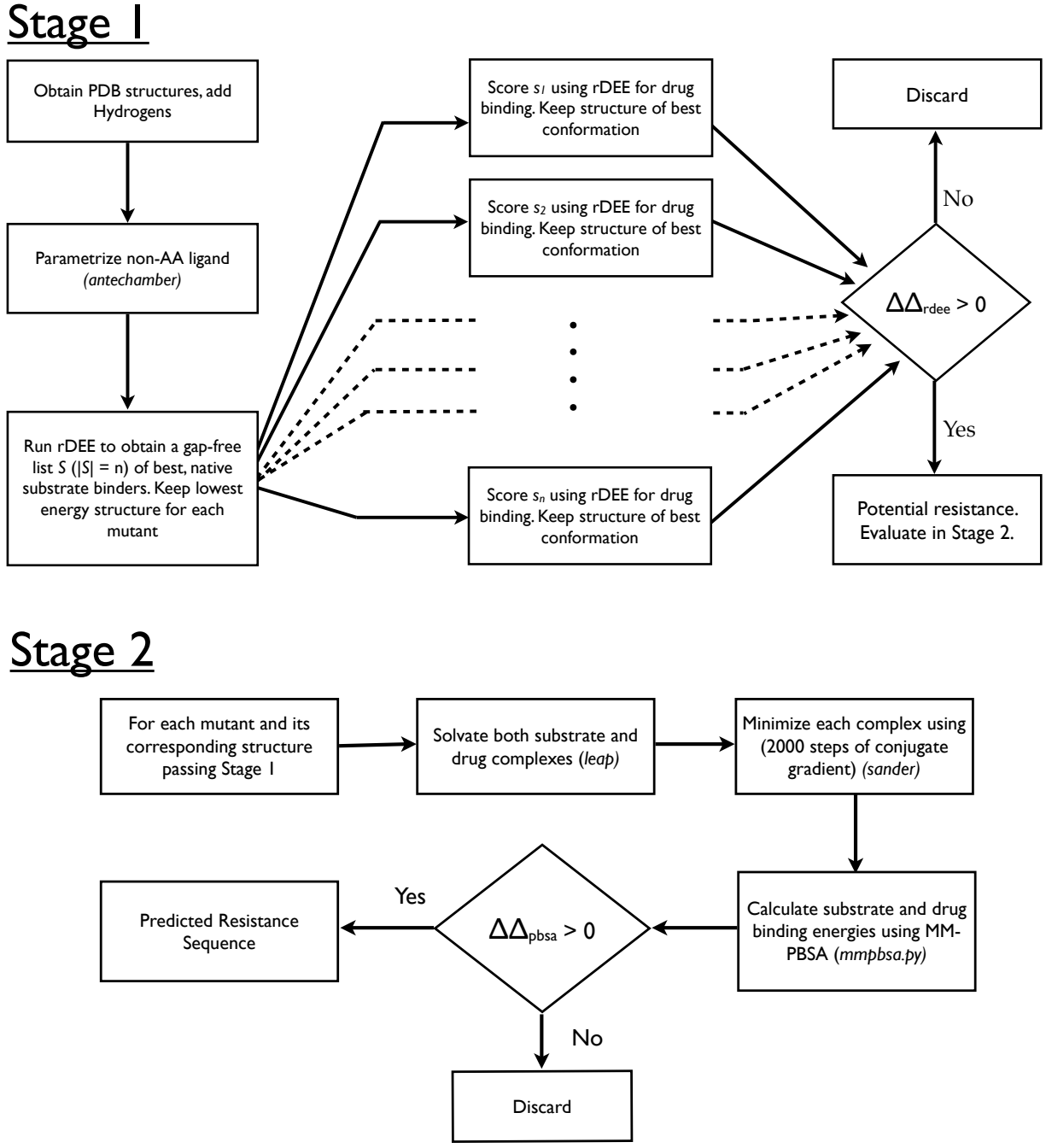

Figure 4.2: **Flowchart of Methods.** The two stage approach is displayed. (A) Stage 1. DEE is used to search and score potential mutants. Complex structures with both substrate and drug corresponding to lowest energy conformation for each selected mutant are generated. (B) Stage 2. Mutants that pass Stage 1, are solvated and energy minimized. A PBSA based approach is used to recalculate binding energies.

S are screened for drug binding. Individual DEE searches are used to identify the lowest energy conformation for each mutation sequence $s \in S$. In these DEE searches, residues are modeled as flexible or rigid, but because the mutations away from the wild type are implicit in s, no further mutations are permitted. The best binding energy between the drug and any conformation of s, as computed by the AMBER energy function (see Additional Modeling Details), is saved as the drug-interaction energy of sequence s. Finally, for each of these mutation sequences, the best interaction energy with the native substrate and the best interaction energy with the drug are used to compute a $\Delta\Delta_{\text{rdee}}$ score (see Scoring Function below).

Additional DEE Search Details: For both the positive and negative design passes, protein systems consist of all residues with at least one atom within 10 Å of the native substrate. The subset of these residues, which together compose the active site, are modeled as both flexible and mutable. These active site residues are identified from previous structural studies (Cowan-Jacob et al., 2007; Dias et al., 2007; Golovin and Henrick, 2008; Schweitzer et al., 1989; Volpato et al., 2009) and PDBeMotif (Protein Data Bank Europe,). Residues that are included (*i.e.*, those within 10 Å) but are not part of the active site form a steric shell that is modeled as rigid and non-mutable during the DEE based search.

Finally, while stochastic methods of protein redesign offer efficient ways of probing the search space, our choice of a DEE based method for resistance modelling is motivated by a desire to maximize coverage. Ideally, a resistance prediction algorithm should aim to predict all resistance conferring mutants for a given drug target protein and drug. DEE allows a systematic method of searching the conformational space where all solutions within a user-specified Δ of the optimal solution are returned. Thus, by altering this Δ, a user can control the percentage of search space near the optimal solution (and hence, number of potential candidates) that is returned, and then further examined for resistance conferring abilities. Furthermore, any mutant not returned by DEE is guaranteed to have an energy difference greater than Δ from the optimal.

4.2.2 Stage 2: Rescoring with MM-PBSA

The first stage serves as a filter to identify candidate resistant mutation sequences and their corresponding low energy structures. In the second stage, these candidates are rescored using the more accurate and more computationally expensive MM-XBSA models as implemented in AMBER 11.0 (Case et al., 2005; Pearlman et al., 1995) (XBSA refers to either the PBSA or GBSA methods). While molecular modeling approaches are far from perfect (Pearlman, 2005), both the MM-PBSA and the MM-GBSA techniques have demonstrated reasonable accuracy on a number of different test systems (Ferrari et al., 2007; Guimaraes and Cardozo, 2008; Raju et al., 2010; Wang et al., 2006). For the results reported in this chapter, we utilized MM-PBSA as it is typically more accurate than MM-GBSA at the cost of added runtime (Hou et al., 2011). We performed

experiments using both MM-PBSA and MM-GBSA in Stage 2 and found no significant differences in results; however, MM-PBSA takes approximately four times longer to compute. Therefore, if optimizing runtime is crucial, MM-GBSA could be used instead of MM-PBSA in Stage 2.

In Stage 2, we rescore each mutation sequence by considering the two structures generated in Stage 1: the lowest energy (best) conformation of the protein bound with the drug and the lowest energy conformation of the protein bound with the native substrate. The rescoring process utilizes all protein residues, including those beyond the steric shell and active site. There are three steps to this rescoring.

First, the mutant structure generated by rDEE in Stage 1 is parametrized using the *leap* module in AMBER 11.0, and solvated in an octahedral box of TIP3P water molecules extending 12 Å beyond the protein on all sides. Second, the solvated complexes are minimized with 500 steps of steepest descent minimization followed by 1,500 steps of conjugate-gradient minimization. No restraints are used during minimization. A residue-based cutoff of 12 Å for nonbonded terms is used. For comparison, we also evaluated longer minimizations with 5,000 and 10,000 steps; however, in all four protein systems, 2,000 steps (500 + 1,500) was sufficient for the minimization to converge. Finally, we compute the ligand binding energy of the solvated, minimized complex. Poisson-Boltzmann (PBSA) implementation of mmpbsa.py module in AMBER 11.0 is used to compute the ligand binding energy. The overall workflow evaluates binding energies using AMBER PBSA, an explicit solvent model, and energy minimization (but no molecular dynamics). This pipeline was introduced by Ferrari *et al* (Ferrari et al., 2007). The authors report promising correlation (∼0.8) between experimentally determined energies and those predicted by AMBER PBSA using explicit solvent for a set of aldose inhibitors.

Using the bound structures generated from Stage 1, each mutation sequence is scored for both native substrate and drug binding. The $\Delta\Delta_{\text{pbsa}}$ score is calculated (see below). Mutation sequences with positive values for both $\Delta\Delta$ scores ($\Delta\Delta_{\text{rdee}}$ and $\Delta\Delta_{\text{pbsa}}$) are categorized as the predicted resistant mutation sequences. A diagram of the process is shown in Figure 4.2.

Scoring Function: We defined a scoring function to measure resistance. The scoring function computes the difference in binding energy between the wild-type and mutation sequence for both the native substrate and the drug. An 'ideal' resistant mutation sequence would maintain native interaction energy with the native substrate while decreasing the protein's interaction energy with the drug (Figure 4.1). The improvement in binding energy for the native substrate is measured as $\text{E}_{\text{wt,s}} - \text{E}_{\text{mut,s}}$ and the decrease in binding energy for the drug is measured as $\text{E}_{\text{wt,d}} - \text{E}_{\text{mut,d}}$. Where $\text{E}_{x,y}$ is the interaction energy between protein x (wt: wild-type; mut: mutation sequence)

and molecule y (s, native substrate; d, drug). The terms are combined to create a resistance score,

$$\Delta\Delta = (\mathrm{E_{wt,s}} - \mathrm{E_{mut,s}}) - (\mathrm{E_{wt,d}} - \mathrm{E_{mut,d}}) \tag{4.1}$$

Two resistance scores can be computed, one using the interaction energies from Stage 1 and one from the MM-PBSA based energies of Stage 2. The method used to calculate the $\Delta\Delta$ score is indicated in the subscript, thus scores calculated by rDEE and PBSA are referred to as $\Delta\Delta_{\mathrm{rdee}}$ and $\Delta\Delta_{\mathrm{pbsa}}$ respectively.

A positive $\Delta\Delta$ score indicates a possible resistance mutation. A positive score is obtained when a mutation sequence disrupts drug binding more than it disrupts binding the native substrate. A negative score indicates the opposite and is thus unlikely to confer resistance. The scoring function of Eq. 4.1 is intuitively similar to the scoring functions used in a number of published binding selectivity studies (Cheng et al., 2010; Kangas and Tidor, 2000; Sherman and Tidor, 2008).

In summary, a series of filters and scoring methods are applied to each molecular system. First, a positive design pass applies rDEE to the protein and native substrate system. Sequences with sufficiently 'good' energies are then evaluated in a negative design pass to assess drug binding and compute a $\Delta\Delta_{\mathrm{rdee}}$ resistance score. Mutation sequences with a positive resistance score are reevaluated in Stage 2 using MM-PBSA. Mutation sequences with both positive $\Delta\Delta_{\mathrm{rdee}}$ and $\Delta\Delta_{\mathrm{pbsa}}$ scores are output as predicted resistant mutation sequences. All mutations that do not pass Stage 1 or that have a negative $\Delta\Delta_{\mathrm{pbsa}}$ score are considered sensitive to the drug.

Molecular Systems: For all four systems, only the active site residues (specified for each system below) were modeled as flexible and mutable. For rDEE runs, a rigid steric shell consisting of residues with at least one atom within 10 Å of the active site (and not explicitly modeled as active site) was included in energy computation. In Stage 2, all protein residues were included in calculations.

Isoniazid-enoyl-ACP Reductase: Isoniazid is a competitive inhibitor of Mycobacterium tuberculosis enoyl-ACP reductase. We utilized the structures from PDB IDs: 2IDZ (inhibitor bound complex) and 1BVR (native substrate bound complex). Similar to previous work (Dias et al., 2007), the following active site residues were modeled as flexible/mutable: Ile 16, Ile 21, Phe 41, Ile 47, Ser 94, Phe 149, Lys 165, Leu 218, Trp 222. Enoyl-ACP reductase binds NADH and a fatty acyl substrate for its native function. The previously known resistance mutations for isoniazid are clustered within the NADH binding site (Dias et al., 2007).

Ritonavir-HIV Protease: Ritonavir is a protease inhibitor used against HIV. We utilized the structures from PDB IDs: 1N49 (inhibitor bound complex) and 1F7A (native substrate bound complex). The 1F7A structure contains the D25N amino acid substitution but is often used as the native form of HIV protease in structure-based modeling studies (Altman et al., 2007; King et al.,

2004). Similar to previous work (Prabu-Jeyabalan et al., 2000; Weber et al., 1989), the following 11 active site residues were modeled as flexible/mutable: Gly 27, Asp 29, Asp 30, Met 46, Gly 48, Ile 50, Ile 54, Val 82, Ile 84, Gly 126, Ile 146. Mutations in the active site catalytic loop, Asp 25, Thr 26, and Gly 27, are known to adversely impact HIV protease function; therefore, these residues were not allowed to mutate during Stage 1.

Methotrexate-DHFR: Methotrexate is an anti-cancer drug targeting human dihydrofolate reductase (hDHFR). We utilized the structures from PDB IDs: 3EIG (inhibitor bound complex) and 1DRF (native substrate bound complex). The structure 3EIG contains two active site amino acid substitutions (F31R and Q35E). We reverted these substitutions back to wild type using rDEE, followed by 5,000 steps of steepest descent unrestrained energy minimization using the *sander* module in AMBER. The resulting structure was used as our wild-type inhibitor bound complex. Similar to previous work (Schweitzer et al., 1989; Volpato et al., 2009), the following 10 human DHFR residues were modeled as flexible/mutable: Ile 7, Leu 22, Glu 30, Phe 31, Arg 32, Phe 34, Gln 35, Leu 67, Val 115, Thr 136.

Gleevec-ABL Kinase: Gleevec is an anti-cancer drug that inhibits human ABL kinase. We utilized the structure from PDB ID: 2HYY (inhibitor bound complex) to model drug binding. The gleevec-ABL kinase system presented a modeling challenge as a structure of ABL kinase bound to its native substrate has not been published in the PDB database. In place of the missing structure, we generated an unbound wild-type receptor by removing the inhibitor from 2HYY. The resulting PDB structure then was solvated in *leap* and subjected to 5,000 steps of steepest descent unrestrained energy minimization using the *sander* module in AMBER. The resulting unbound ABL kinase structure was used in the native substrate rDEE pass of Stage 1. As the substrate bound complex was missing, the value $(E_{wt,d} - E_{mut,d})$ was used to determine if the mutant disrupted drug binding compared to the wild type (see details below). Similar to previous work (Cowan-Jacob et al., 2007; Protein Data Bank Europe,), the following 14 residues were modeled as flexible/mutable: Tyr 253, Val 256, Lys 271, Glu 286, Met 290, Ile313, Thr 315, Phe 317, Met 318, Ile 360, His 361, Leu 370, Asp 381, Phe 382.

Additional Modeling Details: As a first step to evaluate our search pipeline, we limited our search to only hydrophobic amino acids (or hydrophobic plus polar neutral for enoyl-ACP reductase); and we evaluated our ability to recover known resistance mutations involving these amino acids in our search space. The approach we took allows us to decouple the performance of our resistance scoring from the individual scoring of molecular interactions of difficult to model residues and provided the best opportunity for validation using known resistance mutations. Residue flexibility in rDEE was modeled using the Lovell, Richardson, and Richardson's side chain rotamer library (Lovell et al., 2000b). Similar to (Georgiev et al., 2006a; Stevens et al., 2006), AMBER energy

Table 4.1: **Predicted Resistance for Isoniazid-TB.** All 16 single mutants predicted resistant by our model are listed. Of the 5 known mutants, 4 were predicted as resistant by our approach. Another 6 of the predicted 16 are highly likely due to their similarity to known mutants.

Mutation	Comments
I16T, I21T, I21V, I47T	Known resistance mutations
I16V,I21A, I21W, I21F, I21Y, I47V	Plausible, similar to known mutations
K165M, K165Q, F149N	Unlikely, K165, F149 mutants may disrupt function (catalytic triad)
F41M, F41L, L218Y	Less likely, F41 is important for NADH binding

function (a sum of electrostatic, vdW, and dihedral energy terms calculated using the AMBER force field) (Weiner et al., 1984b; Cornell et al., 1995b) was used to compute the pairwise energies between residues. The *reduce* module in AMBER 11.0 was used to protonate the input structures in a neutral environment. In the rDEE stage, small molecule ligands were treated as rigid. Finally, the Lovell rotamer library was used to model the peptide ligand of HIV protease.

4.3 Results And Discussion

We used our two stage search and scoring method to predict resistance mutations in four target systems. Three of the four searches started with experimentally determined structures of the substrate and drug bound wild-type complexes (the gleevec-ABL kinase system used only an experimental structure of the drug bound wild-type complex). None of the four searches used knowledge of previously published resistance mutations. The goal of these experiments was to demonstrate the ability of our approach to discover resistance mutations. We evaluated the quality of the identified candidate resistance sequences by comparison to known resistance mutations. This validation is only partial; a predicted mutation may indeed be resistant, but to date, it may not have been experimentally verified nor reported in the literature.

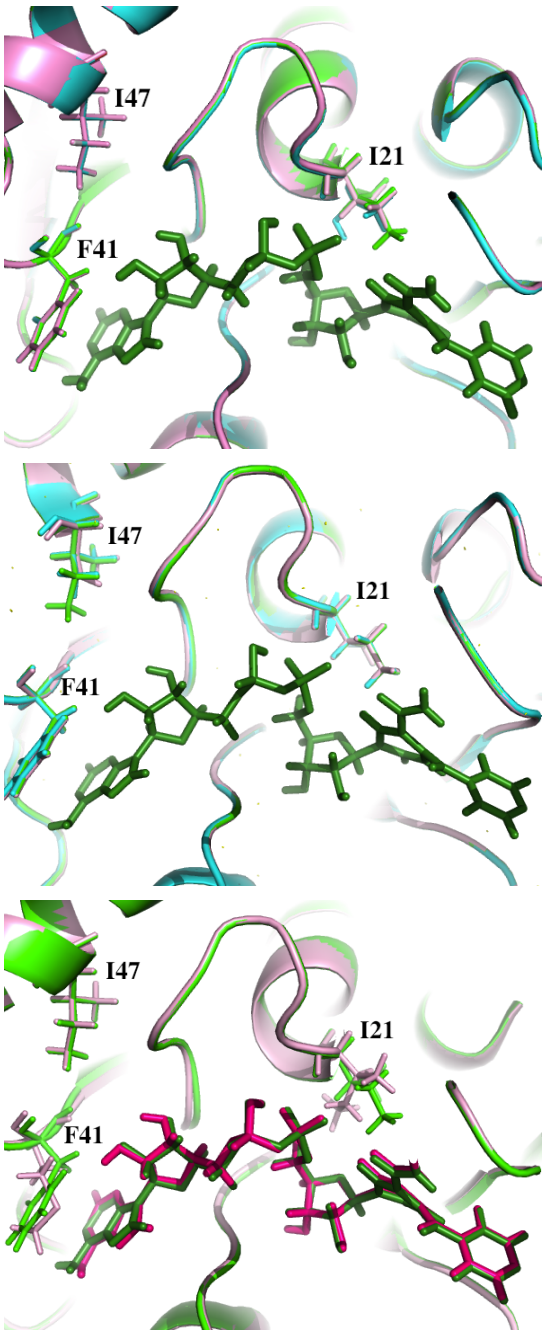

Figure 4.3: **Structures for isoniazid Resistant Mutations.** The Enoyl-ACP reductase protein is shown as cartoon with selected residues and isoniazid rendered in stick form. (Top) Two mutations occurring at Ile 21. Wildtype sequence (green), I21V (pink), I21T (cyan). Both mutations are known and drug binding is predicted to be disrupted by a loss of vdW contacts (∼1.2 kcal/mol) in I21V and a loss of electrostatic interactions (∼6.5 kcal/mol) in I21T. (Center) Two mutations occurring at Ile 47: wild type (green), predicted and plausible I47V mutation (pink), and known I47T (cyan). A loss of electrostatic interactions (∼1 kcal/mol) is predicted to be responsible for the disruption of drug binding. (Bottom) A mutation at Phe 41: wild type (green) and the predicted F41M (pink). Loss of both vdW contacts (∼4 kcal/mol) and electrostatic interactions (∼3 kcal/mol) is predicted. A single isoniazid molecule in dark green is shown in the top and center panels as the drug does not shift significantly between wild-type and mutant structures. In the bottom panel, isoniazid's position in the F41M mutant is shown in pink.

Isoniazid Resistance: Isoniazid remains part of the first line treatment for tuberculosis (TB) worldwide. It is a prodrug activated by the Mycobacterium Tuberculosis's KatG enzyme to form the acyl-NADH complex (INH-NAD) that binds the target enoyl Acyl-Carrier Protein reductase (enoyl-ACP reductase). Enoyl-ACP reductase is a 270 amino acid long protein involved in type II fatty acid synthesis. For wild-type functionality, the TB enoyl-ACP reductase binds a fatty acyl substrate and an NADH cofactor. The NADH and fatty acyl binding are prevented by the competitive inhibitor isoniazid. Numerous point mutations in enoyl-ACP reductase are known to confer isoniazid resistance, these include I16T, I21T, I21V, I47T, V78A, S94A, and I95P (Parikh et al., 1999). Five of these mutations (I16T, I21T, I21V, I47T, S94A) fall within our search space. Our algorithm was used to identify potential resistance mutations in the isoniazid enoyl-ACP reductase system. Most of the isoniazid resistance mutations involve single amino acid changes. We therefore restricted our search to at most one simultaneous mutation. In the Enoyl-ACP reductase system, the 9 active site residues (see Methods) were allowed to mutate to any of the following thirteen amino acid types: Gly, Ala, Leu, Ile, Val, Phe, Tyr, Trp, Met, Asn, Gln, Ser, and Thr. This set includes the nine hydrophobic amino acids A, F, G, I, L, M, V, W, and Y and four polar neutral amino acids N, Q, S, and T. The polar neutral amino acids were included because a number of known resistance mutations involve mutations to Thr.

In this limited system, the search space contains 117 single point mutation sequences. A total of 47 sequences passed the native substrate pass (positive design) of the rDEE search indicating that substrate binding was predicted to lie within acceptable limits. These 47 sequences were then scored for isoniazid binding using rDEE (negative design). The $\Delta\Delta_{\mathrm{rdee}}$ score was calculated for each sequence. Forty six sequences showed a positive $\Delta\Delta_{\mathrm{rdee}}$ score indicating that these candidate mutations affected drug binding more than they affected the binding of the native substrate. Next, for each of these 46 sequences, the MM-PBSA method was used to calculate the binding energies of both the substrate and isoniazid. Sixteen of the initial 117 single point mutants had positive $\Delta\Delta$ scores for both scoring methods and were categorized as our predicted resistant sequences (see Table 4.1).

