UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational Structure-Based Methods to Anticipate HIV Drug Resistance Evolution and Accelerate Inhibitor Discovery

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philoshopy

in

Bioinformatics

by

Max W. Chang

Committee in charge:

Professor Richard K. Belew, Chair
Professor Philip E. Bourne
Professor Lin Chao
Professor Douglas D. Richman
Professor Bruce E. Torbett
Professor Wei Wang

2008

The dissertation of Max W. Chang is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2008

DEDICATION

For Sabrina, with love and squalor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

During my time as a graduate student, I have been grateful for the opportunity to work with many talented and enthusiastic people. Foremost, I would like to thank Rik Belew, my thesis advisor. As a mentor, his guidance and encouragement have been a constant source of support in my studies. As a scientist, his curiosity and breadth of knowledge have set a great example.

The members of the Molecular Graphics Laboratory have been wonderful colleagues. I thank its director, Arthur Olson, for his wisdom, advice, and support. It has been my pleasure to collaborate with David Goodsell, and I have always admired his passion for science. Ruth Huey's gracious assistance with myriad programming matters has been helpful in my work. Garrett Morris has also taught me much and I thank him especially for his love of all things Apple. For their valuable feedback during many lab meetings, I would like to thank Michel Sanner, Michael Pique, Alex Perryman, Rodney Harris, Alex Gillet, and Qing Zhang. I owe tremendous thanks to former MGL member William "Lindy" Lindstrom, whose patience, generosity, and knowledge were a boon to a young grad student.

I am also grateful for my experiences in the Torbett lab, whose members have been a source of friendship and support. Bruce's advice has been valuable and has led to collaborations on some of the most interesting projects during my studies. I thank Michael Giffin for being a vital part of these collaborations, and for being a great source of knowledge and commentary.

I would like to thank my parents for their love and encouragement. Their struggles through adversity have always been an inspiration to me, and I am happy to follow their example in striving for an advanced education.

Finally, I am deeply grateful for the love and support of my wife, Sabrina. Throughout the highs and lows on the path to graduation, her love has been constant.

Chapters 3.2 and 4 contain material that will be submitted to the journal *Retrovirology* as *Combining molecular docking and sequence analysis to predict resistance mutations for novel inhibitors of HIV protease*. Max W. Chang, Michael J. Giffin, Ying-Chuan Lin, John H. Elder, Arthur J. Olson, Bruce E. Torbett, and Richard K. Belew. I was the primary investigator and author of this work.

Chapter 6.1 is a reprint in full of materials that appeared in *Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries*. Max W. Chang, William Lindstrom, Arthur J. Olson, and Richard K. Belew. *Journal of Chemical Information and Modeling* 2007, **47**(3):1258–1262. I was the primary investigator and author of this paper.

Chapter 6.2 contains material that appeared in *Empirical entropic contributions in computational docking: Evaluation in APS reductase complexes*. Max W. Chang, Richad K. Belew, Kate S. Carroll, Arthur J. Olson, David S. Goodsell. *Journal of Computational Chemistry* 2008, **29**(11):1753–1761. I was the primary researcher of this work, David Goodsell was the primary author of the paper.

Chapter 6.3 is an extended version of a preprint of *Virtual screening for HIV protease inhibitors from a library of antiviral compounds obtained via a cell-based screen*. Max W. Chang, Michael J. Giffin, William M. Lindstrom, Jr., Arthur J. Olson, Richard K. Belew, and Bruce E. Torbett. Submitted to *Journal of Medicinal Chemistry*. I was the primary investigator and author of this work.

Chapter 7 contains material that appeared in *Purposeful retrieval: Applying domain insight for topically-focused groups of biologists*. Richard K. Belew and Max W. Chang. In *Search and Discovery in Bioinformatics: SIGIR 2004 Workshop*. Edited by Javed Mostafa and Padmini Srinivasan. 2004. I was the secondary researcher and author of this work.

VITA

2000-2001	Undergraduate Tutor, Computer Science Department, University of California, San Diego
2001	B.S. in Biology, University of California, San Diego
2001-2002	Bioinformatics Intern, Genomics Institute of the Novartis Research Foundation
2002-2008	Research Assistant, University of California, San Diego
2008	Ph.D. in Bioinformatics, University of California, San Diego

PUBLICATIONS

Empirical entropic contributions in computational docking: Evaluation in APS reductase complexes. Max W. Chang, Richad K. Belew, Kate S. Carroll, Arthur J. Olson, David S. Goodsell. *Journal of Computational Chemistry* 2008, **29**(11):1753–1761.

Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries. Max W. Chang, William Lindstrom, Arthur J. Olson, and Richard K. Belew. *Journal of Chemical Information and Modeling* 2007, **47**(3):1258–1262.

Modeling recombination's role in the evolution of HIV drug resistance. Richard K. Belew and Max W. Chang. In *Artificial Life X: The Tenth International Conference on the Synthesis and Simulation of Living Systems*. Edited by Luis M. Rocha, Mark Bedau, Dario Floreano, Robert Goldstone, Alessandro Vespignani, Larry Yaeger. 2006: 98-104.

Purposeful retrieval: Applying domain insight for topically-focused groups of biologists. Richard K. Belew and Max W. Chang. In *Search and Discovery in Bioinformatics: SIGIR 2004 Workshop*. Edited by Javed Mostafa and Padmini Srinivasan. 2004.

FIELDS OF STUDY

Major Field: Bioinformatics

Studies in evolutionary algorithms and data mining
Professor Richard K. Belew, University of California, San Diego

Studies in molecular docking
Professor Arthur J. Olson, The Scripps Research Institute

Studies in HIV protease evolution
Professor Bruce E. Torbett, The Scripps Research Institute

ABSTRACT OF THE DISSERTATION

Computational Structure-Based Methods to Anticipate HIV Drug Resistance Evolution and Accelerate Inhibitor Discovery

by

Max W. Chang
Doctor of Philoshopy in Bioinformatics

University of California San Diego, 2008

Professor Richard K. Belew, Chair

The evolution of drug resistance in HIV has been a major obstacle in combatting the AIDS epidemic, and development of the next generation of antiviral drugs will depend on improvements in the methodology addressing resistance. This work examines HIV evolution from a structural perspective, focusing on the development of methods to anticipate drug resistance and improve drug discovery efforts.

To understand the evolution of HIV in the presence of inhibitors requires knowledge of viral replication capacity as well as drug resistance. Replication capacity can be predicted with a phylogenetic approach, which estimates impairment in HIV protease activity. Pairing these estimates with a structural model based on molecular docking allows the detection of most major clinically observed protease mutations. Combining structural modeling and analysis of existing protease mutations generates predictions of drug resistance mutations for an experimental protease inhibitor. Mutagenesis experiments validate these predictions, while