We evaluated the performance of our algorithm in the context of known resistance mutations in the isoniazid enoyl-ACP reductase system. There are five known resistance mutations among the modeled sequences. Four of these five mutation sequences are identified among our list of sixteen candidate resistant sequences (I16T, I21T, I21V, and I47T). Of the twelve remaining predicted resistant mutation sequences, six occur at positions I21 (I21A, I21W, I21F, and I21Y), I16 (I16V) or I47 (I47V) and can be considered plausible based on their similarity to the known resistance mutations. Of the remaining six mutations, both F41M and F41L are less likely to confer resistance. F41 is hypothesized to be involved in binding the adenine moiety of both the cofactor NADH and

isoniazid. Therefore, a mutation at F41, which disrupts drug binding may also impair native function. Finally, the remaining three mutations K165M, K165Q, and F149N are unlikely to cause resistance, since both F149 and K165 have been implicated in the catalytic triad for the enoyl ACP reductase. For example, experimental evaluation of a number of single mutants of K165, including K165A and K165M indicates that NADH binding is severely affected, inhibiting native function (Parikh et al., 1999). The structures obtained after the energy minimization step of stage 2 of three known resistance mutations (I21V, I21T, and I47T), one plausible mutation (I47V), and one unlikely resistance mutation (F41M) are shown in Figure 4.3.

Of the five known isoniazid resistant mutation sequences, only one, S94A, was not identified by our method. A review of our analysis identifies that S94A was pruned at the native substrate pass. The mutation was predicted to impair native substrate binding by 3.74 kcal/mol, which was more than the allowed threshold of 1.5 kcal/mol. Interestingly, the S94A mutation is known to confer resistance through the loss of water mediated bonds involving Ser (Pantano et al., 2002). The native substrate pass likely had difficulty modeling this interaction because although solvent is modeled in Stage 2, there is no explicit water model in our rDEE phase of Stage 1. In summary, four of five known active site resistance mutations were recovered by our method, while an additional six predicted mutations have plausible mechanisms of resistance. The predictions are significant at 5% (p-value $\approx$0.0018).

Ritonavir Resistance: HIV protease is an aspartyl protease essential for HIV replication. Most protease inhibitors bind the active site and prevent interaction with the native peptide substrate. Our model consisted of HIV protease, the inhibitor ritonavir, and a native binding peptide. We allowed up to two simultaneous mutations at eleven active site residues (see Methods). Residues were allowed to mutate to the following nine hydrophobic amino acids: Gly, Ala, Leu, Ile, Val, Phe, Tyr, Trp, and Met. A total of 787 mutation sequences out of 3771 passed our native substrate filter. 720 of these sequences made it through the drug pass and were rescored using MM-PBSA. 177 mutation sequences had positive $\Delta\Delta_{\text{rdee}}$ and $\Delta\Delta_{\text{pbsa}}$ scores and were output as predicted resistant mutation sequences. These sequences are discussed below and a complete list can be found in Table 5.6. The extremely large amount of HIV protease sequence and screening data complicates the evaluation of our model. To simplify comparison we constructed two validation sets. Our first validation set is derived from the HIV Drug Resistance Database (Rhee et al., 2003; Shafer, 2006). It contains the 28 known single and double residue mutations that lie within our defined search space and confer at least 2.5-fold resistance to ritonavir (Table 4.2). In this context, z-fold resistance indicates that the IC50 of ritonavir for the mutant is z times higher than that for the wild type. This first set of mutations is the best supported set of known resistance mutations in the literature. We refer to the first set as the Gold Standard Set. Because this is

Table 4.2: **Gold Standard Validation Set for Ritonavir Resistance in HIV protease.** The validation set of 28 known single and double point ritonavir resistance conferring mutations obtained from HIV-DB are listed in order of fold resistance. Only the mutants in modeled residues where the fold-resistance was more than 2.5 are included. The prediction column indicates the prediction result of our algorithm (R: predicted resistance sequence, S: predicted sensitive).

Mutation	Prediction	Fold Res.	Comment
M46I/V82A	S	400	Does not pass substrate filter.
V82A/I84V	R	400	
I54A/V82A	R	212	
I54V/I84V	S	201	Does not pass Stage 2. I54V/I84FL predicted resistant.
I54V/V82F	R	128	
I54L/V82A	R	118	
M46I/I84A	S	67	Does not pass substrate filter.
M46I/I84V	S	48	Does not pass substrate filter. M46L/I84V predicted Resistant.
M46L/V82A	R	45	
I54L/I84V	S	29	Does not pass Stage 2. I54L/I84Y is resistant.
I54V/V82A	R	22	
G48V/V82A	S	15	Does not pass substrate filter.
M46I/V82F	S	15	Does not pass substrate filter.
V82A	R	11	
I54M/V82A	R	9.6	
I54V	S	8.8	Does not pass Stage 2.
M46L/I84V	R	8.4	
I50V	S	8.2	Does not pass Stage 2.
V82F	R	8.0	
M46L/V82L	R	5.8	
M46L/I54L	S	5.5	Does not pass Stage 2. M46L/I54M predicted Resistant.
M46I/I50V	S	5.2	Does not pass substrate filter.
I54L	S	5	Does not pass Stage 2.
I84V	R	4.5	
I54M	S	4.4	Does not pass Stage 2.
M46I	S	4	Does not pass substrate filter.
V82L	R	3.1	
M46L	S	2.5	Does not pass Stage 1

a somewhat conservative list we also created a second set of plausible mutations using 17 single residue mutations. These single residue mutations are known to confer resistance to at least one protease inhibitor and are within our search space. They include: D30GY, M46IL, G48MV, I50V, I54LMV, V82AFLM, I84FLV (Brenner et al., 2000; Rhee et al., 2003; Rhee et al., 2010; Shafer, 2006; Stoffler et al., 2002; Wang et al., 2007). In this notation any single amino acid code after the residue number is a valid single point mutation (*i.e.,* D30GY indicates that D30G and D30Y are both separate resistance mutations). Our second validation set consists of the 138 single and double residue mutations that can be constructed using this set of 17 amino acid substitutions. For example, the double mutant V82A/I84L is in the second set of plausible mutations because both V82A and I84L are resistance conferring. In contrast to the first validation set, not all mutation combinations in the second set have been experimentally verified. The constructed combinations are plausible because the constituent mutations are known to display synergistic resistance. We refer to the second validation set as the Plausible Validation Set.

Gold Standard Validation Set. Of the 28 known resistance mutations in the first validation set, 13 are identified by our method as predicted resistant sequences with positive $\Delta\Delta_{\text{rdee}}$ and $\Delta\Delta_{\text{pbsa}}$ scores. This represents an enrichment factor of approximately 10 (28/3771 positive in the entire search space compared with 13/177 positive in the search results, p-value 0 at 5%). The 15 non-identified mutations were pruned at one of the search stages. Six were eliminated in the native substrate pass because the rDEE based prediction of native binding energy was affected by more than the allowed 1.5 kcal/mol. Figure 4.5 shows a recovery plot and ROC curve. At our threshold of 1.5 kcal/mol approximately 75% of the known resistance mutations yet only 18% of the entire search space makes it through the native substrate pass. By relaxing the native substrate binding threshold from 1.5 to 4.0 kcal/mol the numbers are approximately 85% of the known resistance mutations and 25% of the entire search space respectively. Therefore, by modifying this threshold, the user can adjust the number of sequences screened for resistance.

Plausible Validation Set. Of the 138 single and double mutants in the plausible validation set, 40 are identified among our list of candidate mutation sequences. This represents an enrichment factor of 6.2 (138/3771 positive in the entire search space compared with 40/177 positive in the search results). An interesting phenomenon involves secondary or compensatory mutations. We employ a strict native substrate binding cutoff to ensure that only mutants capable of binding the native substrate are considered for resistance. Some of the known single point mutants such as M46I and D30Y do not pass this filter. However, allowing the freedom to incorporate a second and potentially compensatory mutation allows double mutants greater ability to maintain near native binding. Thus, numerous double point mutants that include a known single point resistance mutation pass the substrate filter and are further screened. A total of 89 mutation sequences or 50.3 percent of our 177 output sequences fall into this category.

Seven of these mutants are combined with known mutant D30Y and four include known mutant I50V. Fourteen of the 89 compensatory mutation sequences identified by our search include a combination of V82 and I84 (V82A/I84YM, V82Y/I84LF, V82W/I84LF, V82M/I84Y, V82L/I84Y, V82G/I84FLV, V82I/I84F, and V82F/I84YM). There is experimental support for several protease inhibitor resistance mutations involving this pair (Boden and Markowitz, 1998; Hou and Yu, 2007). A total of 48 of the 177 mutation sequences predicted by our model are neither covered by the Plausible Validation Set nor by the described compensatory phenomenon. These 48 sequences may be false positives or unknown novel true positives. A list of all 177 mutants can be found in Table 5.6.

Methotrexate Resistance: Human DHFR is a frequent chemotherapeutic target. It plays an important role in cell proliferation through its involvement in folic acid metabolism and the production of purines. In its native state, DHFR catalyzes the production of tetrahydrofolate from dihydrofolate and the electron donor NADPH. The chemotherapeutic agent methotrexate (MTX) inhibits cell proliferation by binding DHFR approximately one thousand times more tightly than the native folate. Resistance to methotrexate is widespread and arises both from the upregulation of DHFR and from the introduction of point mutations in the DHFR protein. Unfortunately, it is not clear which amino acid substitutions are primarily responsible for conferring resistance. Instead, methotrexate resistance is associated with a range of single and double amino acid mutations at several active site 'hotspots'.

As in the previous two cases, each active site residue could mutate to the hydrophobic amino acids: Gly, Ala, Leu, Ile, Val, Phe, Tyr, Trp, and Met. We allowed up to two simultaneous mutations at the ten active site residues (see Methods). This defines a search space of 3258 mutation sequences. Resistance conferring mutations are known to occur at DHFR residues Ile 7, Leu 22, Phe 31, Phe 34, Asp 35, and Val 115. Four of these residues, positions 22, 31, 34, and 35, are mutation hotspots where empirical observation suggests that individual mutations at these residues can confer MTX resistance (Ercikan-Abali et al., 1996; Volpato et al., 2007b; Fossati et al., 2008; Volpato et al., 2009). Two-point mutations, where each mutation occurs in these hotspots, are also known to be MTX resistant (Volpato et al., 2007b; Volpato et al., 2009). In the absence of a verified and concise list of known resistant mutations, we define a Plausible Validation Set as the mutation sequences with one or two amino acid substitutions involving only residues 22, 31, 34, and 35. The Plausible Validation Set contains 441 or 13.5% of the search space's 3258 mutation sequences.

Our search produces 272 mutation sequences with a positive $\Delta\Delta_{rdee}$ score. 75 single and double point mutations had positive $\Delta\Delta$ scores for both rDEE and MM-PBSA scoring methods and compose our set of predicted resistant sequences. The Plausible Validation Set contains 18 of these 75 sequences (24%). This represents an approximately two-fold enrichment in identified sequences over their native abundance (13.5%). The enrichment is statistically significant at 5%

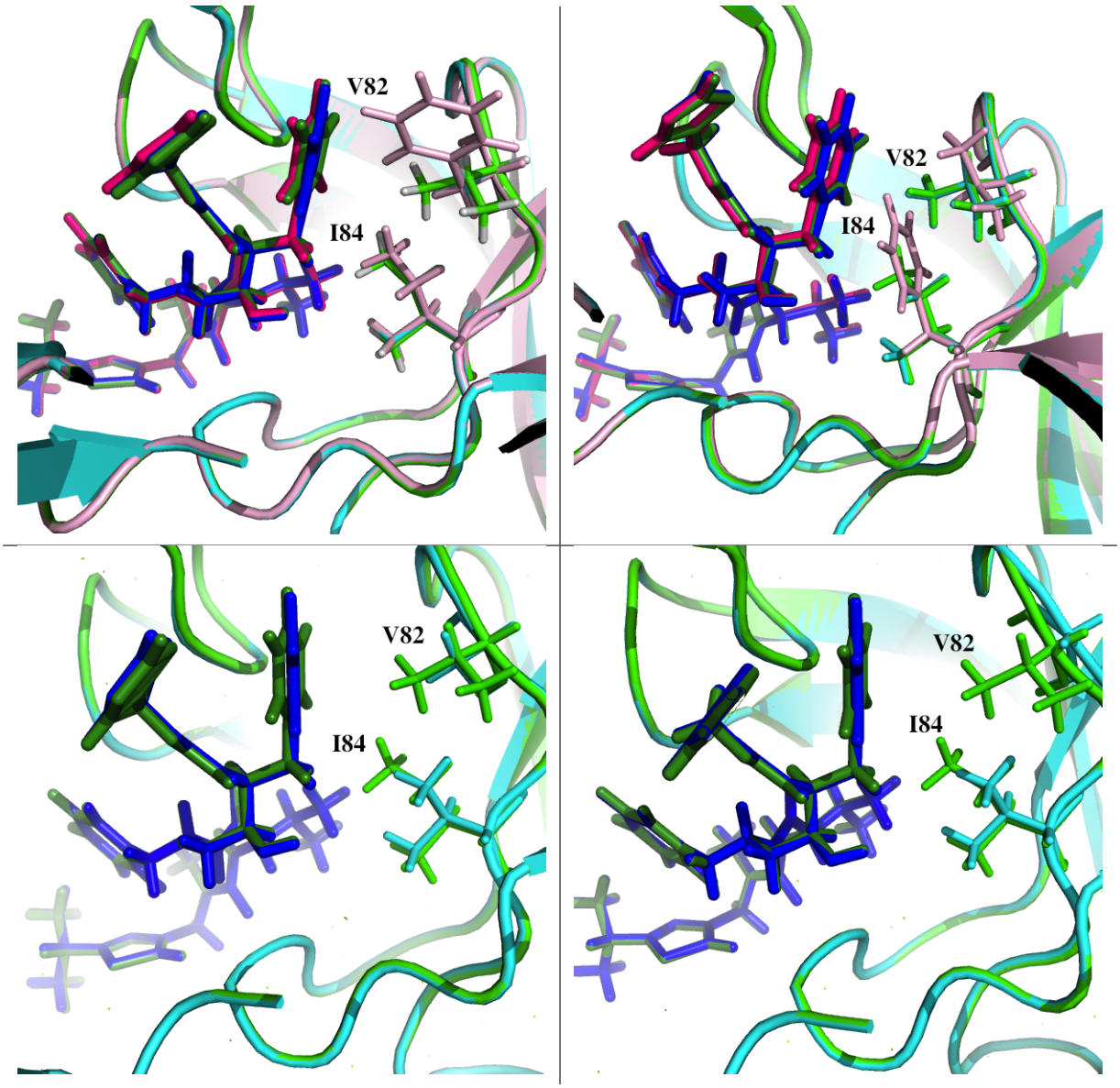

Figure 4.4: **Structures for Ritonavir Resistance Mutants.** HIV protease is shown as a cartoon with selected residues and ritonavir in stick form. In all panels, the mutant structures have been superimposed on the wild type structure (green). In all panels, ritonavir drawn in dark green corresponds to the wild type; otherwise its color reflects the corresponding mutant. (Top Left) Known single point mutants V82A (cyan) and V82F (pink) are displayed. For V82A, loss of vdW interactions (∼1.5 kcal/mol) is predicted to be the cause of disrupted ritonavir binding. Small changes in both vdW and electrostatic interactions are what cause disrupted binding in V82F. (Top Right) Known single point mutants I84V (cyan) and I84F (pink) are displayed. For both mutants loss of vdW interactions (∼1 kcal/mol) is the predicted cause of impaired ritonavir binding. (Bottom Left) The structure of known double mutant V82A/I84V (cyan) is shown. Major loss of vdW (∼2.5 kcal/mol) in the mutant structure along with a small loss of electrostatics (∼1 kcal/mol) is predicted to cause disruption of drug binding. (Bottom Right) The structure of predicted double mutant V82G/I84V (cyan) is shown. Major loss of vdW (∼3 kcal/mol) as well as a small loss of electrostatics (∼1 kcal/mol) is predicted to cause disruption in ritonavir binding.

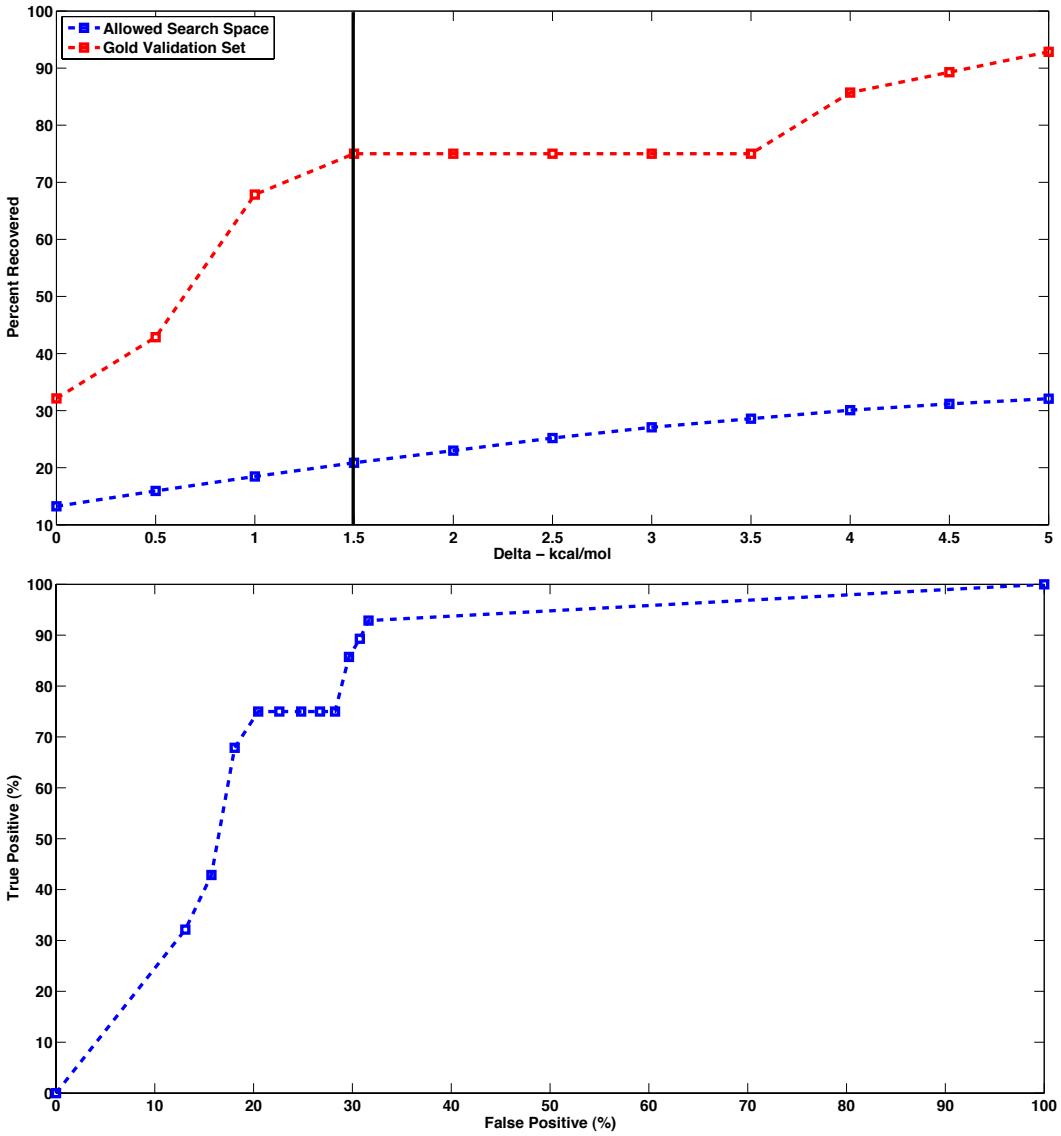

Figure 4.5: **Retrieval of HIV mutants.** (Top) Percent of retrieved known mutants from Gold Validation set (red curve) as well as all the mutants included in the search space (blue curve). The x-axis represents the change in native substrate binding energy of the mutant compared to the wild type. Also shown is the native substrate pass threshold, set to 1.5 kcal/mol from the wild type (vertical black bar). A higher x value indicates a greater loss of binding compared to the wild type. The value at $x = 0$ indicates that these sequences were predicted to have a higher than wild-type affinity for the native substrate in the substrate pass. (Bottom) Percent of true positives (*i.e.*, known mutants from Gold Validation set) is drawn as a function of false positives (*i.e.*, all other mutants from search space).

(p-value ≈ 0.009). A complete list of all 75 predicted mutations appears in Table 4.5. Beyond the mutation hotspots listed above, the profile of MTX resistance becomes a gray area. For example, mutations in Ile 7 are known to pair with hotspot mutations at positions F31, F34, and D35 to produce resistant mutants (Schweitzer et al., 1989; Volpato et al., 2009). Of our 75 predicted resistant sequences, 24 contain a mutation in residue 7. This may be a true positive, or may simply be noise. In summary, our model was able to identify an enriched set of resistance mutations corresponding to known resistance hotspots in the methotrexate- DHFR system.

Gleevec Resistance: In the treatment of chronic myelogenous leukemia (CML), the tyrosine kinase domain of the BCR-ABL fusion protein is inhibited by the highly selective chemotherapeutic drug gleevec. Point mutations in the kinase domain can cause gleevec resistance. The most well known of these mutations is the T315I "gatekeeper mutation" that confers resistance to both gleevec and all second generation tyrosine kinase inhibitors. We conducted a search of single point mutations among the fourteen active site residues (see Methods). As in the previous cases, we allowed residues to mutate to the following nine hydrophobic amino acid types: Gly, Ala, Leu, Ile, Val, Phe, Tyr, Trp, and Met.

Unlike the previous three systems, there is no experimentally determined structure of the protein bound to its native substrate. Therefore, the gleevec system serves as a test of the feasibility of our method to identify potential resistance mutations in the not uncommon case where a structure of the native substrate protein complex is unavailable. Our approach for Stage 1 was to replace the use of the native substrate complex with the unbound apo protein. This approach is somewhat conservative. It will only identify mutation sequences that negatively affect drug binding without affecting the inherent structure of the active site. There is no guarantee that the identified mutation sequences will maintain native binding. The list of predicted resistant sequences is therefore likely to contain an increased number of false positives; but, the sensitivity for identifying resistant sequences should be approximately the same. In the Stage 1 native substrate pass, a gap free list of mutation sequences with energies no worse than 1.5 kcal/mol of the wild-type apo protein is identified. In the Stage 1 drug pass and Stage 2, the $\Delta\Delta$ score is approximated as $\mathrm{E_{mut,d}} - \mathrm{E_{wt,d}}$, where, a positive score indicates a mutant with weaker affinity for the drug. This is the approach taken in the remainder of this section.

There are 117 sequences considered in the mutation search. A total of 32 sequences passed the apo protein stage of the rDEE search (*i.e.,* by not affecting the energy of the apo protein by more than 1.5 kcal/mol). Surviving sequences were evaluated for gleevec binding and a total of 19 sequences had a positive $\Delta\Delta_{\mathrm{rdee}}$ score. Of these 19 mutants, 13 had a positive $(\mathrm{E_{mut,d}} - \mathrm{E_{wt,d}})$ score (Stage 2) and were categorized as our predicted resistant sequences (see Table 4.3).

The set of predicted resistant sequences correctly includes the gatekeeper mutation T315I.

Table 4.3: **Predicted Resistance for Gleevec-ABL Kinase** All 13 single mutants predicted resistant by rDEE and MM-PBSA are given. The clinically well-known T315I gatekeeper mutation is predicted to confer resistance to gleevec by our approach. Two of the predicted mutants are known to be resistant *in vitro* and an additional two are highly likely due to their similarity to known mutants.

Mutation	Comments
T315I	Known gleevec resistance mutation
T315V, T315M	Experimentally confirmed to confer resistance in *in vitro* studies (Corbin et al., 2002; US Patent No. 7326534,)
Y253L, Y253M	Plausible, Y253F causes decreased susceptibility to gleevec
M290L, E286LMY	Unlikely, possible ATP binding site(Corbin et al., 2002)
V256L, L370M, H361M, F382W	Unlikely, residue function is not well-known

In addition, the set also contains two sequences, T315V and T315M, reported to confer gleevec resistance in *in vitro* studies (Corbin et al., 2002; US Patent No. 7326534,). The set also contains four mutations at positions 290 and 286, namely M290L and E286LMY. There is evidence to suggest that both M290 and E286 are possibly involved in ATP binding; therefore, mutations in these residues while impairing gleevec binding, can also result in an inactive kinase (Corbin et al., 2002). This prediction is understandable as our search did not have access to the native substrate complex and therefore only considered gleevec binding. The predictions are statistically significant at 5% (p-value $\approx$0.006).

Ponatinib Resistance: The T315I gatekeeper mutation is the dominant mechanism of tyrosine kinase resistance; however at least one drug, ponatinib, has been shown to overcome this mutation (Zhou et al., 2011). As a short second experiment, we performed a single point mutation resistance analysis for ponatinib. Similar to experimental studies (Zhou et al., 2011), the following nine residues were modeled as flexible: Tyr 253, Glu 286, Thr 315, Phe 317, Met 318, Ile 360, His 361, Asp 381, and Phe 382 (PDB ID: 3OXZ); and each was allowed to mutate to the following set of nine hydrophobic residues: Ala, Phe, Gly, Ile, Leu, Met, Val, Trp, and Tyr. As in the gleevec experiment, a native substrate bound structure of tyrosine kinase was not available; therefore, we predicted resistance using the same modified scoring procedure we utilized for gleevec. Our approach predicts that T315I is indeed sensitive to ponatinib. A total of six single point mutants were categorized as resistant to ponatinib and included: F317LM, D381I, F382ILM. In summary, despite not having an experimental structure of the protein bound to its native substrate, our method was able to identify several known resistance mutations for gleevec, including the T315I gatekeeper mutation. Furthermore, in additional experiments with ponatinib, our method correctly identified T315I as ponatinib sensitive.

Finally, in order to further quantify the effects of a missing substrate bound structure, similar to the case of ABL-kinase, we repeated the Isoniazid-enoyl ACP Reductase experiments without the

bound substrate. A total of 49 sequences passed Stage 1 of the pipeline. Compared to the initial 16 predicted mutants, a total of 21 were predicted to be resistant to isoniazid in absence of the substrate. These included the following: I16V, I21AFGTVW, I47TV, F41LM, S94N, F149LNT, K165MNQ, L218Y, W222MY. As expected, there was a slight increase in the number of false positives when compared to the full experiment.

4.4 Conclusion

This chapter introduced a two-stage structure-based search and scoring procedure for identifying resistance conferring mutations. This technique pairs an efficient restricted Dead-End Elimination based search with the more accurate MM-PBSA scoring method. Positive design ensures that candidate mutations maintain the protein's native function while negative design identifies mutations with significantly reduced affinity for the inhibitor. The output of the two-pass search is an enriched list of possible resistance mutations. It is more efficient to validate this 'short list' than the brute-force approach of testing all possible active site mutations. It is our hope that the type of approach presented in this chapter can provide *a priori* knowledge of drug resistance. The model can also be used to prioritize lead compounds and protein targets based on the ease with which resistance can arise.

Computational methods, like the one presented in this chapter, are most effective if they can maximize the number of true positives while minimizing the number of false positives. In our case, this corresponds to not predicting a resistant mutant as being sensitive. Of course, in the absence of exhaustive knowledge of experimentally verified true positives, it is difficult to completely assess performance. In our testing, we constructed a number of validation sets using known or likely mutation sequences. We demonstrated our technique using four well-known systems. In all four cases, our technique produced a set of predicted mutation sequences enriched with known resistant sequences. Our method was successful in both single point and double point mutation searches. Three of the four systems utilized experimentally determined structures of the protein in complex with the native substrate as well as with the drug. One system did not have an available experimental structure of the native substrate protein complex. For this system, we replaced the missing complex with the structure of the apo (unbound) protein and approximated the $\Delta\Delta$ score. The fact that the method was still able to recover well known resistance mutations suggests that in some cases resistant mutants can be predicted despite having only partial structural information. We also note that individual components of the method presented here have been previously validated in a number of ways. First, dead-end elimination is provably accurate and guarantees to find the global optimum. Second, previous studies have reported a number of successful DEE based re-

designs (Filikov et al., 2002; Gielens et al., 2007; Looger et al., 2003a; Maglia et al., 2008; Novoa de Armas et al., 2007). Plus, AMBER and MMPBSA have been extensively used in literature to study molecular dynamics and binding. Finally, the pipeline used to obtain binding energies using MMPBSA and AMBER in Stage 2 has previously been validated by (Ferrari et al., 2007) and reported to have reasonable correlation with experimental binding energies.

The technique presented here is one step towards a general purpose computational tool. As such, it is not without limitations. Structural models employing approximations of interaction energies should not be interpreted as ground truth. Computational models are generally most useful when used in conjunction with wetlab testing. There are three primary directions for future work. First, we want to refine our ability to gracefully handle the situation where, as in our tyrosine kinase example, an experimental structure of the native substrate protein complex is not available. Second, the ability to model larger scale and allosteric type conformational changes could allow identification of mutations beyond the active site. Finally, it would be useful to close-the-loop by coupling our computational model with feedback from wetlab testing of the predicted mutation sequences. For a detailed discussion and description of possible future work, see Chapter 6.

In conclusion, our results on four diverse drug-target systems indicate that structure-based methods like the one presented in this chapter can be useful in identifying resistance mutations in drug targets. Since no prior knowledge of resistance is needed, our approach can be employed as a first step to probe resistance in systems where information regarding drug resistance is minimal or nonexistent.

Table 4.4: **Predicted Resistance for HIV.** All 177 mutants predicted as resistant by our model.

D29Y/I84V	D29Y/V82FGW	I146GMV
D30A/I146M	D30A/V82LM	D30F/I146V
D30F/I54M	D30G/I146M	D30G/I50FLM
D30G/I54LM	D30G/V82LM	D30L/I146M
D30L/I84Y	D30L/V82M	D30M/I84Y
D30V/I146M	D30W/I146G	D30Y/I146MW
D30Y/I50FLMY	D30Y/I54LM	D30Y/I84LM
D30Y/V82LM	I50A/V82M	I50F/I146G
I50F/I184Y	I50F/V82I	I50G/I146W
I50G/V82L	I54G/I146M	M46L/I84M
I50G/I84LM	I50V/I146W	I50G/I54M
I50V/V82AFLW	I50L/I146VG	I50L/I84Y
I50L/V82FWY	I50V/I146GVW	I54A/V82AM
I54G/V82L	I54L/I84Y	I54L/V82AGWY
I54M/I146GLV	I54M/I84ALVY	I54M/V82AFGLMWY
I54V/I146GV	I54V/I84FL	I54M/I146G
I54M/I84V	I54L/I146G	I54V/V82AFGLWY
I84A/I146M	I84FV	I84F/I146MV
I84L/I146GVL	I84M/I146G	I84V/I146V
I84Y/I146MW	M46F/I54M	M46L/I146V
M46L/154M	M46L/I84VM	M46L/V82AGLM
M46W	M46W/I146MV	M46W/I54MV
M46W/V82AILM	M46Y/I146M	M46Y/I54L
M46Y/I54M	M46Y/I84LM	M46Y/V82LM
V82AFGYL	V82A/I146GMV	V82A/I84LMVY
V82F/I146MV	V82F/I84FLMY	V82G/I146MV
V82G/I84FLMVY	V82I/I146GV	V82I/I84F
V82L/I146GLV	V82L/I84FY	V82L/I146M
M46L/150L	V82M/I146G	V82M/I84Y
V82W/I146M	V82W/I84FLMY	V82Y/I84FLM
V82Y/I146M		

Table 4.5: **Predicted Resistance for DHFR.** All 75 mutants predicted as resistant by our model.

E30M/Q35Y	F31M/Q35F	F31W/F34L
E30M/V115M	F34M/Q35FLY	I7A/F34M
F34M/L67M	F34Y/V115G	I7L/V115M
I7M/F34M	I7M/T136V	I7MV
I7M/E30M	I7M/L22IMV	I7M/L67IM
I7M/Q35LWY	I7M/V115GIM	L22M/F31M
I7VW/L22M	I7W/V115IM	I7V/V115IM
L22M/Q35W	L22M/T136V	L22A/F31Y
L22A/V115M	L22F	L22F/E30LM
L22F/F31LMY	L22F/F34M	L22F/L67M
L22F/Q35Y	L22F/V115IM	L22G/V115M
L22M/F31MY	L22M/L67IV	L22M/Q35LY
L22M/V115AG	L67I/V115M	L67V/V115M
Q35F/V115M	Q35L/V115M	Q35M/V115M
Q35W/V115M	Q35Y	Q35Y/L67M
Q35Y/T136V	Q35Y/V115IM	V115IM
V115M/T136A	L22M/V115I	

Chapter 5

Evolution of Drug Resistance

5.1 Introduction

The structure based methods of Chapter 4 probe binding energy shifts between the native substrate and drug, arising as a result of point mutations, to model resistance in a drug-target system. However, the mechanism through which resistance mutations are acquired is primarily stochastic in nature, namely evolution. Point mutations arise in the wild type DNA, and resistance evolves via these point mutations. Thus, irrespective of how resistant the final mutant is, the path taken from the wild-type to the final mutant holds additional information about resistance. For example, the length of the path from the wild type to a resistance mutation can determine, in part, how likely the mutant is in nature. A mutant that requires fewer mutations can evolve quickly and is more likely to be seen than a mutant that requires a larger number of mutants. In addition, the quality of the paths i.e. the *fitness* of the intervening mutants can be another factor that determines the likelihood of a resistance mutation. Consider, for example, the following scenario where two double point mutants m_1 and m_2 are equally resistant. However, all paths from wild type to m_1 contain an intermediate single point mutant that is susceptible to the drug. Mutation to this single point mutant will be immensely deleterious to the pathogen under drug pressure. Thus, it would be unlikely for the pathogen to sample m_1 since all paths from the wild type to m_1 are *blocked* by drug sensitive mutants. Since the structure-based models of previous chapters do not take into account the mutational mechanism by which the resistance mutation is accumulated, both m_1 and m_2 will be considered equally resistant and plausible by them.

In this chapter, I employ a simple Markov-chain based model to evolution of resistance. This model combines a mutational mechanism at the DNA level with the structural models of Chapter 4. This evolutionary model is then applied to the case of resistance evolution in HIV protease.

As resistance to a single inhibitor develops, treatment is often geared towards combination

therapy where multiple inhibitors are combined to create a drug cocktail (Majori, 2004; Spanagel and Vengeliene, 2013). Combination of multiple drugs can theoretically cover a large section of the available mutational space such that development of resistance is significantly harder than in the case of a single inhibitor. Furthermore, adherence to a drug regimen is often one of the aspects affecting treatment success and evasion of drug resistance as well (Huang et al., 2011; Krakovska and Wahl, 2007a; Rosenbloom et al., 2012; Tam et al., 2008). As patients fail to take the prescribed drugs, the selective pressure exerted by the drug is temporarily lifted. In such a case, the virus might have an opportunity to sample new mutations which were previously prevented under drug pressure. This can create new paths likely to be sampled by evolution such that mutants that were previously unreachable, might be accessible now. Thus, under less-than-perfect drug adherence, the virus has an increased opportunity to mutate and create drug resistant variants. Finally, the evolutionary model presented in this chapter will be applied to probe evolution of drug resistance in HIV protease under drug cocktails and various drug adherence strategies.

5.2 Methods

Evolution can be intuitively considered as a random walk on the genotypic space (mutation space) where DNA mutations cause transitions between different genotypic states (mutants) (McCandlish, 2011). Under drug pressure the ability of each mutant to confer resistance to the drug determines how beneficial the mutant is for the pathogen, if sampled by the mutational process. Thus resistance acts as a measure of natural selection in such an evolutionary process, controlling how likely a mutant is to survive. This evolutionary model is described below.

5.2.1 A Markov-Chain Based Model

To model the underlying DNA mutations, each mutant is represented as its genotype. Thus for each mutant at the amino acid level, there are multiple mutants at the genotype level. For instance, a mutation to phenylalanine at the amino acid level translates to two distinct mutants at the genotype level, one for each coding genotype of phenylalanine: TTT and TTC (see Chapter 1). Each genotype is represented by a single state in the Markov chain (thus for each amino acid mutation there are multiple genotypes). A transition or a bi-directional edge exists between two states a and b if b can be reached by making a single DNA point mutation in a, and vice versa. Since we are working in the genotype space, there are only four point mutations corresponding to the DNA bases: A,T,G and C. All single point DNA mutations are considered equally likely. Hence, all transitions, if one exists, are equally likely (transition probability is 0 otherwise). During the course of simulations, the next state b is chosen according to the distribution of transition

probabilities at the current state a. Thus, in the scenario where all DNA bases are equally likely, the next state b is chosen according to a uniform distribution.

Once a transition is made, the move is accepted or rejected using the probability of fixation p_{fix} described in equation (5.2). The formula in equation (5.2) can be obtained using a diffusion approximation to the evolutionary Wright-Fisher process and is taken from (Fisher, 1930; Wright, 1931; Kimura, 1962). This fixation probability is a function of $s_{i\to j}$, which is a measure of natural selection for genotype j relative to genotype i or its *selection coefficient* relative to i. In this way, by employing relative selection coefficients, the probability of fixation intuitively models whether or not evolution would select for the particular mutant j just chosen by the transition. Thus a mutant j that increases chances of survival i.e. increases resistance compared to the current mutant i, has a higher chance of being selected as compared to a mutant that either marginally improves or reduces chances of survival (Fisher, 1930; Kimura, 1962). It is worth noting that selection coefficients are always pairwise and calculated against or relative to another genotype. Thus selection coefficient is always a measure of fitness for a genotype j relative to another genotype (mutant or wild type). In this work, the notation $s_{i\to j}$ is used to indicate the selection coefficient of genotype j relative to genotype i; whereas s_j represents the selection coefficient for genotype j relative to the wild type. Finally, the terms fitness and s_i are used interchangeably.

Under drug resistance, natural selection can be represented by the degree of resistance of a mutant, and governs how likely it is to be selected for by evolution, once visited. In Chapter 4, binding energies of the mutant with its native substrate and the drug were used to determine if the mutant was resistant or susceptible. Here, these binding energies for a mutant i are used to generate a selection coefficient that measures its resistance conferring potential and represents its relative fitness to the wild type. The following equation follows (Wylie and Shakhnovich, 2011) closely.

$$s_i = e^{(\mathrm{E_{wt,s}}-\mathrm{E_{i,s}})-(\mathrm{E_{wt,d}}-\mathrm{E_{i,d}})} - 1 \tag{5.1}$$

Equation 5.1 assumes that mutations in the gene being modelled determine the overall fitness of the organism. However, when the gene being modelled is an essential gene e.g. HIV protease, this is a reasonable assumption since mutants that affect function of the protease gene are likely to have an effect on overall viral fitness.

The selection coefficients are then converted into a probability of fixation of mutant j (Kimura, 1962), when a mutation in i changes it into j. The intuition is that a beneficial mutant has a much higher probability of being selected/fixed, whereas a deleterious mutant is seldom fixed by evolution.

$$p_{fix} = \frac{1 - e^{-2s_{i\to j}}}{1 - e^{-2N_e s_{i\to j}}} \tag{5.2}$$

where N_e is the effective population size and $s_{i\to j}$ is the relative difference in fitness of mutants j

and i, and is given by:

$$s_{i \to j} = \frac{s_j + 1}{s_i + 1} - 1 \tag{5.3}$$

where s_i and s_j are selection coefficients of i and j relative to the wild type calculated using Equation 5.1. A derivation of Equations 5.1 and 5.3 is provided in the Appendix. An N_e of 1000 was used for simulations reported in this chapter which is in agreement with the known effective population size of HIV under drug pressure (Leigh-Brown, 1997; Drummond et al., 2002; Seo et al., 2002; Achaz et al., 2004; Shriner et al., 2004; Althaus and Bonhoeffer, 2005). For the sake of completion, we also note that simulations were repeated for different values of N_e including 10^4, 10^5, 10^6, 10^7, and similar results were obtained.

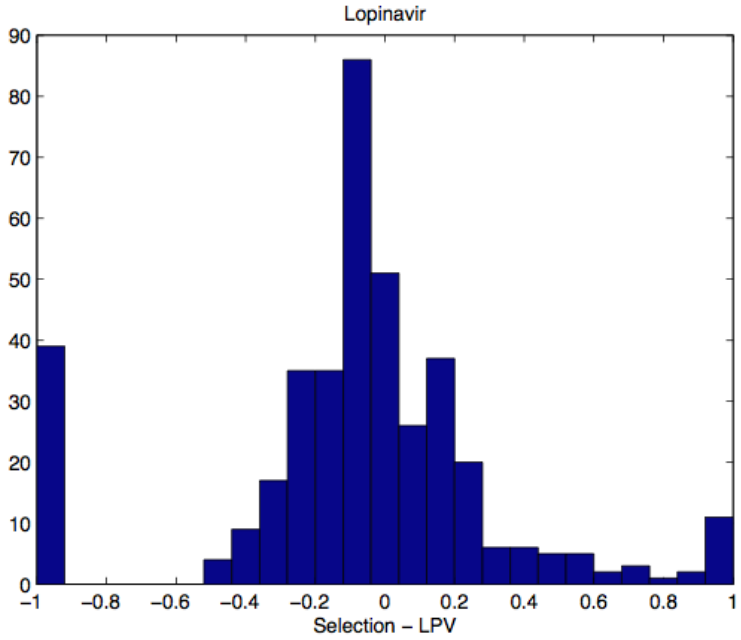

Figure 5.1: **Distribution of Selection Coefficients.** A histogram of selection coefficients for the protease inhibitor lopinavir is presented. Selection coefficients for all possible mutants in the V82-I84 model are shown. The selection coefficients are generated using equation 5.1. A large percentage of mutations is deleterious (selection < 0), and a significant number is lethal (selection = -1).

The evolutionary Markov-chain based method described above was used to simulate evolution in HIV protease. Two sets of experiments were performed: first, on a smaller test-system consisting of only two active site positions for HIV for validation purposes; and second a full scale active site system for HIV protease. Use of first system was intended to help verify our model.

V82-I84 System For this system, we used two active site positions from HIV protease namely I84 and V82. As the total number of nucleotide triplets (codons) is 61, a total of 61*61 or 3721

genotypes were present in this system. Each of these genotypes represented a state in our model. These 3721 genotypes translate to a total of 400 possible mutants that are distinct at the amino acid or protein level. Selection coefficients for all these mutants were calculated using equation 5.1. A histogram of the selection coefficients for protease inhibitor lopinavir is shown in Figure 5.1 as an example. A large number of the mutants are deleterious with selection coefficients less than 0, and a significant number are lethal. This is consistent with the observation that a significant portion of possible mutations do not confer resistance to the drugs.

Figure 5.2 shows a 3D map of this V82-I84 mutational system of HIV protease, both with and without the drug pressure. This map represents a *fitness landscape* (Wright, 1932; Kauffman, 1993; Stadler, 2002; Armstrong and Tidor, 2012). Each point on this fitness landscape represents a genotype (mutant or wild type), and the z-axis represents the selection coefficients or the fitness of that genotype. Each of the 3721 genotypes are represented in this landscape. In the presence of a drug the selection coefficients used were calculated using Equation 5.1. Selection in the absence of drug was calculated as :

$$s_i = e^{(\mathrm{E_{wt,s}} - \mathrm{E_{mut,s}})} - 1 \tag{5.4}$$

The landscape under no drug shows does not show significant positive peaks, indicating that without drug pressure wild type is among the fittest genotypes. This behaviour changes drastically as the drug is introduced and we see fitness peaks for certain mutants indicating that under drug pressure, resistance conferring genotypes have an advantage over the wild type. However, a large number of mutants are still susceptible to the drug and are deleterious. Lastly, the results of simulations on the V82-I84 system under lopinavir are shown in Figure 5.3. We used an ensemble of 10,000 runs and the simulation started in a wild type genotype in each run. A maximum of 1000 steps were allowed. As expected, the virus improves its fitness as it takes steps in the Markov chain. The fitness at step s across n runs was calculated using the following formula:

$$F_s = \sum_m p_m s_m \tag{5.5}$$

where p_m is the fraction of times mutant m was observed in n runs, and s_m is its selection coefficient. For the lopinavir results above, $n = 10,000$.

Full HIV Active Site System The next set of experiments were performed on a larger system consisting of the full active site of HIV protease as modelled in Chapter 4. The data generated by resistance prediction pipeline of Chapter 4 for HIV protease was used. Since each mutant genotype represents a state in our model, the full scale active site system consisting of 11 residues corresponds to a total of 4^{33} distinct genotypes. However, the methods of Chapters 3 and 4

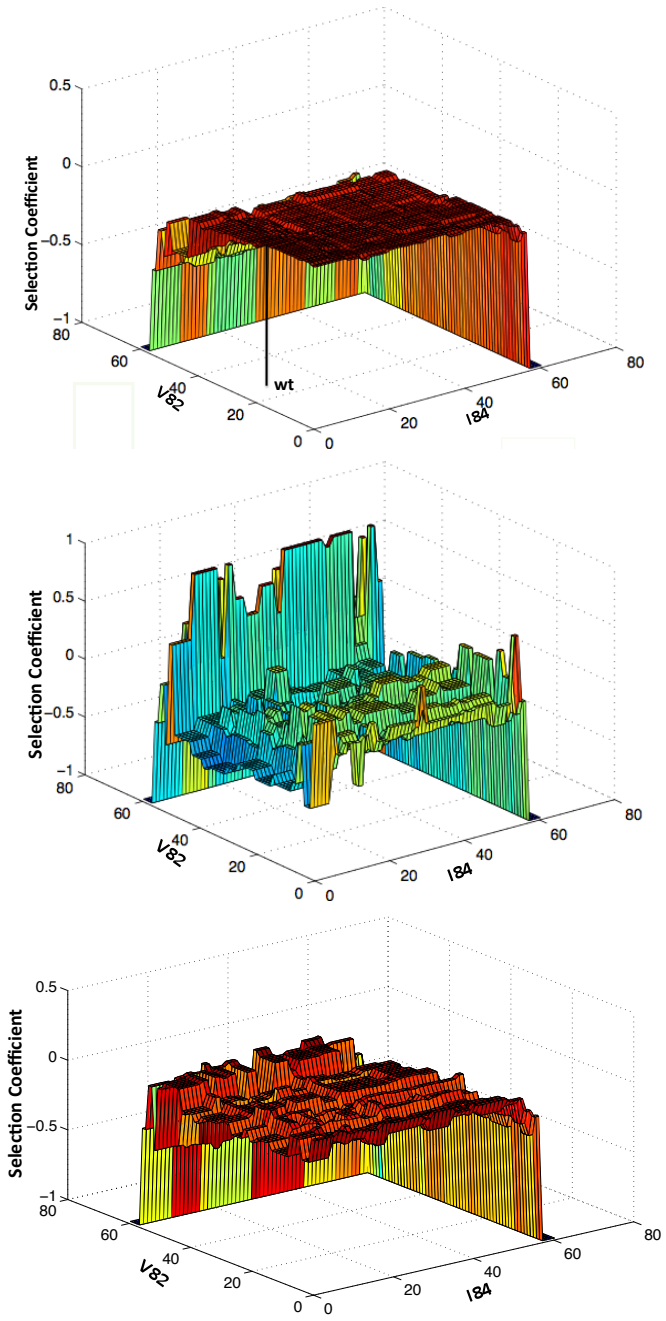

Figure 5.2: **Fitness Landscape** Top: Fitness landscape without the drug pressure is shown. The selection coefficients are calculated using only the substrate. Middle: Fitness landscape under lopinavir. The high selection coefficient ridge represents mutants V82D and V82G; under lopinavir both mutants combine well with mutations at I84 in our model. Bottom: Fitness landscape under ritonavir. Administration of the drugs significantly alters the fitness landscape. All mutants in V82-I84 system are included. All 61 coding codons at each position are included so each point in the landscape represents the genotype of the mutant formed by combining the codon at position V82 with codon at position I84. The low fitness extrema at codons 58-61 in all three landscapes represent prolines.

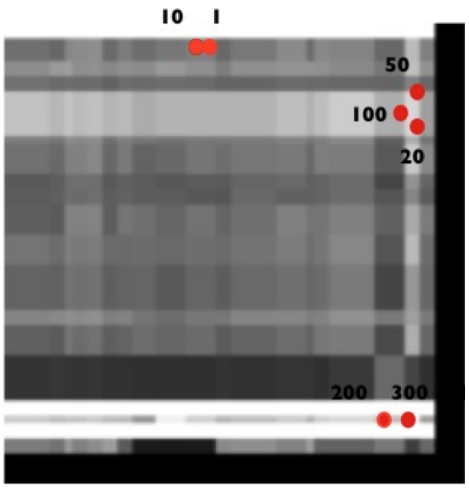

Figure 5.3: **A Walk on Fitness Landscape**. A greyscale representation of the lopinavir landscape is drawn. Each point represents a genotype; white implies a highly beneficial mutant, whereas black is lethal and corresponds to a selection coefficient of -1. The red dots indicate where the virus is at a particular point in simulation. The walk starts at the wild type (selection coefficient of 0;grey) and steadily improves fitness until it reaches the highly fit region around step 200 and stays there for the rest of the simulation.

employ efficient DEE based pruning techniques to quickly eliminate large portions of this space from consideration. This implies that for a large section of the possible mutation space, binding energies are never calculated explicitly. Hence, selection coefficients for these mutants cannot be computed with the available data. However, the structural models only eliminate mutants that are considered deleterious or sensitive to the drug. Thus, these mutants are highly unlikely to be selected by evolution and can possibly be treated as a single object in evolutionary simulations.

To deal with the large, highly deleterious sections of the mutational space, we introduce a low-fitness state in our model. This low-fitness state is assigned a selection coefficient of -1.0, reflecting its highly deleterious nature. A large number of mutant genotypes are collapsed into this single low-fitness state in our full active site model. These low-fitness mutant genotypes include all mutations for which selection coefficients cannot be calculated because they were pruned and not explicitly evaluated by the resistance prediction pipeline described in Chapter 4.

To determine the effects of this approximate low-fitness state on the evolutionary simulations, a low-fitness state was incorporated in the smaller V82-I84 system. Since, selection data for all mutants included in V82-I84 system was available and all mutants were initially modelled explicitly, the effect of a low-fitness state approximation could be evaluated in this system. A number of experiments (see Figure 5.4) were performed on this system, and the results indicate that collapsing of mutant genotypes into a single low-fitness state is a well-tolerated approximation for our evolutionary model.

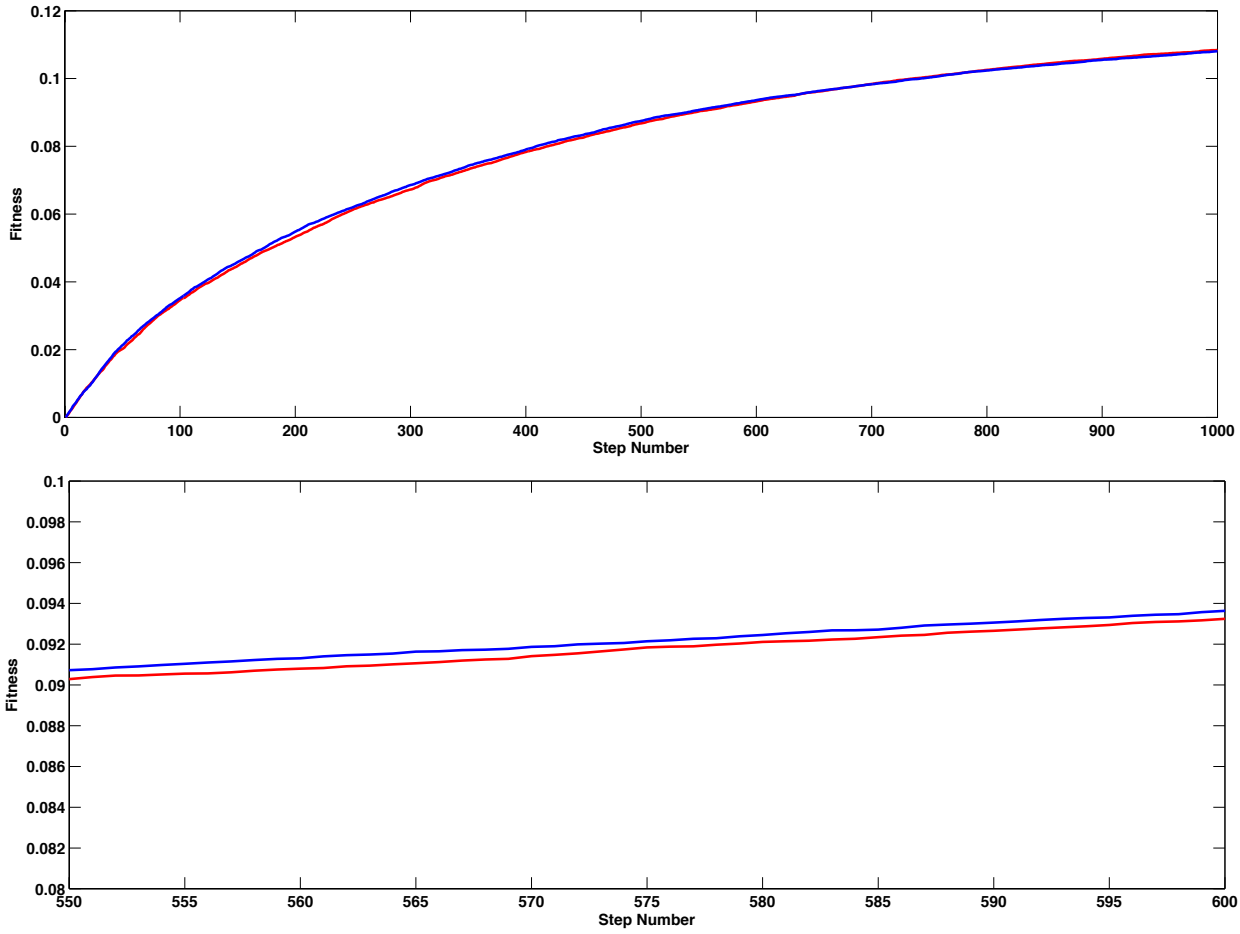

Figure 5.4: **Introduction of a Low Fitness State** Fitness of the virus at each iteration under ritonavir is drawn. An ensemble of 5000 runs of the simulation was used. Each run had 1000 steps. Top: Fitness of the full model (blue) and the low-fitness approximation (red) is displayed. Bottom: A zoomed in view of the Top plot, showing fitness for 50 steps. The blue curve represents the full model and the red is low-fitness approximation as before. All mutant genotypes with selection coefficients below -0.7 were collapsed into a single low-fitness state. The low-fitness approximation behaves similar to the full simulation. A number of other thresholds for low-fitness were also tried (data not shown). Under all thresholds, the low-fitness approximation was well tolerated. Only mutations at two residues, namely V82 and I84, are modelled.

Table 5.1: **Gold Standard Validation Set for Ritonavir Resistance in HIV protease.** The validation set of 28 known single and double point ritonavir resistance conferring mutations obtained from HIV-DB are listed. Only the mutants in modelled residues where the fold-resistance was more than 2.5 are included.

M46I/V82A	V82A/I84V	I54A/V82A
I54V/I84V	I54V/V82F	I54L/V82A
M46I/I84A	M46I/I84V	M46L/V82A
I54L/I84V	I54V/V82A	G48V/V82A
M46I/V82F	V82A	I54M/V82A
I54V	M46L/I84V	I50V
V82F	M46L/V82L	M46L/I54L
M46I/I50V	I54L	I84V
I54M	M46I	V82L
M46L		

5.3 Results

The Markov chain based evolutionary simulations described in Methods were used to observe the evolutionary behaviour of resistance in HIV protease. All results described in this section pertain to the full active site of HIV protease. All simulations started in a wild type genotype. The simulations were allowed to run for a total of 3000 steps, at which point all simulations had converged. An ensemble of 5000 individual simulation runs has been used to report the results unless otherwise noted. We note that we also performed simulations for 4000 and 5000 steps and similar results were obtained. The system under consideration was the full active site of the HIV protease as modelled in Chapter 4. For an active site of 11 residues, a total of 22000 mutant sequences, which are distinct at the amino acid level and have a maximum of two mutations, fall under this search space ($\binom{11}{2}$ * 20 * 20). The methods of Chapter 4 only screen mutants further if the substrate binding of the mutant is not impaired by more than 1.5 kcal/mol compared to the wild type binding. This corresponds to all mutants that pass the native substrate pass of Stage 1. All single and double point mutants that fall under this category were included in the evolutionary simulations (787 mutants). Equation 5.1 was used to determine the selection coefficient of each mutant. The binding energies were derived using MMPBSA similar to Stage 2 of Chapter 4. The resulting landscape had 185 mutants that were resistant to ritonavir i.e. had a selection coefficient greater than 0. A list of these mutants is found in Table 5.6 at the end of this chapter. The rest were sensitive to ritonavir. A selective coefficient of 0 was assigned to the wild type genotypes.

Figure 5.5 displays the fitness landscape under ritonavir used for evolutionary simulations. Each node in the graph is a mutant present in the landscape. For visual simplicity, the landscape is drawn at the amino acid level even though the evolutionary simulations are performed at a genotypic level.

Table 5.2: **Mutual Information between Models of Resistance and HIV DB** Mutual information between different models of resistance and the gold standard dataset from Stanford HIV DB is measured. Structure refers to the structural algorithm presented in Chapter 4 (see Section 5.3), whereas Evolution is the evolutionary model presented in this chapter. As a control, mutual information between Mutation and HIV db is also listed. Mutation here refers to evolution in the absence of selection (see text).

Mutation (alone)	Structure (alone)	Evolution (Structure+Mutation)
0.01	0.22	0.32

There are 787 nodes, each of which represents a mutant of HIV protease. An edge between nodes indicates that it is possible to mutate from one node to the other using a single DNA mutation. All edges are bi-directional. The colours on the nodes indicate whether a mutant is resistant or not, and are a function of the selection coefficients. Red nodes indicate a deleterious mutation (ritonavir sensitive) whereas green ones are resistant and beneficial. All mutants considered resistant by the structural models (and hence with selection coefficients greater than zero) are thus represented as green in this landscape. A list of these mutants is found in Table 5.6. Finally, the low-fitness state is not displayed.

Predicting Resistance using Evolution The structural models of Chapter 4 treat resistance as an interplay of binding energies between the native substrate and the drug. If a candidate mutant is able to disrupt the drug binding more than it disturbs the native substrate binding, it is considered a potential candidate for resistance. However, as indicated earlier, the paths that a pathogen needs to take from the wild type genotype to the target mutant hold additional information about the feasibility of resistance. Thus incorporating such evolutionary information can potentially improve the prediction and categorization of resistance mutations by the structural methods. In order to test this hypothesis, we compared the performance of the structural model of Chapter 4 against the evolutionary model described in Section 5. A list of HIV protease mutants predicted to be resistant by the structural model is provided in Table 5.6. For the evolutionary model, all mutants selected by evolutionary simulations were considered to be *resistant* by the model. Table 5.3 lists these mutants.

For model evaluation, the set of ritonavir resistant HIV protease mutations from HIV DB (see Table 5.1) was used as the set of true positives. Any mutant not appearing in this set was considered a false positive. Table 5.2 shows the mutual information calculated between the HIV DB gold standard and the predictions by different models. The structural model alone is a reasonable predictor of resistance when compared against the gold standard; however a significant improvement in this performance is seen when a mutational model is combined with the structural model to generate an evolutionary framework. A mutational process in the absence of selection was also simulated for comparison purposes. Since no selection was involved, a transition to state i in these

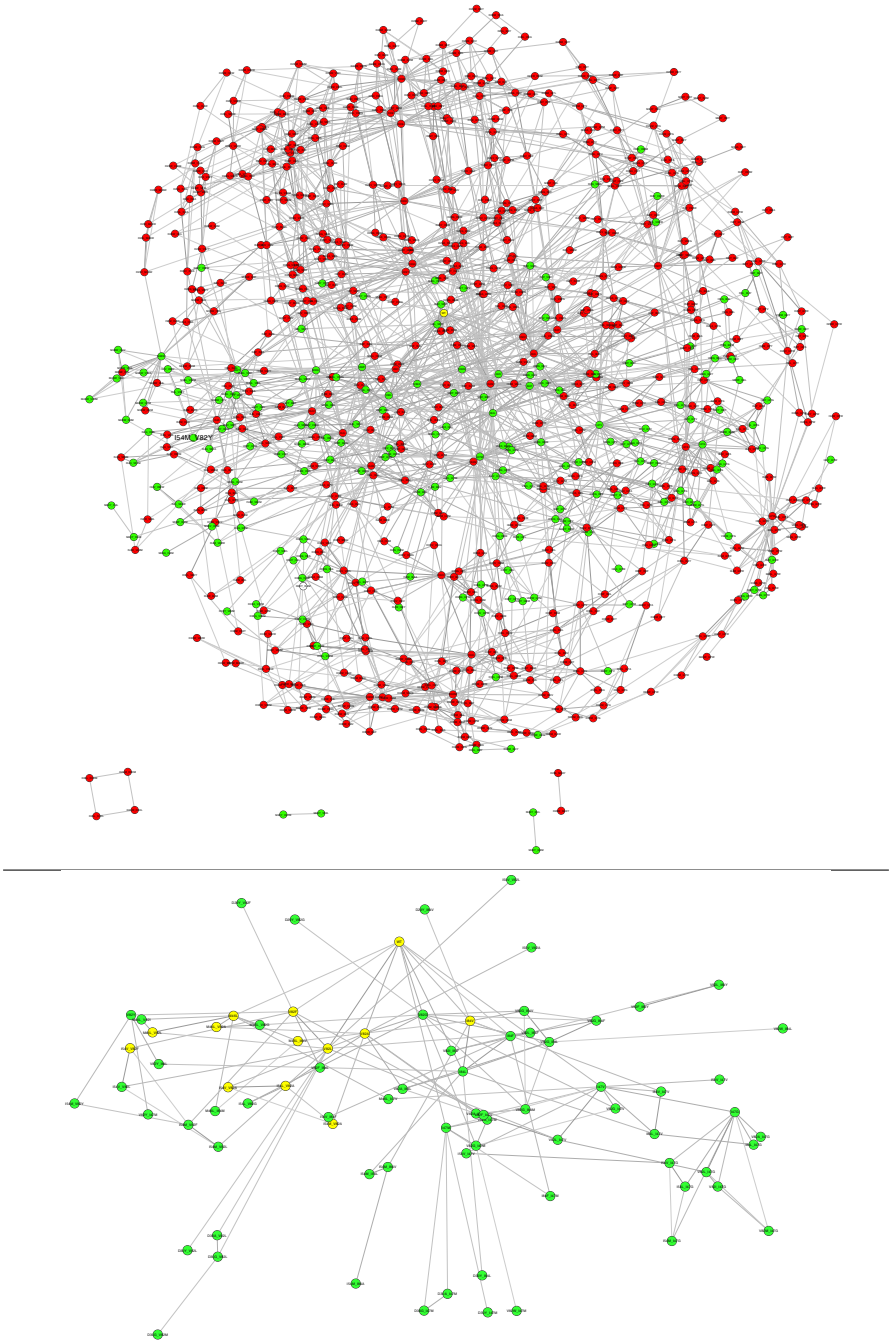

Figure 5.5: **HIV Protease Fitness Landscape Under Ritonavir** Top: The HIV fitness landscape is displayed as a graph where each node is a mutant of HIV protease. An edge between nodes indicates that a single DNA mutation can convert one into the other. The colour of the nodes represents the selection coefficient of the mutant under ritonavir calculated using Equation 5.1, red: sensitive and green: resistant. Wildtype is highlighted in yellow. All mutants with substrate binding within 1.5 kcal/mol of the wild type are displayed. Bottom: Part of the HIV fitness landscape selected by evolution is shown. Only a subset of beneficial mutants i.e. those resistant to ritonavir are sampled. True positives sampled by evolutionary simulations are shown in yellow. The networks were generated using Cytoscape (Shannon et al., 2003).

mutation only simulations was accepted or rejected with equal probability. It is evident that the random mutational process alone, i.e. mutation that does not take into account the structural aspects of resistance, is a poor indicator of resistance, as expected.

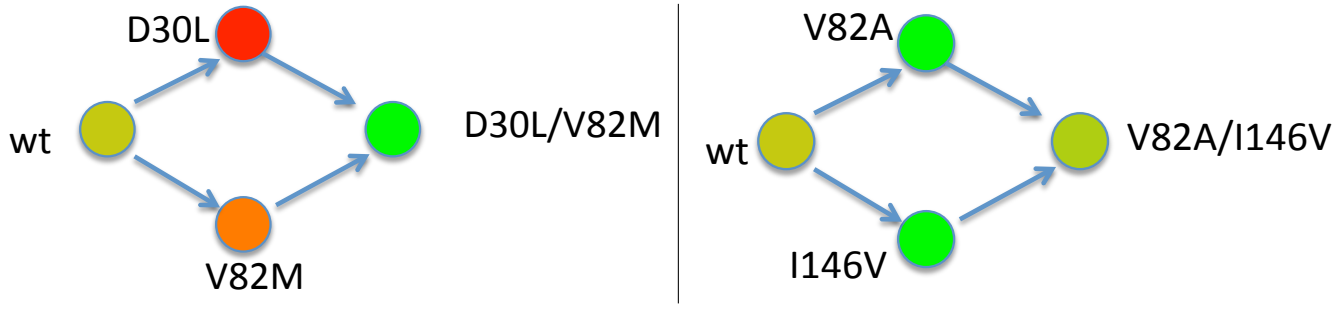

Figure 5.6: **Beneficial, Unreachable Double Mutants.** Two of the scenarios in which a beneficial double mutant is unlikely to be sampled by evolution are shown. Left: The path to the double mutant passes through deleterious (drug-sensitive) single mutants. Right: The intervening single mutants are more beneficial than the double mutant.

Figure 5.5 shows the entire fitness landscape that was included in the simulations. As mentioned earlier, this landscape contains 787 single and double point mutants. The part of this landscape selected by the evolutionary simulations is shown in Figure 5.5, Bottom panel. A few things are immediately clear. First, as expected, the evolutionary simulations only select the mutants that have a selective advantage i.e. the ritonavir resistant mutants. Second, evolution only selects a small percentage of the beneficial mutants. A total of 185 mutants in the original network were ritonavir resistant or evolutionary beneficial under ritonavir. However, evolutionary simulations only selected ≈80 of these mutants. A list of these mutants is found at the end of this chapter (see Table 5.3).

It is worth noting that the evolutionary model selects 12 of the 14 true positives from the structural model (i.e. all the gold standard resistance mutations that were predicted as resistant by the structural model). On the other hand, a 60 percent reduction in false positives between the structural and evolutionary model was seen. These false positives are mutations that appear to disrupt drug binding while maintaining near native substrate binding, and hence are categorized as resistant by the structural model. However these mutations do not appear in the HIV DB, indicating that the mutations have either not yet been seen or reported in an isolate or are rarely seen.

There can be a number of potential causes for these false positives. First, structural modelling is an approximation of the underlying biophysical realities. Hence, some mutants predicted to be resistant might not be resistant in practice due to inaccuracies in modelling. Second, the mutant might be a rare mutant or has a side effect due to which it is not seen commonly and/or is not deposited in the HIV DB etc. However, the significant reduction in false positives when an

evolutionary model is applied suggests that a large number of these false positives might be mutants that are not easily reachable from the wild type via single mutations.

This can happen in a number of different ways. For instance, in a scenario where all paths to the resistant target mutant go through fitness valleys, it is highly unlikely that the target mutant will be sampled by evolution. This can happen when the target double mutant is highly resistant, however the intermediate single mutants are both sensitive to the drug, thereby giving rise to fitness valleys. Figure 5.6 (Left) depicts such a scenario: the double mutant D30L/V82M is categorized as resistant by the structural model. However, both of the intervening single mutants D30L and V82M are categorized as sensitive by the structural model. Thus no highly likely path to D30L/V82M exists, and the mutant is not sampled by the evolutionary simulations. A similar situation arises when the paths to the target double mutant go through fitness peaks. In this scenario, the intervening single point mutants have a selective advantage over the target double point mutant, therefore making it highly unlikely that the double mutant will be sampled. Figure 5.6 (Right) depicts this scenario. The double mutant V82A/I146V is categorized as resistant, and has a positive selection coefficient. However, the paths to V82A/I146V go through the single mutants V82A and I146V, which are both highly resistant single mutants. Therefore, the mutational step from these single mutants to the double mutant requires a fitness sacrifice, making it unlikely for the evolutionary process to sample it.

A non-equlibrium process It was postulated earlier on in this chapter that the clinically well known mutations are possibly selected since they are easier to reach by an evolutionary process. Figure 5.7 shows the distribution of the clinically known and unknown mutants as observed during the course of our evolutionary simulations. In this context, a known mutation is one that appears in the HIV DB (or the gold standard set of mutations), whereas an unknown mutation is one that is categorized as resistant by the structural models and selected by the evolutionary simulations. It is evident that as the simulation progresses, the percentage of known mutations compared to the unknown ones increases, until it plateaus and then decreases. Thus the known mutations are sampled earlier in the evolutionary simulations, indicating that these mutations are possibly easier to reach by evolution and appear quickly, and are thus more widely seen in clinical practice. The unknown mutations that are resistance conferring according to our structural models tend to dominate the simulation towards the end. This can indicate that majority of these mutations are harder to reach by evolution. It is also worth noting that in clinical practice, prevalence of drug resistant mutations in a patient will render the treatment ineffective, thereby prompting a change in the prescribed drug. Thus, the evolutionary process in practice might never have the chance to explore these otherwise promising resistance candidates that are harder to reach.

Patient Trajectories Patient trajectories from HIV DB were also used to analyze results of the evolutionary simulations (Rhee et al., 2003; Shafer, 2006). These trajectories consist of HIV

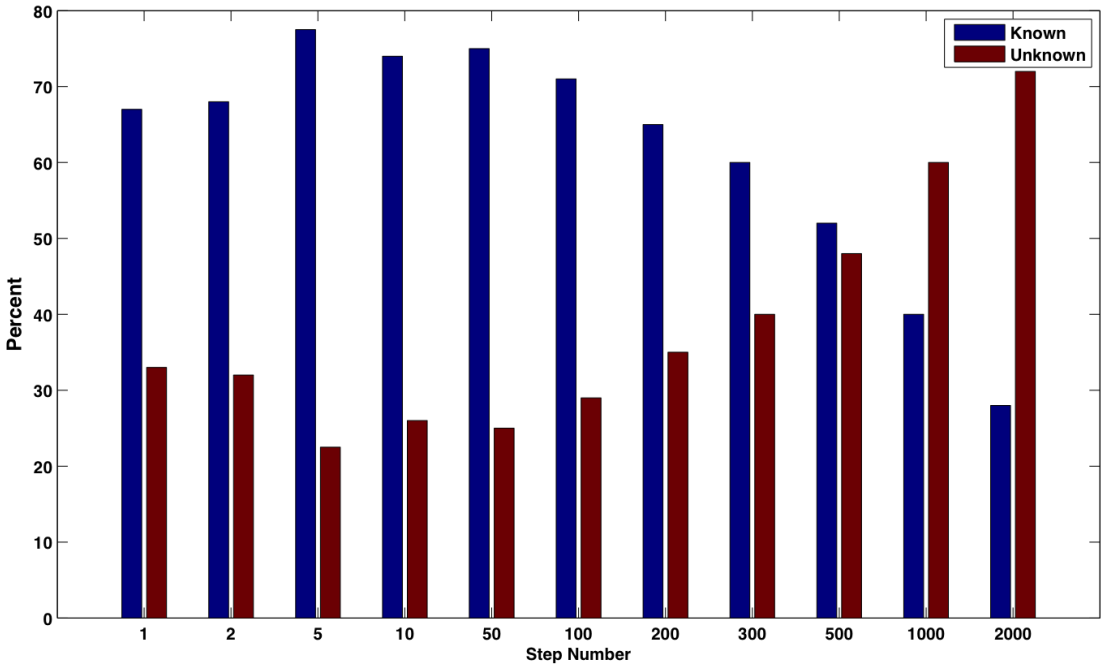

Figure 5.7: **Occurrence of Known Mutations in Evolution.** Percentage of known and unknown mutations selected by evolutionary simulations is plotted. In this context, known mutations are limited to those occurring in the gold standard set from HIV DB. Known mutations dominate the beginning of the evolutionary simulation indicating that these are easier to reach. As the evolutionary simulation progresses, more unknown mutations are sampled.

protease sequence data obtained from infected patients over a span of time and represent the evolutionary paths taken by the virus to acquire resistance. In order to facilitate comparison with the evolutionary simulations performed in this work, patient trajectories where mutations occurred in the modelled protease active site and consisted of the gold standard mutations (Table 5.1) were selected. Furthermore, any trajectories where multiple mutations appeared between two sequences, were excluded. Since the available patient trajectory data is sparse for ritonavir, and significant cross resistance is often observed between protease inhibitors (Kozal, 2004), patient data under other protease inhibitors was also included. The resulting set had 5 patient trajectories namely: wt $\rightarrow$ V82A, wt $\rightarrow$ M46L, wt $\rightarrow$ M46L $\rightarrow$ M46L/V82A, wt $\rightarrow$ I50V and wt $\rightarrow$ V82A $\rightarrow$ I54V/V82A. Of these, M46L is the most highly visited single mutant, making wt $\rightarrow$ M46L as the most frequently sampled transition in our simulations. Similarly, wt $\rightarrow$ V82A is the top 12th trajectory sampled in the simulation, whereas wt $\rightarrow$ M46L $\rightarrow$ M46L/V82A is the 17th most frequently sampled trajectory. Similarly, wt $\rightarrow$ V82A $\rightarrow$ I54V/V82A is found among the top 50 most frequently sampled trajectories. The evolutionary simulations did not sample wt $\rightarrow$ I50V. This is possibly due to a miscategorization of I50V as ritonavir sensitive by the structural method. Finally, it is worth noting that for an active site of 11 residues, a total of $\approx$8910 trajectories that consist of single and double point mutants that are hydrophobic, exist (11*9*10*9). Since 5 of these trajectories are reported in HIV DB, we expect $\approx$0.028 trajectories to show up in the 50 trajectories selected by evolution by random chance (5/8910 * 50). On the other hand, four of these trajectories are represented in the top 50 trajectories sampled by our evolutionary simulations. This represents an enrichment factor of 143 and is statistically significant at 5% (p-value 0).

While these results comparing the actual patient sequences against the evolutionary simulations are limited in nature, the fact that these trajectories could be retrieved by our evolutionary simulation indicates that significant information about evolution of resistance in patients can be obtained, even when using simple evolutionary models. A list of top paths from the evolutionary simulation is presented at the end of this chapter (see Table 5.5).

Compensatory Mutations Evolutionary simulations like those presented in this Chapter might also be used to identify the primary and compensatory mutations in a mutant. In this context, a primary mutation is a single point mutation that can confer resistance on its own. On the other hand, a compensatory mutation is a secondary mutation that can augment the resistance conferring capacity of the first mutant either by further impairing drug binding, improving substrate binding or both (Levin et al., 2000; Handel et al., 2006). However, such a compensatory mutation is not resistant on its own. In the HIV protease simulations performed, resistance conferring primary mutations in residues M46, I50, I54, V82 and I84 (see Table 5.6) were often paired with compensatory mutants. We find that in a scenario where a double mutant comprises of a primary and compensatory mutation, the primary mutation appears first, whereas the compensatory mutation

occurs next. A total of 25 such double point mutations were sampled during the evolutionary simulations. The trajectories were modelled using a binomial distribution, with success defined as the accumulation of the primary resistance mutation first. Probability of success was calculated independent of selection (i.e. by using the underlying mutational process alone). All observations were significant at P = 0.05 with Bonferroni correction for multiple hypothesis testing applied. A list of these mutations can be found in Table 5.4 presented at the end of this chapter, with the primary resistant conferring mutations highlighted in bold.

5.3.1 Evolution of Resistance under Drug Cocktails:

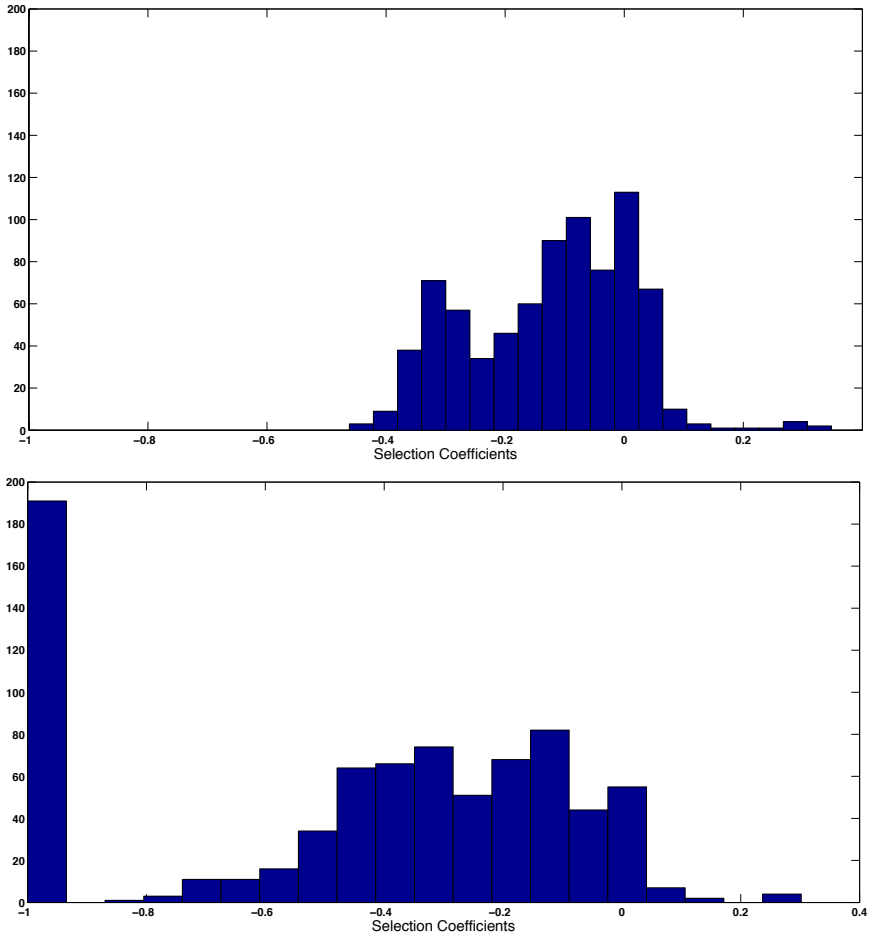

Figure 5.8: **Distribution of Selection Coefficients** Histograms of selection coefficients under ritonavir (Top) and ritonavir-nelfinavir-lopinavir cocktail (Bottom) are shown. A significant peak at -1 is found for the cocktail indicating that a large number of previously resistant mutants are sensitive to the cocktail.

Use of drug cocktails is an important part of therapy for HIV. These drug cocktails combine a

number of different drugs for increased efficacy and to overcome resistance. In order to study the evolution of resistance under a drug cocktail, Equation 5.1 is modified as follows:

$$s_i = e^{(\mathrm{E_{wt,s}} - \mathrm{E_{i,s}}) - \min_{\mathrm{d}}[(\mathrm{E_{wt,d}} - \mathrm{E_{i,d}})]} - 1 \tag{5.6}$$

Under a single drug, a mutant confers resistance if the drug binding is affected more than the substrate binding. Thus, in order to confer resistance to a drug cocktail, the mutant must be resistant to all the drugs. Furthermore, under a drug cocktail, multiple drugs are competing for binding, both with the substrate and with each other; thus the substrate binding should be compared against the drug that binds the mutant with the lowest binding energy (the most effective drug in the cocktail against the mutant). The min over the second part in the exponent captures this behaviour.

A protease inhibitor cocktail consisting of ritonavir, lopinavir and nelfinavir (RNL) was used in the evolutionary simulations. Addition of these additional protease inhibitors changes the fitness landscape of the protease. Similar to the previous section, all 787 mutants predicted to bind the native substrate within 1.5 kcal/mol of the wild type were used in resistance calculations. Using Equation 5.6, a total of 57 mutants were found to be resistant i.e. they had positive selection coefficients. This is in contrast to the 185 resistance conferring mutants found for ritonavir, indicating a reduction in resistance when a cocktail is applied. Figure 5.8 shows the distribution of selection coefficients for both ritonavir and the RNL cocktail on the simulated landscape. A significant shift towards the deleterious region in general (selection coefficients less than 0) can be seen for the cocktail.

The evolutionary landscape for RNL cocktail is shown as a graph in Figure 5.9. A total of 787 mutants are displayed, with each mutant being a single node. The colour of the nodes indicates the fitness or the selection coefficient of the mutant, with red being lethal and green being beneficial. Compared to the evolutionary landscape under ritonavir (see Figure 5.5), a significant increase in the number of red nodes (and hence negative selection coefficients) is seen reflecting the general shift towards increased sensitivity under the cocktail. Figure 5.9 shows the section of this landscape selected by the evolutionary simulations. Similar to the evolutionary simulations under ritonavir alone, it is obvious that a significantly smaller portion of the landscape is selected by evolution compared to that predicted to be beneficial by structure alone. Under ritonavir, a total of ≈ 80 or about 44 percent of the beneficial landscape was sampled by evolution. On the other hand, 11 out of 57 nodes or about 19 percent of the beneficial mutational space is explored under the cocktail. This points to an interesting property of the cocktails so far as resistance is concerned. A cocktail does not work by simply reducing the number of resistance conferring mutants. Instead, the use of multiple drugs in a cocktail seems to eliminate the mutational paths needed by evolution to sample resistance conferring mutants. Thus a cocktail prevents resistance by blocking off beneficial

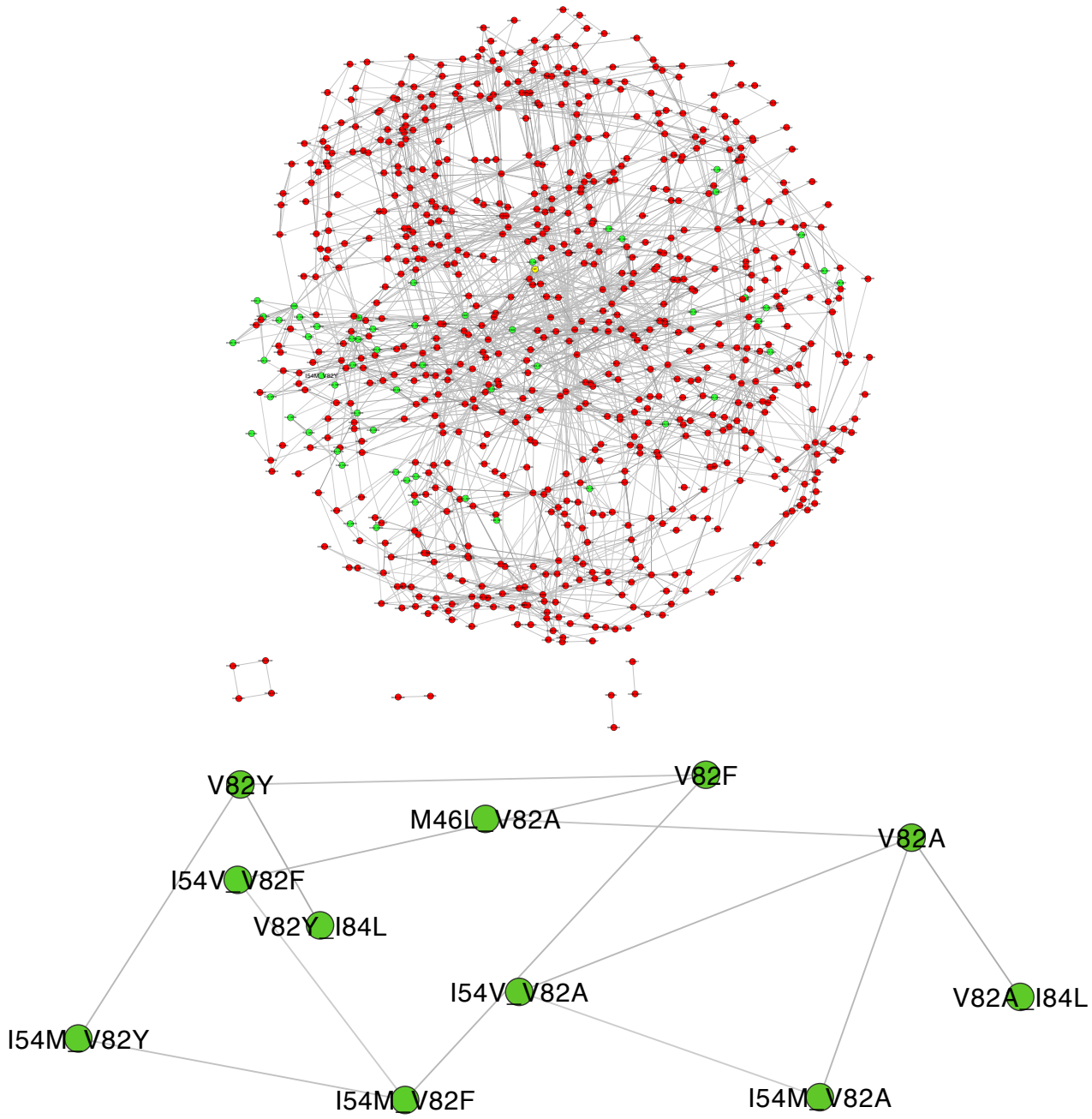

Figure 5.9: **HIV Protease Fitness Landscape Under a Protease Cocktail** Top: The HIV fitness landscape is displayed as a graph where each node is a mutant of HIV protease. An edge between nodes indicates that a single DNA mutation can convert one into the other. The colour of the nodes represents the selection coefficient of the mutant under a cocktail of ritonavir, nelfinavir and lopinavir (Equation 5.6) with green as beneficial and red as deleterious. Wildtype is highlighted in yellow. All mutants with substrate binding within 1.5 kcal/mol of the wild type are displayed. Bottom: Part of the HIV fitness landscape selected by evolution is shown. Only a subset of beneficial mutants i.e. those resistant to the ritonavir-nelfinavir-lopinavir cocktail are sampled. The networks were generated using Cytoscape (Shannon et al., 2003).

segments of the evolutionary landscape. A similar trend is also shown by Figure 5.10. The figure plots the average fitness across 5000 runs, through the evolutionary simulation. It is evident that under the cocktail, fitness evolves much slowly compared to the fitness under a single drug.

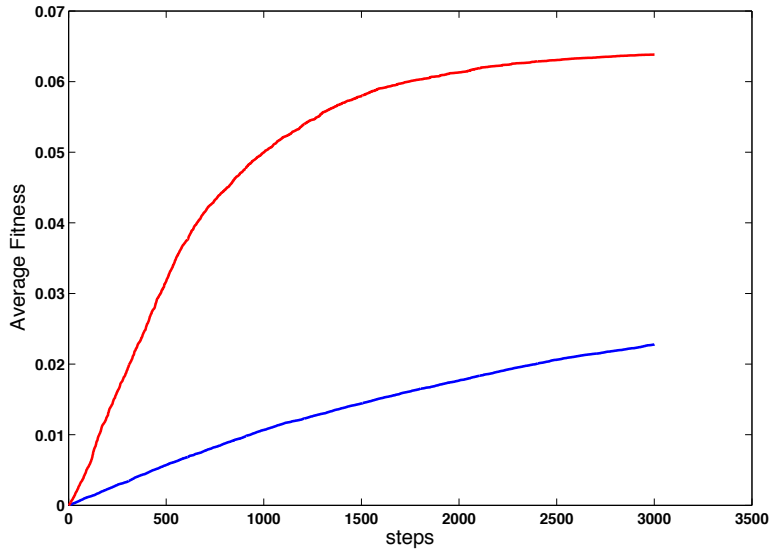

Figure 5.10: **HIV Fitness under a Cocktail** Average fitness (Equation 5.5) through the evolutionary simulation is plotted. The red line is fitness under ritonavir alone. Fitness under the ritonavir-nelfinavir-lopinavir cocktail is displayed in blue. As expected, fitness rises much slower under the cocktail pointing towards increased efficacy of the cocktail compared against ritonavir alone.

5.3.2 Resistance under Varying Levels of Drug Adherence:

Adherence to a drug regimen can potentially effect therapy outcome (Huang et al., 2011; Krakovska and Wahl, 2007a; Rosenbloom et al., 2012; Tam et al., 2008). Less than perfect drug adherence creates periods of fluctuating selection during the evolutionary process; as the drug doses are missed, the underlying landscape is under no drug selection, whereas as drug is resumed again, selective pressure due to the drug resumes. In order to study the effects of this fluctuating selective pressure, the evolutionary simulations were repeated with an additional input parameter k to specify the absence ($k = 0$) or presence of drug ($k = 1$) at each step of the Markov chain. In the presence of a drug, the selection coefficients used were calculated as before (see Equation 5.1). To represent the case when the drug was missed, selection was calculated as :

$$s_i = e^{(\mathrm{E_{wt,s}} - \mathrm{E_{mut,s}})} - 1 \tag{5.7}$$

Equation 5.7 is a modified version of Equation 5.1 with the drug component excluded. Adher-

ence levels of 70, 75, 80, 85 and 90 percent were simulated, where adherence level indicates how often the drug was taken. Hence, an adherence level of 90 indicates that 10 % of times, drug doses was missed. Other than the overall adherence level, the consecutive numbers of time a drug dosage is missed or the *gap* length might also affect the overall outcome of the drug regimen. Gap lengths of 10, 20, 50, 100 and 200 were used with all adherence levels. Thus a total of 25 simulation experiments (5 adherence levels each with 5 gap lengths) were performed.

Previous work exploring the effects of drug adherence on therapy outcome has mostly focused on reverse transcriptase and integrate inhibitors, and prolonged gaps in therapy or fluctuating adherence is known to adversely affect therapy outcome (Huang et al., 2011; Li et al., 2012; Krakovska and Wahl, 2007b; Rosenbloom et al., 2012; Tam et al., 2008). However, somewhat differing opinions about adherence and protease inhibitors have been reported. For instance, (Krakovska and Wahl, 2007b; Rosenbloom et al., 2012; Tam et al., 2008) report an increase in virological failure for decreased adherence for various protease inhibitor based therapies. On the other hand (Rosenbloom et al., 2012; Tam et al., 2008) have reported a minimal effect on therapy outcome when adherence is affected for some protease inhibitors. In the experiments performed, size of the visited landscape grew as adherence was decreased from 100 percent and gaps were introduced. However, in the current application of our evolutionary model to evolution under ritonavir, no correlation of the explored evolutionary landscape with the gap length or adherence levels was found (see Figure 5.11).

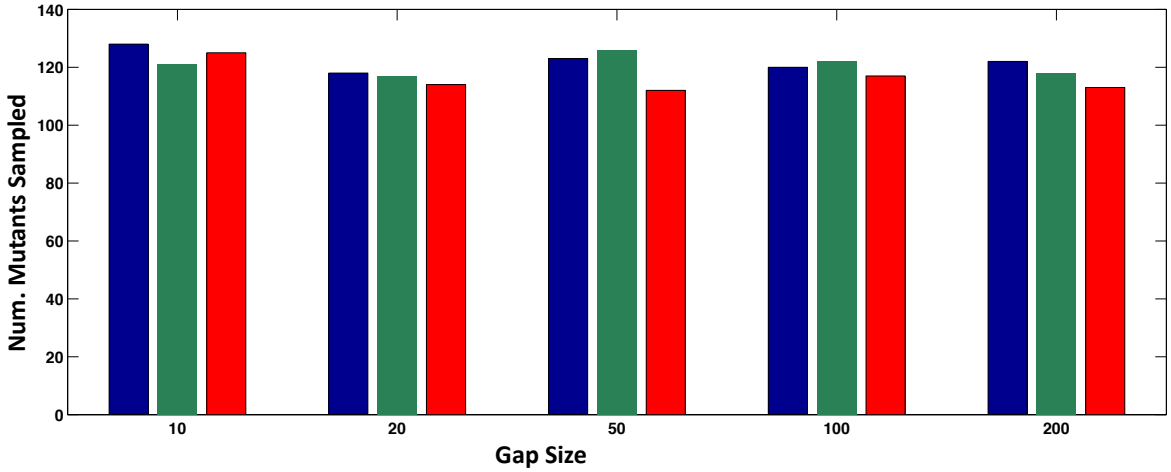

Figure 5.11: **Mutants explored for varying adherence levels and gap sizes** Number of mutants explored as the adherence level and gap size are changed. Adherence levels of 70 percent (blue), 80 percent (green) and 90 percent (red) are displayed. No correlation was found between adherence level, gap size and size of explored landscape and fitness.

5.4 Conclusion

In this chapter a simple evolutionary model to probe resistance was presented. This model uses a Markov chain based implementation of the evolutionary process, and incorporates resistance mutation scores obtained via structural models of Chapter 4 to calculate evolutionary fixation probabilities. Thus the model combines both the mutational and structural aspects of resistance. Adding mutational information can be crucial in some cases since paths leading from the wild type to a target mutant can offer insights into the feasibility of a particular resistance candidate. Use of this evolutionary model on HIV protease as a test system demonstrated that the evolutionary model is indeed a better model of drug resistance than structure alone. This is in part owing to the fact that evolutionary model quickly rules out the mutants that appear structurally promising but lie at the end of highly unlikely evolutionary paths e.g. those involving deleterious mutants or fitness sacrificing steps. This evolutionary model was also used to examine evolution of resistance under drug cocktails and under suboptimal adherence strategies. We find that under drug cocktails, as expected, portions of landscape beneficial under a single inhibitor are converted to those deleterious under the cocktail. More importantly, we find that a much smaller portion of that beneficial landscape under the cocktail is accessible. This hints at the possibility that a cocktail is efficient not simply because more of the evolutionary space is covered, but also because most previously likely evolutionary paths become unlikely indicating that good resistance candidates are harder to reach. This observation can possibly be included in optimal cocktail design (see Chapter 6).

While dependence of fitness on adherence was also tested in a number of experiments, we find that adherence has little to no effect under ritonavir. This observation is in qualitative agreement with a few previous studies that note that perfect adherence might not be determinant of treatment success under some protease inhibitors.

While the evolutionary method presented here improves on the structural methods of Chapter 4 to model drug resistance, it is not without its limitations. First, since this method uses binding energies generated using structural methods to calculate selection coefficients, it is also susceptible to the limitations affecting the structural components. Second, the method presented in this chapter is an approximation of a full evolutionary population dynamics simulation. Instead, this model treats the entire population as a single viral particle that moves from the wild type to different mutants. Finally, the formulae presented in this chapter do not explicitly take into account the dosage and half lives of the drugs involved. This potentially affects the scoring function for a drug cocktail as well as the adherence simulations. Incorporation of these two features, while not straight forward, can provide further insights. For a detailed description of future work, see Chapter 6.

Table 5.3: **Mutants Visited by Evolution** All HIV protease mutants selected by evolution are listed. See Figure (5.5, Bottom).

D29Y/I84V	D29Y/V82F	D29Y/V82G
D30A/I146M	D30A/V82L	D30G/I146M
D30G/V82L	D30G/V82M	D30Y/I146M
D30Y/I84L	D30Y/V82L	I146G
I146M	I146V	I50V/I146V
I50V/V82A	I50V/V82L	I54L/I146G
I54L/V82A	I54L/V82G	I54M/I146G
I54M/I84A	I54M/I84L	I54M/I84V
I54M/V82A	I54M/V82F	I54M/V82L
I54M/V82Y	I54V/I146G	I54V/I146V
I54V/I84F	I54V/V82A	I54V/V82F
I54V/V82L	I84F	I84F/I146M
I84L	I84L/I146G	I84L/I146V
I84V	I84V/I146V	M46L
M46L/I146V	M46L/I84M	M46L/I84V
M46L/V82A	M46L/V82G	M46L/V82I
M46L/V82L	V82A	V82A/I146G
V82A/I146M	V82A/I84L	V82A/I84M
V82Y/I84L	V82F	V82F/I146V
V82F/I84L	V82F/I84Y	V82G
V82G/I146M	V82G/I146V	V82G/I84F
V82G/I84L	V82G/I84M	V82G/I84V
V82I/I146G	V82I/I84F	V82L
V82L/I146G	V82L/I146V	V82L/I84F
V82L/I84Y	V82M/I146G	V82W/I146M
V82W/I84L	V82Y	V82Y/I146M

Table 5.4: **Double Mutants with Compensatory Single Mutants**

D30A/**I146M**	D30G/**I146M**
D30Y/**I146M**	I50V/**I146V**
I54V/**I146V**	I54V/**I84F**
V82I/**I84F**	D30Y/**I84L**
I54M/**I84L**	D29Y/**I84V**
I54M/**I84V**	**M46L**/I84M
M46L/V82I	I50V/**V82A**
I54L/**V82A**	I54M/**V82A**
I54V/**V82A**	**V82A**/I84M
D29Y/**V82F**	I54M/**V82F**
I54V/**V82F**	D29Y/**V82G**
I54L/**V82G**	**V82G**/I84M
D30A/**V82L**	D30G/**V82L**
D30Y/**V82L**	I50V/**V82L**
I54M/**V82L**	I54V/**V82L**

Table 5.5: **Top Trajectories in Evolutionary Simulations**

wt→M46L	wt→I84V
wt→M46L→M46L/V82L	wt→V82F
wt→I84V→D29Y/I84V	wt→V82L
wt→I146M	wt→M46L→M46L/I84M
wt→I146V	wt→I146M→I146M/D30Y
wt→I84F	**wt→V82A**
wt→I84F→I84F/I146M	wt→V82G
wt→V82F→D29Y/V82F	wt→M46L→M46L/V82G
wt→V82F→V82F/I54V	**wt→M46L→M46L/V82A**
wt→I146V→I146V/V82L	wt→I84V→D29Y/I84V
wt→M46L→M46L/V82G	wt→V82F→I54V/V82F
wt→M46L→M46L/V82A	wt→I146V→V82L/I146V
wt→V82F→V82F/I146V	wt→V82A→I50V/V82A
wt→I84V→I84V/I146V	wt→ I146V→I50V/I146V
wt→ V82A→M46L/V82A	wt→ V82A→V82A/I84L
wt→ V82G→D29Y/V82G	wt→ I146V→V82G/I146G
wt→ I84F→V82I/I84F	wt→ V82G→M46L/V82G
wt→ I146V→V82F/I146V	wt→ V82F→V82F/I84L
wt→ I84F→I54V/I84F	wt→ I84L→D30Y/I84L
wt→ V82L→D30Y/V82L	wt→ I146M→D30G/I146M
wt→ I146V→I84V/I146V	wt→ I146V→V82L/I146V
wt→ V82F→I54M/V82F	wt→ I146V→V82L/I146V
wt→ V82G→V82G/I84L	wt→ V82A→I54L/V82A
wt→ I146M→I84F/I146M	wt→ V82G→V82G/I146V
wt→ I84L→V82A/I84L	wt→ V82L→D30G/V82L
wt→ I84V→I54M/I84V	wt→ V82F→I54M/V82F
wt→I146V→I54V/I146V	wt→ I84V→M46L/I84V
wt→ V82G→V82G/I84M	wt→ M46L→M46L/V82I
wt→ V82G→V82G/I146M	wt→ V82L→M46L/V82L
wt→ M46L→M46L/V82I	wt→V82G→V82G/I84L
wt→I84V→I54M/I84V	**wt→V82A→I54V/V82A**
wt→V82A→I54M/V82A	wt→V82L→D30A/V82L

Table 5.6: **Predicted Resistance for HIV.** All mutants considered resistant in the HIV landscape.

D29Y/I84V	D29Y/V82FGW	I146GMV
D30A/I146M	D30A/V82LM	D30F/I146V
D30F/I54M	D30G/I146M	D30G/I50FLM
D30G/I54LM	D30G/V82LM	D30G/I146W
D30L/I146M	D30L/I84Y	D30L/V82M
D30M/I84Y	D30V/V82M	D30V/I146M
D30W/I146G	D30Y/I146MW	D30Y/I50FLMY
D30Y/I54LM	D30Y/I84LM	D30Y/V82LM
I50A/V82M	I50F/I146G	I50F/I184Y
I50F/V82I	I50G/I146W	I50G/V82L
I54G/I146M	M46L/I84M	I50G/I84LM
I50V/I146W	I50G/I54M	I50V/V82AFLW
I50L/I146VG	I50L/I84Y	I50L/V82FWY
I50V/I146GVW	I54A/V82AM	I54G/V82L
I54L/I84Y	I54L/V82AGWY	I54M/I146GLV
I54M/I84ALVY	I54M/V82AFGLMWY	I54V/I146GV
I54V/I84FL	I54M/I146G	I54M/I84V
I54L/I146G	I54V/V82AFGLWY	I84A/I146M
I84FV	I84F/I146MV	I84L
I84L/I146GVL	I84M/I146G	I84V/I146V
I84Y/I146MW	M46F/I54M	M46L/I146V
M46L/V82I	M46AV/I54M	M46L/150L
M46L/154M	M46L/I84VM	M46L/V82AGLM
M46LW	M46L	M46L/I54V
M46W/I146MV	M46W/I54MV	M46W/V82AILM
M46Y/I146M	M46Y/I54L	M46Y/I54M
M46Y/I84LM	M46Y/V82LM	V82AFGYL
V82A/I146GMV	V82A/I84LMVY	V82F/I146MV
V82F/I84FLMY	V82G/I146MV	V82G/I84FLMVY
V82I/I146GV	V82I/I84F	V82L/I146GLV
V82L/I84FY	V82L/I146M	V82M/I146G
V82M/I84Y	V82W/I146M	V82W/I84FLMY
V82Y/I84FLM	V82Y/I146M	

Chapter 6

Future Directions

6.1 Conclusions

The goal of this thesis was the development of methods that predict resistance *a priori*. In particular, prediction of resistance conferring mutations occurring in the active site of the target protein was desired. Furthermore, the aim was also to design a method that can predict this resistance with limited or no prior knowledge about drug resistance in a particular system.

In Chapter 3, an efficient protein redesign algorithm was presented that can be applied to smaller or restricted redesigns and consequently to drug resistance. In Chapter 4, this algorithm was combined with MMPBSA in a hierarchical approach to predict drug resistance. The efficacy of this approach was demonstrated on four different drug target systems. In all four cases, this method was able to correctly predict majority of the known mutants as resistant. In all cases, additional novel mutants were also predicted. Finally, in Chapter 5, an evolutionary model that combines sequence with structure was presented. This evolutionary model incorporates the process of mutation occurring at the sequence level along with its structural effects to better model drug resistance. It was shown that resistance mutations that are likely to occur in clinical practice are often those that can be easily reached during evolution. Thus, overall the methods presented in this thesis form a pipeline that can be used to predict resistance *a priori* (i.e. before it is observed in patients/clinical practice) in a drug target system. Finally, the approach presented is general in nature i.e. it can be applied to any drug target system where structures of the target protein, native substrate and the drug are available, and it does not rely on prior information regarding resistance in the system under consideration.

While the results of the current approach are promising, it is worth noting that a number of approximations were made by the structural and evolutionary methods. Consequently, a few limitations to this work exist. First, the results presented in this thesis were limited to single and

double mutants only. However, the methods presented in this thesis can be applied to probe mutants farther away from the wild type as well. The mutational search space grows combinatorially as the number of allowed mutations grows, and consequently the structural model will take longer to complete. However, as demonstrated in Chapter 3, even for these larger number of allowed mutations, rDEE search finished in less than 12 hours for the studied systems. It is worth noting that resistance is often limited to small number of residue changes. Furthermore, previously known resistance mutations known for most systems consist of 1 or 2 mutations only, thus limiting available validation data to single or double mutants. Second, for the systems tested for resistance in this thesis, the allowed set of residues was limited to either hydrophobic or hydrophobic plus polar neutral. This was motivated by a number of factors including the difficulty of scoring functions to score charged residues as well as scarcity of available validation data. Future work should expand on the possibility of including these harder to score residues to determine possible pitfalls, as well as possible residue-specific corrective measures for resistance predictions. One possible direction for this may be experimentation with various energy functions and force fields to determine the best ways of evaluating charged residues in context of resistance. Finally, proteins are flexible entities that sometimes undergo large conformational changes to accomplish a task (Prabu-Jeyabalan et al., 2006). The methods presented in the current work do not take into account protein flexibility at this scale. It is, however, worth noting that some degree of flexibility is incorporated by methods of Chapter 4 in the form of side chain rotamers and minimization. On occasion, approximations in modelling such as this can result in erroneous calculations of the binding energy and affect accuracy. An example of this phenomenon would be a highly flexible protein, where a mutant seems to disrupt substrate binding, however significant conformational changes in the protein can allow the substrate to bind better. If this conformational change is not captured by a computational method, the substrate binding energy is likely to be miscalculated and the mutant flagged as lethal or deleterious when, in reality, it is not so. Section 6.2 presents some extensions to the current work to address this issue. Third, the evolutionary model of Chapter 5 does not involve full scale population dynamics, instead it treats the entire population as a single particle. In addition, a single mutation at the nucleotide level is allowed during each step of the Markov chain and its effects on fitness are evaluated. In practice, multiple simultaneous mutations, though unlikely, might occur as well.

Finally, resistance can be conferred in a number of different ways. Of these, the most direct and common way is by the introduction of point mutations in the active site of the target protein. This is the mechanism of resistance that has been modelled in this work. However, the goal of this thesis did not involve modelling of other mechanisms of resistance such as compensatory mutations away from the active site (Foulkes-Murzycki et al., 2007; Mittal et al., 2012), over expression of efflux pumps, and over or under expression of compensatory genes and drug breakdown by bacterial

enzymes. Thus, the methods presented in this thesis are not directly applicable to resistance arising through these mechanisms. However, Section 6.2 presents possible extensions to this work to address resistance emerging in other ways.

6.2 Future Work

As the limitations of the approach presented have been discussed in the previous section, future directions addressing some of these limitations as well as building on the work performed in this thesis are provided in this section.

The methods presented in Chapter 4 model resistance as a function of binding energies for the drug and native substrate. During modelling, a single substrate is explicitly accounted for and used. For a number of drug target proteins, multiple ligands bind at the same location or close to the native substrate binding site. An example of this is the ATP binding site that overlaps the native substrate binding site for various proteins. This implies additional constraints on the mutant protein, since it needs to maintain near wild type binding with all the native substrates, while disrupting drug binding. Thus, knowledge of these additional ligands can be incorporated in the modelling process to improve accuracy. A possible approach to achieve this would be the addition of a filtering step that analyzes bindings of the predicted resistant mutants against other ligands and eliminates the ones that disturb binding with these additional ligands. Another example of such multi-ligand binders are proteins that bind ligands possessing a specific template. An example of these are the viral proteases that bind a wide range of peptide ligands to cleave them. These proteases often recognize a specific template in these peptides. This behaviour can be captured in a number of ways by a scoring process. First, one can create a template ligand structure for native substrate binding evaluation. For instance, a peptide template might be created by converting all residues that are not in the template to alanines or glycines. Alternately, the scoring procedure itself can be modified to emphasize interactions with the template locations higher than those for non-template locations: thereby capturing the binding with template ligand better.

Incorporating protein flexibility during structural modelling is crucial for accuracy, yet it remains computationally prohibitive. However, employing full fledged molecular dynamics (MD) encompassing large time scales (i.e. those that capture large scale conformational changes) can be inefficient, and in most cases, such large scale conformational changes might be unlikely to occur. The methods of Chapter 4 sacrifice some modelling accuracy for efficiency by using a hierarchical approach. Side chain flexibility in the form of rotamers is allowed during Stage 1; and while Stage 2 incorporates additional flexibility including minimization, full scale MD is not incorporated. Furthermore, even if Stage 2 was modified to incorporate MD, large scale conformational changes can occur on time scales which will make it infeasible to perform a full MD simulation for each mutant.

This is particularly the case for proteins such as membrane proteins and various kinases that undergo large conformational changes. Thus, mutations occurring in the protein must be compatible with all the target conformations. A mutant that disrupts one conformation is unlikely to occur in nature since it is likely to inhibit protein function. Finally, proteins might occupy multiple states such as phosphorylated or unphosphorylated, protonated or neutral etc. Along with conformational changes, these state changes can also be crucial such that a candidate mutant has to be compatible with all states of the protein. Thus, redesign algorithms that take conformational flexibility and multiple biochemical states into account can be employed for improved accuracy. For instance, multiState DEE (Yanover et al., 2007) is an attempt to incorporate multiple target states when looking for possible redesigns. Similarly, algorithms can be designed that take into account multiple conformational targets for mutation searching. These multi-target approaches can then be used as the first stage of a hierarchical resistance modelling algorithm, such as that presented in Chapter 4. Alternately, the algorithm presented in Chapter 4 can be executed in parallel, with each execution searching for mutants compatible with a unique conformational target. Then, an intersection of these parallel pipelines will be the mutants compatible with all target folds/conformations.

Drug cocktails are an important part of therapy in many diseases including HIV and various cancers. Chapter 5 presented a scoring function to determine whether a particular mutant confers resistance to a drug cocktail or not. The intuition behind the scoring function was that a mutant has to escape all drugs in the cocktail to be truly resistant to the cocktail. While this logic is intuitive in nature, the scoring function presented in Chapter 5 weighs all drugs in a cocktail equally. However, in practice, different doses of drugs are often combined to create a cocktail implying potentially varying contributions to the scoring system. A scoring function that uses differing contributions from each drug, possibly weighed by drug dosages, can be more useful. Future work can thus focus on finding the optimal dosage for each contributing drug in a given cocktail to minimize resistance to the cocktail.

Design of an optimal cocktail that minimizes emergence of drug resistance can aid therapy significantly. Chapter 5 presented evolutionary fitness landscapes as observed under various drugs and drug cocktails. In addition, we find that under drug cocktails, evolutionary paths are *blocked* by deleterious mutations, so that improving fitness under a cocktail becomes harder as compared to under a single drug. This points to another goal for optimal cocktail design. Future work can be directed at the design of optimal cocktails that focus on limiting the evolutionary mutant space available for the pathogen to escape to. Use of evolutionary landscapes and *a priori* knowledge of drug resistance can be potentially combined with stochastic search algorithms to search for an optimal cocktail that evades resistance.

Another exciting area for future work is the design of escape proof treatment strategies i.e. those that preclude drug resistance. In this context, an escape proof treatment strategy can be

designed to drive the viral population into a predetermined set of mutants M. This set of mutants M can be chosen such that all mutations leading the virus out of M are highly unlikely or lethal. In this way, the viral population on M is evolutionarily *trapped*. If the target set of mutants M is so chosen that it is susceptible to an existing drug (or cocktail) D, viral population on M can essentially be eliminated using D, without allowing it to develop resistance.

A further application of methods presented in Chapter 5 can be that of vaccine design or escape proof vaccines. An example of this is provided using HIV vaccine. The vaccine is a cocktail of antigens (e.g from HIV envelope protein) that produce HIV antibodies in the patients and strengthen the immune response to an HIV infection (Rerks-Ngarm et al., 2009; Gamble and Matthews, 2010). One can imagine creating an evolutionary landscape for the HIV envelope gene (or the antigens included in the vaccine) to examine regions of this landscape where the virus can escape to. Each antibody produced as a result of the vaccine injection covers a part of this landscape (i.e. strains susceptible to the antibody). Thus, the antibodies generated as response to the vaccine can be considered as a drug cocktail that exerts a selective pressure on the virus. We can then examine parts of this landscape that are not sensitive to any available antibody. These areas of the landscape represent HIV mutant strains that will escape the HIV vaccine since no antibody affects them. This knowledge can then be used to identify further antigens that can be added to the vaccine such that the antibodies produced cover the maximum area of the evolutionary landscape, making HIV vaccine escape proof.

Adherence is modelled in Chapter 5 as a fluctuating selective pressure. Our model does not take into account questions of drug half life, its dosage or its interactions with other drugs being administered. Future work should attempt at creation of a hybrid model that incorporates the evolutionary simulations described in this thesis with a mathematical model (Huang et al., 2011; Rosenbloom et al., 2012). Such a hybrid model will thus combine explicit structural information with a model of drug half life and dosage to better model when and how resistance starts to arise along the course of treatment.

Finally, prediction of *a priori* resistance in drug targets can guide the drug design process. First, it can aid in ranking lead compounds (candidate drugs) such that the lead compounds for which resistance evolves easily are ranked lower. In addition, knowledge of resistance *a priori* can potentially be used to redesign lead compounds so that it becomes difficult to evolve resistance to them. Future work can be aimed at such lead redesigns. The structural methods presented in this thesis provide bound structures of the mutants and drug as their output. These structures can potentially be analyzed to identify areas of the lead compound (drug) contributing to drug resistance. A redesign method could then incorporate this knowledge to redesign the lead compound and eliminate or improve these resistance susceptible areas.

Appendix A

Approximate Calculation of Selection Coefficients

For a wild type enzyme wt and a mutant a, the rate of reaction under Michaelis-Menten kinetics is given by:

$$V_{0,wt} = \frac{V_{max}[S]}{K_{m,wt}(1 + [I]/K_{i,wt}) + [S]} \tag{A.1}$$

$$V_{0,a} = \frac{V_{max}[S]}{K_{m,a}(1 + [I]/K_{i,a}) + [S]} \tag{A.2}$$

where K_m is the Michaelis constant, K_i is the dissociation constant for the inhibitor and $[S]$ and $[I]$ are the substrate and inhibitor concentrations respectively. Furthermore, V_{max} is the maximum reaction velocity and is given by $V_{max} = E_0 k_{cat}$, where E_0 is the enzyme concentration and k_{cat} is the turnover number.

Similar to (Wylie and Shakhnovich, 2011), we define selection to be a function of the wild type and mutant fitness/growth rate:

$$s = \frac{b_a}{b_{wt}} - 1 \tag{A.3}$$

where b_a and b_{wt} represent the fitness or growth rates of the mutant a and wild type respectively.

Assuming that fitness of the virus is dependent on the reaction rate, selection becomes a function of $V_{0,wt}$ and $V_{0,a}$.

$$s \approx \frac{V_{0,a}}{V_{0,wt}} - 1 \tag{A.4}$$

Furthermore, we assume that the V_{max} is the same for both the wild type and mutant proteins. This is a reasonable assumption since a resistant conferring mutant should maintain native function. (Furthermore, for the case of HIV protease, experiments conducted in Chapters 4 and 5 did not

allow any mutations in the catalytic triad to avoid disruption of k_{cat}.)

Substituting values of $V_{0,wt}$ and $V_{0,a}$ into Equation A.4, we have

$$s = \frac{K_{m,wt}(1 + [I]/K_{i,wt}) + [S]}{K_{m,a}(1 + [I]/K_{i,a}) + [S]} - 1 \tag{A.5}$$

At large inhibitor concentrations, Equation A.5 can be approximated by:

$$s \approx \frac{K_{m,wt}/K_{i,wt}}{K_{m,a}/K_{i,a}} - 1 \tag{A.6}$$

$$= \frac{K_{m,wt}K_{i,a}}{K_{i,wt}K_{m,a}} - 1 \tag{A.7}$$

Finally, at equilibrium K_m equals the dissociation constant. Assuming $K_{m,g} \propto e^{E_{g,s}}$ as well as $K_{i,g} \propto e^{E_{g,d}}$ where g is either wild type or mutant protein and $E_{g,s}$ is the energy of the protein with its native substrate and $E_{g,d}$ is the binding energy with the drug, we get:

$$s \approx \frac{e^{E_{wt,s}}e^{E_{a,d}}}{e^{E_{wt,d}}e^{E_{a,s}}} - 1 \tag{A.8}$$

Rearranging, we get

$$s_a \approx e^{(E_{wt,s}-E_{a,s})-(E_{wt,d}-E_{a,d})} - 1 \tag{A.9}$$

While the derivation provided above uses the wild type, without loss of generality, the formula can be used for any two mutants as well. In such a case, Equation A.9 reduces to:

$$s_{a\to b} \approx e^{(E_{a,s}-E_{b,s})-(E_{a,d}-E_{b,d})} - 1 \tag{A.10}$$

We claim that $s_{a\to b}$ can be determined by using the selection coefficients of a and b with respect to the wild type alone by using the following alternate formula:

$$s_{a\to b} = \frac{s_b + 1}{s_a + 1} - 1 \tag{A.11}$$

Use of Equation A.11 is computationally attractive, since the selection coefficients for all mutants need to be computed with respect to wild type alone. These selection coefficients can then be combined using Equation A.11 to calculate the selection coefficient $s_{a\to b}$ as a mutation is made from a mutant a to another mutant b.

Using Equation A.9, the selection coefficient of the mutants a and b with respect to wild type is given by:

$$s_a = e^{(E_{wt,s}-E_{a,s})-(E_{wt,d}-E_{a,d})} - 1 \tag{A.12}$$

$$s_b = e^{(E_{wt,s}-E_{b,s})-(E_{wt,d}-E_{b,d})} - 1 \tag{A.13}$$

Substituting these values in Equation A.11, we get

$$\begin{aligned} s_{a\to b} &= \frac{e^{(E_{wt,s}-E_{b,s})-(E_{wt,d}-E_{b,d})} - 1 + 1}{e^{(E_{wt,s}-E_{a,s})-(E_{wt,d}-E_{a,d})} - 1 + 1} - 1 \\ &= e^{(E_{wt,s}-E_{b,s})-(E_{wt,d}-E_{b,d})-(E_{wt,s}-E_{a,s})+(E_{wt,d}-E_{a,d})} - 1 \\ &= e^{E_{wt,s}-E_{b,s}-E_{wt,d}+E_{b,d}-E_{wt,s}+E_{a,s}+E_{wt,d}-E_{a,d}} - 1 \\ &= e^{-E_{b,s}+E_{b,d}+E_{a,s}-E_{a,d}} - 1 \\ &= e^{(E_{a,s}-E_{b,s})-(E_{a,d}-E_{b,d})} - 1 \end{aligned}$$

which is the formula provided by Equation A.10.

Appendix B

Resistance Mutations Predicted by MM-GBSA

Table B.1: **Predicted Resistance for Gleevec.** All mutants predicted resistant by MM-GBSA are listed.

Y253FLM	V256L	E286LY
M290L	I313V	T315IMV
I360V	L370M	D381F
F382W		

Table B.2: **Predicted Resistance for Isoniazid.** All mutants predicted resistant by MM-GBSA are listed.

I16TV	121AFGLTVW	F41ILMTV
S94TV	F149LT	K165M
L218Y	W222FLM	

Table B.3: **Predicted Resistance for Ritonavir.** All mutants predicted resistant by MM-GBSA are listed.

D29Y/I84V	D29Y/V82FGWY	D30A/I146M
D30A/V82LM	D30F/I146V	D30F/I54M
D30G/I146MW	D30G/I50FLM	D30G/I54LM
D30G/V82LM	D30L/I146M	D30L/I84Y
D30L/V82M	D30M/I84Y	D30V/I146M
D30V/V82M	D30W/I146G	D30Y/I146MW
D30Y/I50FLMY	D30Y/I54LM	D30Y/I84LM
D30Y/V82ILM	I146GMV	I50A/V82M
I50F/I146G	I50F/I84Y	I50F/V82I
I50G/I146W	I50G/I54M	I50G/I84LM
I50G/V82LM	I50L/I146GV	I50L/I84Y
I50L/V82FWY	I50V/I146GVW	I50V/V82AFLW
I54A/V82AM	I54G/I146M	I54G/V82L
I54L/I146G	I54L/I84Y	I54L/V82AGWY
I54M/I146GLV	I54M/I84ALMVY	I54M/V82AFGLMWY
I54V/I146GV	I54V/I84FL	I54V/V82AFGLWY
I84A/I146M	I84FLV	I84F/I146MV
I84L/I146GV	I84M/I146GL	I84V/I146V
I84Y/I146MW	M46A/I54M	M46F/I54M
M46LW	M46L/I146V	M46L/I50LMV
M46L/I84MV	M46L/V82AGILM	M46V/I54M
M46W/I146MV	M46W/I54MV	M46W/V82AILM
M46Y/I146M	M46Y/I54LM	M46Y/I84LM
M46Y/V82LM	V82AFGLY	V82A/I146GMV
V82A/I84LMVY	V82F/I146MV	V82F/I84FLMVY
V82G/I146MV	V82G/I84FLMVY	V82I/I146GV
V82I/I84F	V82L/I146GLMV	V82L/I84FY
V82M/I146G	V82M/I84Y	V82W/I146M
V82W/I84FLMWY	V82Y/I146M	V82Y/I84FLM

Table B.4: **Predicted Resistance for Methotrexate.** All mutants predicted resistant by MM-GBSA are listed.

E30A/F34L	E30I/V115M	E30M/Q35Y
F34A/Q35Y	F34G/V115LM	F34LM/L67M
F34Y/V115G	I7L/V8IM	I7MVW
I7M/E30LM	I7M/L22FIMV	I7M/L67IMV
I7M/Q35LMWY	I7M/V115AGIM	I7V/L22F
I7V/L67M	I7V/Q35Y	I7W/L22M
I7W/V115I	I7W/V115M	L22A/V115M
L22F	L22F/E30LM	L22F/F34ALM
L22F/L67M	L22F/Q35WY	L22F/T136LV
L22F/V115ILM	L22G/V115M	L22M/L67IMV
L22M/Q35FLMY	L22M/V115AG	L22V/Q35Y
L22V/V115I	L22W/E30M	L22W/L67M
L22W/Q35Y	L22Y/Q35Y	L67I/V115IM
L67M	L67M/T136V	L67M/V115I
L67V/V115M	Q35F/V115M	Q35L/V115IM
Q35M/V115M	Q35W/V115IM	Q35Y
Q35Y/L67M	Q35Y/T136V	Q35Y/V115I
Q35Y/V115M	V115IG	V115M/T136A
L22M/F31Y	L22F/F31LMY	L22A/F31MY
L22M/F31M	E30L/F31ML	E30I/F31ML
F31W/F34L	F31M/Q35Y	F31L/Q35YM
F31LM/Q35F	F31LM/L67M	F31M/L67I
F31L/L67V	F31LM/T136A	

Bibliography

Achaz, G., Palmer, S., Kearney, M., Maldarelli, F., Mellors, J. W., Coffin, J., and Wakeley, J. (2004). A robust measure of hiv-1 population turnover within chronically infected individuals. *Mol. Biol. Evol.*, 21:1902–1912.

Adams, C. P. and Brantner, V. V. (2006). Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff (Millwood).*, 25:420–428.

Althaus, C. L. and Bonhoeffer, S. (2005). Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1. *J. Virol.*, 79:13572–13578.

Altman, M., Ali, A., Reddy, G., Nalam, M., Anjum, S., Cao, H., Chellappan, S., Kairys, V., Fernandes, M., Gilson, M., Schiffer, C., Rana, T., and Tidor, B. (2008). HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J. Am. Chem. Soc.*, 130:6099–6113.

Altman, M., Nalivaika, E., Prabu-Jeyabalan, M., Schiffer, C., and Tidor, B. (2007). Computational design and experimental study of tighter binding peptides to an inactivated mutant of HIV-1 protease. *Proteins*, 70:678–694.

Ambroggio, X. and Kuhlman, B. (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. *J. Am. Chem. Soc.*, 128:1154–1161.

Andersson, H., Fridborg, K., Lowgren, S., Alterman, M., Muhlman, A., Bjorsne, M., Garg, N., Kvarnstrom, I., Schaal, W., Classon, B., Karlen, A., Danielsson, U., Ahlsen, G., Nillroth, U., Vrang, L., Oberg, B., Samuelsson, B., Hallberg, A., and Unge, T. (2003). Optimization of p1-p3 groups in symmetric and asymmetric hiv-1 protease inhibitors. *Eur. J. Biochem.*, 270:1746–1758.

Armstrong, K. and Tidor, B. (2012). Computationally mapping sequence space to understand evolutionary protein engineering. *Biotechnol. Prog.*, 24:62–73.

Arnold, C., Westland, L., Mowat, G., Underwood, A., Magee, J., and Gharbia, S. (2005). Single-nucleotide polymorphism-based differentiation and drug resistance detection in mycobacterium tuberculosis from isolates or directly from sputum. *Clin. Microbiol. Infect*, 11:122–130.

Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J., Stoddard, B. L., and Baker, D. (2006). Computational redesign of endonuclease dna binding and cleavage specificity. *Nature*, 441:656–659.

Baddeley, A., Dean, A., Dias, H., Falzon, D., Floyd, K., Garcia, I., Glaziou, P., Hiatt, T., Law, I., Lienhardt, C., Nguyen, L., Sismanidis, C., Timimi, H., van Gemret, W., and Zignol, M. (2013). Global tuberculosis report 2013. *World Health Organization*.

Bae, J. H., Rubini, M., Jung, G., Wiegand, G., Seifert, M. H., Azim, M. K., Kim, J. S., Zumbusch, A., Holak, T. A., Moroder, L., Huber, R., and Budisa, M. (2003). Expansion of the genetic code enables design of a novel gold class of green fluorescent proteins. *J. Mol. Biol.*, 328:1071–81.

Bang, H., Park, S., Hwang, J., Jin, H., Cho, E., Kim, D., Song, T., Shamputa, I., Via, L., Barry, C., Cho, S., and Lee, H. (2011). Improved rapid molecular diagnosis of multidrug-resistant tuberculosis using a new reverse hybridization assay, reba mtb-mdr. *J. Med. Microbiol.*, 60:1447–1454.

Beerenwinkel, N., Daumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., and Walter, H. (2003). Geno2pheno: estimating phenotypic drug resistance from hiv-1 genotypes. *Nucleic Acids Res.*, 31:3850–3855.

Bembom, O., Petersen, M. L., Rhee, S., Fessel, W., Sinisi, S., Shafer, R., and van der Laan, M. (2009). Biomarker discovery using targeted maximum-likelihood estimation: application to the treatment of antiretroviral-resistance hiv infection. *Statist. Med.*, 28:152–172.

Blanchard, J. S. (1996). Molecular mechanisms of drug resistance in mycobacterium tuberculosis. *Annu. Rev. Biochem.*, 65:215–239.

Boden, D. and Markowitz, M. (1998). Resistance to human immunodeficiency virus type 1 protease inhibitors. *Antimicrob. Agents Chemother.*, 42:2775–2783.

Borst, P. (1991). Genetic mechanisms of drug resistance. A review. *Acta Oncol.*, 30:87–105.

Brenner, B., Wainberg, M., Salomon, H., Rouleau, D., Dascal, A., Spira, B., Sekaly, R., Conway, B., and Routy, J. (2000). Resistance to antiretroviral drugs in patients with primary HIV-I infection. Investigators of the quebec primary infection study. *Int. J. Antimicrob. Agents*, 16:429–434.

Buendia, P., Cadwallader, B., and DeGruttola, V. (2009). A phylogenetic and markov model approach for the reconstruction of mutational pathways of drug resistance. *Bioinformatics*, 25:2522–2529.

Cai, Y., Yilmaz, N., Myint, W., Ishima, R., and Schiffer, C. (2012). Differential flap dynamics in wild-type and a drug resistant variant of hiv-1 protease revealed by molecular dynamics and nmr relaxation. *J Chem Theory Comput.*, 8:3452–3462.

Case, D. A., Cheatham, T., Darden, T., Gohlke, H., Luo, R., Merz, K. M. J., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. (2005). The Amber biomolecular simulation programs. *J. Comp. Chem.*, 26:1668–1688.

Cerny, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J. Optimiz. Theory App.*, 45:4151.

Chakrabarti, R., Kilbanov, A. M., and Friesner, R. A. (2005). Computational prediction of native protein ligand-binding and enzyme active site sequences. *Proc. Natl. Acad. Sci. USA*, 102:10153–10158.

Chazelle, B., Kingsford, C., and Singh, M. (2004). A semidefinite programming approach to side-chain positioning with new rounding strategies. *INFORMS Journal on Computing, Computational Biology Special Issue*, 16(4):380–392.

Cheatham, T. and Young, M. (2001). Molecular dynamics simulation of nucleic acids: Successes, limitations and promise. *Biopolymers*, 56:232256.

Chen, B. J., Causton, H. C., Mancenido, D., Goddard, N. L., Perlstein, E. O., and Pe'er, D. (2009a). Harnessing gene expression to identify the genetic basis of drug resistance. *Mol. Syst. Biol.*, 5:310.

Chen, C., Georgiev, I., Anderson, A. C., and Donald, B. (2009b). Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U S A*.

Chen, S., Zhang, D., and Seelig, G. (2013). Conditionally fluorescent molecular probes for detecting single base changes in double-stranded dna. *Nature Chemistry*, 5:782–789.

Chen, Y. Z., Gu, X. L., and Cao, Z. W. (2001). Can an optimization/scoring procedure in ligand-protein docking be employed to probe drug-resistant mutations in proteins? *J. Mol. Graph. Model.*, 19:560–570.

Cheng, A., Eksterowicz, J., Geuns-Meyer, S., and Sun, Y. (2010). Analysis of kinase inhibitor selectivity using a thermodynamics-based partition index. *J. Med. Chem.*, 53:4502–4510.

Conti, E., Stachelhaus, T., Marahiel, M., and Brick, P. (1997). Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of Gramicidin S. *EMBO J.*, 16:4174–4183.

Corbin, S., Buchdunger, E., Pascal, F., and Druker, B. (2002). Analysis of the structural basis of specificity of inhibition of the abl kinase by STI571. *J. Biol. Chem.*, 277:32214–32219.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995a). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1995b). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, 117:5179–5197.

Cowan-Jacob, S. W., Fendrich, S., Floersheimer, A., Furet, P., Liebetanz, J., Rummel, G., Rheinberger, P., Centeleghe, M., Fabbro, D., and Manley, P. W. (2007). Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta. Crystallogr. D Biol. Crystallogr.*, 63:80–93.

Daley, C. and Caminero, J. (2013). Management of multidrug resistant tuberculosis. *Semin. Respir. Crit. Care Med.*, 34:44–59.

Daniela, N., Schneider, V., Pialouxd, G., Krivinee, A., Grabarf, S., Nguyend, T., Girard, P., Rozenbaum, W., and Salmona, D. (2003). Haart interruption and drug resistance: Emergence of hiv-1 mutated strains after interruption of highly active antiretroviral therapy in chronically infected patients. *AIDS*, 17:2126–2129.

Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: An n?log(n) method for ewald sums in large systems. *J. Chem. Phys.*, 98:10089.

de Vos, M., Mller, B., Borrell, S., Black, P., van Helden, P., Warren, R., Gagneux, S., and Victor, T. (2013). Putative compensatory mutations in the rpoc gene of rifampin-resistant mycobacterium tuberculosis are associated with ongoing transmission. *Antimicrob Agents Chemother.*, 57:827–832.

Desjarlais, J. D. and Handel, T. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci.*, 4:2006–2018.

Desmet, J., Maeyer, M., Hazes, B., and Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539 – 542.

Dias, M. V., Vasconcelos, I. B., Prado, A. M., Fadel, V., Basso, L. A., de Azevedo, W. F., and Santos, D. S. (2007). Crystallographic studies on the binding of isonicotinyl-NAD adduct to wild-type and isoniazid resistant 2-trans-enoyl-ACP (CoA) reductase from mycobacterium tuberculosis. *J. Struct. Biol.*, 159:369–380.

DiMasia, J., Hansenb, R., and Grabowskic, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 22:151185.

Dixit, A., Torkamani, A., Schork, N. J., and Verkhivker, G. (2009). Computational modeling of structurally conserved cancer mutations in the RET and MET kinases: the impact on protein structure, dynamics, and stability. *Biophys. J.*, 96:858–874.

Drawz, S. M. and Bonomo, R. A. (2010). Three decades of beta-lactamase inhibitors. *Clinical Microbiology Reviews*, 23:160200.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161:1307–1320.

Dwyer, M. A., Looger, L. L., and Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science*, 304:1967–1971.

Eboumbou Moukoko, E. C., Bogreau, H., Briolant, S., Pradines, B., and Rogier, C. (2009). Molecular markers of plasmodium falciparum drug resistance. *Med. Trop (Mars)*, 69:606–612.

Eldridge, M., Murray, C., Auton, T., Paolini, G., and Mee, R. P. (1997). Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided. Mol. Des.*, 11:425–445.

Ercikan-Abali, E. A., Mineishi, S., and Tong, Y. (1996). Active site-directed double mutants of dihydrofolate reductase. *Cancer Res.*, 56:4142–4145.

Erickson, J. W. and Burt, S. K. (1996). Structural mechanisms of HIV drug resistance. *Annu. Rev. Pharmacol. Toxicol.*, 36:545–571.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.

Ferrari, A., Degliesposti, G., Sgobba, M., and Rastelli, G. (2007). Validation of an automated procedure for the prediction of relative free energies of binding on a set of aldose reductase inhibitors. *Bioorg. Med. Chem.*, 15:7865–7877.

Filikov, A., Hayes, R., Luo, P., Stark, D., Chan, C., Kundu, A., and Dahiyat, B. (2002). Computational stabilization of human growth hormone. *Prot. Sci.*, 11:1452–1461.

Fisher, R. A. (1930). *The genetical theory of natural selection.* Clarendon, Oxford.

Fjell, C. D., Jenssen, H., Hilpert, K., Cheung, W. A., Pante, N., Hancock, R. E., and Cherkasov, A. (2009). Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.*, 52:2006–2015.

Fogel, D. B. (1998). Evolutionary computation: The fossil record.

Fossati, E., Volpato, J., Poulin, L., Guerrero, V., Dugas, D., and Pelletier, J. (2008). 2-tier bacterial and in vitro selection of active and methotrexate-resistant variants of human dihydrofolate reductase. *J. Biomol. Screen.*, 13:504–514.

Foulkes-Murzycki, J., Scott, W., and Schiffer, C. (2007). Hydrophobic sliding: a possible mechanism for drug resistance in human immunodeficiency virus type 1 protease. *Structure*, 15:225–233.

Frenkel, D. and Smit, B. (2001). *Understanding Molecular Simulation: From Algorithms to Applications.* Academic Press, Massachussets, USA.

Frey, K. M., Georgiev, I., Donald, B. R., and Anderson, A. C. (2010). Predicting resistance mutations using protein design algorithms. *Proc. Natl. Acad. Sci. USA*, 107:13707–13712.

Frieboes, H., Edgerton, M. E., Fruehauf, J., Rose, F. R., Worrall, L. K., Gatenby, R. A., Ferrari, M., and Cristini, V. (2009). Prediction of drug response in breast cancer using integrative experimental/computational modeling. *Cancer Res.*, 69:4484–4492.

Gallagher, T., Alexander, P., Bryan, P., and Gillilad, G. L. (1994). Two crystal structures of the β1 immunoglobin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry*, 33:4721–4729.

Gamble, L. and Matthews, Q. (2010). Current progress in the development of a prophylactic vaccine for hiv-1. *Drug Des. Devel. Ther.*, 5:9–26.

Garret, T., Clingeleffer, D., Guss, J., Rogers, S., and Freeman, H. (1984). The crystal structure of poplar apoplastocyanin at 1.8-a resolution. the geometry of the copper-binding site is created by the polypeptide. *J. Biol. Chem.*, 259:2282–2825.

Georgiev, I. and Donald, B. (2007). Dead-end elimination with backbone flexibility. *Bioinformatics*, 23(13):185–194.

Georgiev, I., Keedy, D., Richardson, J., Richardson, D., and Donald, B. (2008a). Algorithm for backrub motions in protein design. *Bioinformatics*, 13:196–204.

Georgiev, I., Lilien, R., and Donald, B. (2006a). Improved pruning algorithms and divide-and-conquer strategies for dead-end elimination, with application to protein design. *Bioinformatics*, 22:e174–83.

Georgiev, I., Lilien, R., and Donald, B. (2006b). A novel minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. In *Proc. of the 11th Ann. Intl. Conf. on Research in Comput. Biol. (RECOMB)*, pages 530–545.

Georgiev, I., Lilien, R., and Donald, B. (2008b). The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *J. Comp. Chem.*, 29(10):1527–1542.

Gielens, C., Idakieva, K., De Maeyer, M., Van de Bergh, V., Siddiqui, N. I., and Compernolle, F. (2007). Conformational stabilization at the active site of molluskan (Rapana thomasiana) hemocyanin by a cysteine-histidine thioether bridge a study by mass spectrometry and molecular modeling. *Peptides*, 28(4):790–797.

Godzik, A. (1995). In search of the ideal protein sequence. *Protein Eng.*, 8:409–416.

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295:337–356.

Goldstein, R. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.*, 66:1335.

Golovin, A. and Henrick, K. (2008). MSDmotif: exploring protein sites an motifs. *BMC Bioinformatics*, 9:312.

Gordon, D., Hom, G., Mayo, S., and Pierce, N. (2003). Exact rotamer optimization for protein design. *J. Comp. Chem.*, 24:232–243.

Gordon, D. and Mayo, S. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comp. Chem.*, 19:1505–1514.

Guimaraes, C. and Cardozo, M. (2008). MMGB/SA rescoring of docking poses in structure based lead optimization. *J. Chem. Inf. Model.*, 48:958–970.

Handel, A., Regoes, R. R., and Antia, R. (2006). The role of compensatory mutations in the emergence of drug resistance. *PLoS Comput. Biol.*, 2:e137.

Hartl, D. and Clark, A. (2007). *Principles of Population Genetics, 4th edition.* Sinauer Associates., Massachussets, USA.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J. Mol. Evol.*, 22:160–174.

Hayes, R., Bentzien, J., Ary, M., Hwang, M., Jacinto, J., Vielmetter, J., Kundu, A., and Dahiyat, B. (2002). Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. USA*, 99:15926–15931.

Hayik, S., Dunbrack, R., and Merz, K. (2010). A mixed qm/mm scoring function to predict protein-ligand binding affinity. *J Chem Theory Comput.*, 6:30793091.

Heider, D., Verheyen, J., and Hoffmann, D. (2010). Predicting bevirimat resistance of HIV-1 from genotype. *BMC Bioinformatics*, 11:37.

Hellinga, H. and Richards, F. M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. USA*, 91:5803–5807.

Holland, J. H. (1975). Adaptation in natural and artificial systems.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence.* MIT Press, Cambridge, MA, USA.

Holohan, C., van Schaeybroeck, S., Longley, D., and Johnston, P. (2013). Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer*, 13:714–726.

Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.*, 51:69–82.

Hou, T. and Yu, R. (2007). Molecular dynamics and free energy studies on the wild-type and double mutant HIV-1 protease complexed with amprenavir and two amprenavir-related inhibitors: mechanism for binding and drug resistance. *J. Med. Chem.*, 50:1177–1188.

Hu, X., Wang, H., Ke, H., and Kuhlman, B. (2008). Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design. *Structure*, 16:1799–17805.

Huang, Y., Wu, H., Holden-Wiltse, J., and Acosta, E. P. (2011). A dynamic bayesian nonlinear mixed-effects model of hiv response incorporating medication adherence, drug resistance and covariates. *Ann. Appl. Stat.*, 5:551–577.

Huggins, D. J., Sherman, W., and Tidor, B. (2012). Rational approaches to improving selectivity in drug design. *J. Med. Chem.*, 55:1424–1444.

Ito, M., Fukuzawa, K., Mochizuki, Y., Nakano, T., and Tanaka, S. (2008). Ab initio fragment molecular orbital study of molecular interactions between liganded retinoid X receptor and its coactivator; part II: influence of mutations in transcriptional activation function 2 activating domain core on the molecular interactions. *J. Phys. Chem. A.*, 112:1986–98.

Jain, A. N. (1996). Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided. Mol. Des.*, 10:427–440.

Jiang, X., Farid, H., Pistor, E., and Farid, R. S. (2000). A new approach to the design of uniquely folded thermally stable proteins. *Protein Sci.*, 9:403–416.

Jones, D. T. (1994). De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.*, 3:3567–3574.

Jouaux, E. M., Timm, B. B., Arndt, K. M., and Exner, T. E. (2009). Improving the interaction of myc-interfering peptides with myc using molecular dynamics simulations. *J. Pept Sci.*, 15:5–15.

Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. *New York: Academic Press*, pages 21–132.

Junaid, M., Lapins, M., Eklund, M., Spjuth, O., and Wikberg, J. (2010). Proteochemometric modeling of the susceptibility of mutated variants of the hiv-1 virus to reverse transcriptase inhibitors. *PLoS One*, 5:e14353.

Kangas, E. and Tidor, B. (2000). Electrostatic specificity in molecular ligand design. *J. Chem. Phys.*, 112:9120–9132.

Kauffman, S. (1993). *The origins of order: self organization and selection in evolution.* Oxford University Press., New York, USA.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.

King, N. M., Prabu-Jeyabalan, M., Nalivaika, E., Wigerinck, P., Bethune, M., and Schiffer, C. (2004). Structural and thermodynamic basis for the binding of tmc114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *J. Virol.*, 78:12012–12021.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. ((1983). Optimization by simulated annealing. *Science*, 220:671680.

Knobler, S., Lemon, S., Najafi, M., and Burroughs, T. (2003). *The Resistance Phenomenon in Microbes and Infectious Disease Vectors: Implications for Human Health and Strategies for Containment.* The National Academies Press.

Kollman, P., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000). Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.*, 33:889–897.

Konig, J., Mller, F., and Fromm, M. F. (2013). Transporters and drug-drug interactions: important determinants of drug disposition and effects. *Pharmacol Rev.*, 65:944–966.

Kozal, M. (2004). Cross-resistance patterns among hiv protease inhibitors. *AIDS Patient Care STDS.*, 18:199–208.

Kraemer-Pecore, C., Lecomte, J., and Desjarlais, J. (2003). A de novo redesign of the ww domain. *Protein Sci.*, 12:2194–2205.

Krakovska, O. and Wahl, L. M. (2007a). Optimal drug treatment regimens for hiv depend on adherence. *J. Theor. Biol.*, 246:499–509.

Krakovska, O. and Wahl, L. M. (2007b). Optimal drug treatment regimens for hiv depend on adherence. *J. Theor. Biol.*, 7:499–509.

Kuhlman, B. and Baker, D. (Proc. Natl. Acad. Sci. USA). Native protein sequences are close to optimal for their structures. *2000*, 97:10383–10388.

Kuhlman, B., Dantas, G., Ireton, G., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, 302:1364–1368.

Lapins, M., Eklund, M., Spjuth, O., Prusis, P., and Wikberg, J. (2008). Proteochemometric modelling of hiv protease susceptibility. *BMC Bioinformatics*, 9:181.

Lapins, M. and Wikberg, J. E. (2009). Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors. *J. Chem. Inf. Model.*, 49:1202–1210.

Lasters, I. and Desmet, J. (1993). The fuzzy-end elimination theorem: correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Prot. Eng.*, 6:717–722.

Le Bras, J. and Durand, R. (2003). The mechanisms of resistance to antimalarial drugs in plasmodium falciparum. *Fundam. Clin. Pharmacol.*, 17:147–153.

Leach, A. and Lemon, A. (1998). Exploring the conformational space of protein side chains using dead-end elimination and the A^* algorithm. *Proteins*, 33:227–239.

Leigh-Brown, A. J. (1997). Analysis of hiv-1 env gene sequences reveals evidence for a low effective number in the viral population. *Proc. Natl. Acad. Sci. USA*, 94:1862–1865.

Levin, B. R., Perrot, V., and Walker, N. (2000). Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics*, 154:985–997.

Li, J. Z., Paredes, R., Ribaudo, H. J., Svarovskaia, E. S., Kozal, M. J., Hullsiek, K. H., Miller, M. D., Bangsberg, D. R., and Kuritzkes, D. R. (2012). Relationship between minority nonnucleoside reverse transcriptase inhibitor resistance mutations, adherence, and the risk of virologic failure. *AIDS*, 26:185–92.

Liberles, D., Teichmann, S., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L., de Koning, A., Dokholyan, N., Echave, J., Elofsson, A., Gerloff, D., Goldstein, R., Grahnen, J., Holder, M., Lakner, C., Lartillot, N., Lovell, S., Naylor, G., Perica, T., Pollock, D., Pupko, T., Regan, L., Roger, A., Rubinstein, N., Shakhnovic, E., Sjolander, K., Sunyaev, S., Teufel, A., Thorne, J., Thornton, J., Weinreich, D., and Whelan, S. (2012). The interface of protein structure, protein biophysics, and molecular evolution. *Protein Sci*, 21:769–785.

Lilien, R., Stevens, B., Anderson, A., and Donald, B. (2004). A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the gramicidin synthetase A phenylalanine adenylation enzyme. *Proc. of the 8th Ann. Intl. Conf. on Research in Comput. Mol. Biol. (RECOMB) San Diego, CA, March 2004*, pages 46–57.

Lilien, R., Stevens, B., Anderson, A., and Donald, B. (2005). A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the Gramicidin Synthetase A Phenylalanine Adenylation Enzyme. *J. Comput. Biol.*, 12:740–61.

LoBue, P., Sizemore, C., and Castro, K. (2009). Plan to combat extensively drug-resistant tuberculosis: recommendations of the federal tuberculosis task force. *MMWR Recomm. Rep.*, 13:1–43.

Looger, L., Dwyer, M., Smith, J., and Hellinga, H. (2003a). Computational design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190.

Looger, L. L., Dwyer, M. A., Smith, J. J., and Hellinga, H. W. (2003b). Computational design of receptor and sensor proteins with novel functions. *Nature*, 423:185–190.

Lovell, S., Word, J., Richardson, J., and Richardson, D. (2000a). The penultimate rotamer library. *Proteins*, 40:389–408.

Lovell, S., Word, J., Richardson, J., and Richardson, D. (2000b). The penultimate rotamer library. *Proteins*, 40:389–408.

Maglia, G., Jonckheer, A., De Maeyer, M., J.M., F., and Engelborghs, Y. (2008). An unusual red-edge excitation and time-dependent stokes shift in the single tryptophan mutant protein DD-carboxypeptidase from Streptomyces: the role of dynamics and tryptophan rotamers. *Prot. Sci.*, 17(2):352–361.

Majori, G. (2004). Combined antimalarial therapy using artemisinin. *Parassitologia.*, 46:85–7.

Marvin, J. S. and Hellinga, H. W. (2001). Conversion of a maltose receptor into a zinc biosensor by computational design. *Proc. Natl. Acad. Sci. USA*, 98:4955–4960.

McCandlish, M. (2011). Visualizing fitness landscapes. *Evolution*, 65:1544–1558.

Meynard, J. L., Vray, M., Morand-Joubert, L., Race, E., Descamps, D., Peytavin, G., Matheron, S., Lamotte, C., Guiramand, S., Costagliola, D., Brun-Vezinet, F., Clavel, F., and Girard, P. M. (2002). Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *AIDS*, 16:727–736.

Miller, L. H., Ackerman, H. C., Su, X. Z., and Wellems, T. E. (2013). Malaria biology and disease pathogenesis: insights for new treatments. *Nat Med.*, 19:156–167.

Mittal, S., Bandaranayake, R., King, N., Prabu-Jeyabalan, M., Nalam, M., Nalivaika, E., Yilmaz, N., and Schiffer, C. (2013). Structural and thermodynamic basis of amprenavir/darunavir and atazanavir resistance in hiv-1 protease with mutations at residue 50. *J Virol.*, 87:4176–4184.

Mittal, S., Cai, Y., Nalam, M., Bolon, D., and Schiffer, C. A. (2012). Hydrophobic core flexibility modulates enzyme activity in hiv-1 protease. *J Am Chem Soc.*, 134:4163–4168.

Nalam, M., Ali, A., Altman, M., Reddy, G., Chellappan, S., Kairys, V., Ozen, A., Cao, H., Gilson, M., Tidor, B., Rana, T., and Schiffer, C. (2010). Evaluating the substrate-envelope hypothesis: structural analysis of novel hiv-1 protease inhibitors designed to be robust against drug resistance. *J Virol.*, 84:5368–78.

Nalam, M., Ali, A., Reddy, G., Cao, H., Anjum, S., Altman, M., Yilmaz, N., Tidor, B., Rana, T., and Schiffer, C. A. (2013). Substrate envelope-designed potent hiv-1 protease inhibitors to avoid drug resistance. *Chem Biol.*, 20:1116–1124.

Nilges, M. and Brunger, A. T. (1991). Automated modelling of coiled coils: application to the gcn4 dimerization region. *Protein Eng.*, 4:649–659.

Noble, M., Endicott, J., and Johnson, L. (2004). Protein kinase inhibitors: Insights into drug design from structure. *Science*, 303:1800–1805.

Norris, J. R. (1997). *Markov Chains*. Cambridge University Press, UK.

Novoa de Armas, H., Dewilde, M., Verbeke, K., De Maeyer, M., and Declerck, P. J. (2007). Study of recombinant antibody fragments and pai-1 complexes combining protein-protein docking and results from site-directed mutagenesis. *Structure*, 15(9):1105–1116.

Obermeier, M., Pironti, A., Berg, T., Braun, P., Daumer, M., Eberle, J., Ehret, R., Kaiser, R., Kleinkauf, N., Korn, K., Kucherer, C., Muller, H., Noah, C., Sturmer, M., Thielen, A., Wolf, E., and Walter, H. (2012). Hiv-grade: A publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. *Intervirology*, 55:102–107.

Ode, H., Neya, S., Hata, M., Sugiura, W., and Hoshino, T. (2006). Computational simulations of hiv-1 proteases–multi-drug resistance due to nonactive site mutation l90m. *J Am Chem Soc.*, 128:7887–7895.

Ode, H., Ota, M., Neya, S., Hata, M., Sugiura, W., and Hoshino, T. (2005). Resistant mechanism against nelfinavir of human immunodeficiency virus type 1 proteases. *J. Phys. Chem. B.*, 109:565–574.

Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A. S., van de Weerdt, C., Hoch, J. C., Prosperi, C., Francois, J. M., Mayo, S. L., and Martial, J. A. (2003). De novo backbone and sequence design of an idealized alpha/beta-barrel protein:evidence of stable tertiary structure. *J. Mol. Biol.*, 325:163–174.

Ohtaka, H., Muzammil, S., Schon, A., Velaquez-Campoy, A., Vega, S., and Freire, E. (2002). Thermodynamic rules for the design of high affinity hiv-1 protease inhibitors with adaptability to mutations and high selectivity towards unwanted targets. *Int. J. Biochem. Cell. Biol.*, 36:1787–1799.

Operario, D. J., Moser, M. J., and St George, K. (2010). Highly sensitive and quantitative detection of the H274Y oseltamivir resistance mutation in seasonal A/H1N1 influenza. *J. Clin. Microbiol.*, pages 3517–3524.

Pantano, S., Alber, F., Lamba, D., and Carloni, P. (2002). Nadh interactions with wt- and s94a-acyl carrier protein reductase from mycobacterium tuberculosis: an ab initio study. *Proteins*, 47:62–68.

Parikh, S., Moynihan, D., Xiao, G., and Tonge, P. (1999). The role of tyrosine 158 and lysine 165 in the catalytic mechanism of InhA, the enoyl-ACP reductase from mycobacterium tuberculosis. *Biochemistry*, 38:13623–13634.

Parisi, G. and Echave, J. (2001). Structural constraints and emergence of sequence patterns in protein evolution. *Mol. Biol. Evol.*, 18:750–756.

Pasomsub, E., Sukasem, C., Sungkanuparph, S., Kijsirikul, B., and Chantratita, W. (2010). The application of artificial neural networks for phenotypic drug resistance prediction: evaluation and comparison with other interpretation systems. *Jpn. J. Infect. Dis.*, 63:87–94.

Pastor, M. and Cruciani, G. (1995). A novel strategy for improving ligand selectivity in receptor-based drug design. *J. Med. Chem.*, 38:4637–4647.

Pearlman, D., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E. I., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Commun.*, 91:1–41.

Pearlman, D. A. (2005). Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 map kinase. *J. Med. Chem.*, 48:7796–7807.

Pierce, N., Spriet, J., Desmet, J., and Mayo, S. (2000). Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comp. Chem.*, 21:999–1009.

Pierce, N. and Winfree, E. (2002). Protein design is NP-hard. *Prot. Eng.*, 15:779–782.

Ponder, J. and Case, D. (2003). Force fields for protein simulations. *Adv. Prot. Chem.*, 66:2785.

Ponder, J. and Richards, F. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences of different structural classes. *J. Mol. Bio.*, 193:775–791.

Prabhu-Jeyabalan, M., Nalivaika, E., and Schiffer, C. (2002). Substrate shape determines specificity of recognition for hiv-1 protease: analysis of crystal structures of six substrate complexes. *Structure*, 10:369–381.

Prabu-Jeyabalan, M., Nalivaika, E., Romano, K., and Schiffer, C. A. (2006). Mechanism of substrate recognition by drug-resistant human immunodeficiency virus type 1 protease variants revealed by a novel structural intermediate. *J Virol.*, 80:3607–3616.

Prabu-Jeyabalan, M., Nalivaika, E., and Schiffer, C. A. (2000). How does a symmetric dimer recognize an asymmetric substrate? a substrate complex of HIV-1 protease. *J. Mol. Biol.*, 301:1207–1220.

Pricl, S., Fermeglia, M., Ferrone, M., and Tamborini, E. (2005). T315I-mutated Bcr-Abl in chronic myeloid leukemia and imatinib: insights from a computational study. *Mol. Cancer Ther.*, 4:1167–1174.

Protein Data Bank Europe. Pdbemotif. URL: `http://www.ebi.ac.uk/pdbe-site/pdbemotif/`.

Raju, R., Burton, N., and Hillier, I. (2010). Modeling the binding of HIV-reverse transcriptase and nevirapine: an assessment of quantum mechanical and force field approaches and predictions of the effect of mutations on binding. *Phys. Chem. Chem. Phys*, 12:7117–7125.

Rathore, R., Sumakanth, M., Reddy, M., Reddanna, P., Rao, A., Erion, M., and Reddy, M. (2013). Advances in binding free energies calculations: Qm/mm-based free energy perturbation method for drug design. *Curr Pharm Des.*, 19:4674–4686.

Reid, C., Bassett, R., Day, S., Larder, B., de Gruttola, V., and Winslow, D. (2002). A dynamics rules-based interpretation system derived by an expert panel is predictive of virological failure. *Antiviral There.*, 7:S91.

Rerks-Ngarm, S., Pitisuttithum, P., Nitayaphan, S., Kaewkungwal, J., Chiu, J., Paris, R., Premsri, N., Namwat, C., de Souza, M., Adams, E., Benenson, M., Gurunathan, S., Tartaglia, J., McNeil, J., Francis, D., Stablein, D., Birx, D., Chunsuttiwat, S., Khamboonruang, C., Thongcharoen, P., Robb, M., Michael, N., Kunasol, P., Kim, J., and Investigators., M.-T. (2009). Vaccination with alvac and aidsvax to prevent hiv-1 infection in thailand. *N. Engl. J. Med.*, 361:2209–2220.

Rhee, S., Gonzales, M., Kantor, R., Betts, B., Ravela, J., and Shafer, R. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, 31:298–303.

Rhee, S., Taylor, J., Fessel, W., Kaufman, D., Towner, W., Troia, P., Ruane, P., Hellinger, J., Shrivani, V., Zolopa, A., and Shafer, R. (2010). HIV-1 protease mutations and protease inhibitor cross-resistance. *Antimicrob. Agents Chemother.*, 54:4253–4261.

Rhee, S., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D., and Shafer, R. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Natl. Acad. Sci. USA*, 103:17355–17360.

Rosenbloom, D., Hill, A. L., Rabi, S. A., Siliciano, R. F., and Nowak, M. A. (2012). Antiretroviral dynamics determines hiv evolution and predicts therapy outcome. *Nat. Med.*, 18:1378–1385.

Rosenzweig, S. A. (2012). Acquired resistance to drugs targeting receptor tyrosine kinases. *Biochem Pharmacol.*, 83:1041–1048.

Safi, M. and Lilien, R. H. (2010). Restricted dead-end elimination: protein redesign with a bounded number of residue mutations. *J. Comput. Chem.*, 31:1207–1215.

Safi, M. and Lilien, R. H. (2012). Efficient a priori identification of drug resistant mutations using dead-end elimination and mm-pbsa. *J. Chem. Inf. Model.*, 52:1529–41.

Schweitzer, B. I., Srimatkandada, S., Gritsman, H., Sheridan, R., Venkataraghavan, R., and Bertino, J. R. (1989). Probing the role of two hydrophobic active site residues in the human dihydrofolate reductase by site-directed mutagenesis. *J. Biol. Chem.*, 264:20786–20795.

Seo, T., Thorne, J., Hasegawa, M., and Kishino, H. (2002). Estimation of effective population size of hiv-1 within a host: A pseudomaximum-likelihood approach. *Genetics*, 160:1283–1293.

Shafer, R. (2006). Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.*, 194:S51–S58.

Shafer, R. W., Stevenson, D., and Chan, B. (1999). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.*, 27:348–352.

Shah, P., Hom, G., and Mayo, S. (1999). Preprocessing of rotamers for protein design calculations. *J. Comp. Chem.*, 25:1797–1800.

Shakhnovic, E. and Gutin, A. M. (1998). A new approach to the design of stable proteins. *Protein Eng.*, 6:793–800.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504.

Shen, Y., Altman, M. D., Ali, A., Nalam, M. N., Cao, H., Rana, T., Schiffer, C., and Tidor, B. (2013). Testing the substrate-envelope hypothesis with designed pairs of compounds. *ACS Chem Biol.*, 8:2433–2441.

Sherman, W. and Tidor, B. (2008). Novel method for probing the specificity binding profile of ligands: applications to hiv protease. *Chem. Biol. Drug. Des.*, 71:387–407.

Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L., and Springer, T. A. (2000). Computational design of an integrin i domain stabilized in the open high affinity conformation. *Nat. Struct. Biol.*, 7:674–678.

Shriner, D., Shankarappa, R., Jensen, M. A., Nickle, D. C., Mittler, J., Margolick, J., and Mullins, J. (2004). Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. *Genetics*, 166:1155–1164.

Slovic, A. M., Kono, H., Lear, J. D., Saven, J. G., and DeGrado, W. F. (2004). Computational design of water-soluble analogues of the potassium channel kcsa. *Proc. Natl. Acad. Sci. USA*, 101:1828–1833.

Spanagel, R. and Vengeliene, V. (2013). New pharmacological treatment strategies for relapse prevention. *Curr Top Behav Neurosci.*, 13:583–609.

Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998). 'continuum solvent studies of the stability of dna, rna and phosphoramidate-dna helices. *J. Am. Chem. Soc.*, 120:9401–9409.

Stachelhaus, T., Mootz, H. D., and Marahiel, M. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, 6:493–505.

Stadler, P. (2002). Fitness landscapes. *Appl. Math. and Comput*, 117:187–207.

Stevens, B. W., Lilien, R. H., Georgiev, I., Donald, B. R., and Anderson, A. C. (2006). Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity. *Biochemistry*, 45:15495–15504.

Stoffler, D., Sanner, M., Morris, G., Olson, A., and Goodsell, D. (2002). Evolutionary analysis of HIV-1 protease inhibitors: Methods for design of inhibitors that evade resistance. *Proteins*, 48:63–74.

Tam, L. W., Chui, C. K., Brumme, C. J., Bangsberg, D. R., Montaner, J. S., Hogg, R. S., and Harrigan, P. R. (2008). The relationship between resistance and adherence in drug-naive individuals initiating haart is specific to individual drug classes. *J. Acquir. Immune Defic. Syndr.*, 49:266–271.

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Mol. Biol. Evol.*, 10:512–526.

Tavare, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86.

Thomas, J., Ramakrishnan, N., and Bailey-Kellogg, C. (2009). Protein design by sampling an undirected graphical model of residue constraints. *IEEE/ACM Trans Comput Biol Bioinform.*, 6:506–516.

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.*, 8:1267–1289.

US Patent No. 7326534. Detection of gleevec resistant mutations. URL: `http://www.patents.com/us-7416873.html`.

van der Borght, K., van Craenenbroeck, E., Lecocq, P., van Houtte, M., van Kerckhove, B., Bacheler, L., Verbeke, G., and van Vlijmen, H. (2011). Cross-validated stepwise regression for identification of novel non-nucleoside reverse transcriptase inhibitor resistance associated mutations. *BMC Bioinformatics*, 12:386.

Van der Vaart, A., Gogonea, V., Dixon, S., and Merz, K. M. (2000). Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. *J. Comput. Chem.*, 21:14941504.

van Laethem, K., de Luca, A., Antinori, A., Cingolani, A., Perna, C. F., and Vandamme, A. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in hiv-1 infected patients. *Antiviral There.*, 7:123–129.

Van Laethem, K., Schrooten, Y., Lemey, P., Van Wijngaerden, E., De Wit, S., Van Ranst, M., and Vandamme, A. (2005). A genotypic resistance assay for the detection of drug resistance in the human immunodeficiency virus type 1 envelope gene. *J. Virol. Methods*, 123:25–34.

van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1:1–30.

Velazquez-Campoy, A., Muzammil, S., Ohtaka, H., Schon, A., Vega, S., and Freire, E. (2003). Structural and thermodynamic basis of resistance to HIV-1 protease inhibition: implications for inhibitor design. *Curr. Drug Targets Infect. Disord.*, 3:311–328.

Voigt, C., Gordon, D., and Mayo, S. (2000a). Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *Journal of Molecular Biology*, 299:789–803.

Voigt, C., Gordon, D., and Mayo, S. (2000b). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.*, 299:789–803.

Volpato, J., Fossati, E., and Pelletier, J. (2007a). Increasing methotrexate resistance by combination of active-site mutations in human dihydrofolate reductase. *J. Mol. Biol.*, 373:599–611.

Volpato, J., Fossati, E., and Pelletier, J. (2007b). Increasing methotrexate resistance by combination of active-site mutations in human dihydrofolate reductase. *J. Mol. Biol.*, 373:599–611.

Volpato, J. P., Yachnin, B. J., Blanchet, J., Guerrero, V., Poulin, L., Fossati, E., Berghuis, A. M., and Pelletier, J. (2009). Multiple conformers in active site of human dihydrofolate reductase F31R/Q35E double mutant suggest structural basis for methotrexate resistance. *J. Biol. Chem.*, 284:20079–20089.

Wahab, H. A., Choong, Y. S., Ibrahim, P., Sadikun, A., and Scior, T. (2009). Elucidating isoniazid resistance using molecular modeling. *J. Chem. Inf. Model.*, 49:97–107.

Wang, J., Hou, T., and Xu, X. (2006). Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. *Curr. Comput-Aid. Drug.*, 2:95–103.

Wang, X., Tong, X., Tang, H., Liu, P., Zhang, W., and Yang, R. (2007). Study on genotypic resistance mutations to antiretroviral drugs on HIV strains of treated and treatment-naive HIV-1 infectious patients in hubei province. *Zhonghua Liu Xing Bing Xue Za Zhi*, 11:1112 – 1115.

Warshel, A; Levitt, M. (1976). Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227249.

Weber, I. T., Miller, M., Jaskolski, M., Leis, J., Skalka, A. M., and Wlodawer, A. (1989). Molecular modeling of the HIV-1 protease and its substrate binding site. *Science*, 243:928–931.

Weiner, S., Kollman, P., Case, D., Singh, U., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984a). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc*, 106:765–784.

Weiner, S., Kollman, P., Case, D., Singh, U., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984b). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16:97–159.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the sixth International Congress of Genetics, Brooklyn, New York.*, pages 356–366.

Wylie, C. S. and Shakhnovich, E. I. (2011). A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci. USA*, 108:9916–9921.

Xia, Y. and Levitt, M. (2004). Simulating protein evolution in sequence and structure space. *Curr. Opin. Struct. Biol.*, 14:202–207.

Yang, Y. and Liu, H. (2006). Genetic algorithms for protein conformation sampling and optimization in a discrete backbone dihedral angle space. *J. Comput. Chem.*, 27:1593–1602.

Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39:105–111.

Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–314.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press., USA.

Yanover, C., Fromer, M., and Shifman, J. M. (2007). Dead-end elimination for multistage protein design. *J. Comput. Chem.*, 28:2122–2129.

Zhang, J., Hou, T., Wang, W., and Liu, J. S. (2010). Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proc. Natl. Acad. Sci. U S A.*, 107:1321–1326.

Zhou, T., Commodore, L., Huang, W., Wang, Y., Thomas, M., Keats, J., Xu, Q., Rivera, V., Shakespeare, W., Clackson, T., Dalgarno, D., and Zhu, X. (2011). Structural mechanism of the Pan-BCR-ABL inhibitor ponatinib (AP24534): lessons for overcoming kinase inhibitor resistance. *Chem. Biol. Drug. Des.*, 77:1–11.

Zhu, X. L., Ge-Fei, H., Zhan, C. G., and Yang, G. F. (2009). Computational simulations of the interactions between acetyl-coenzyme-A carboxylase and clodinafop: resistance mechanism due to active and nonactive site mutations. *J. Chem. Inf. Model.*, 49:1936–1943.

Zumla, A., Nahid, P., and Cole, S. T. (2013). Advances in the development of new tuberculosis drugs and treatment regimens. *Nat Rev Drug Discov.*, 12:388–404.

Zwanzig, R. (1954). High temperature equation of state by a perturbation method. i. nonpolar gases. *J. . Chem. Phys.*, 22:1420–1426.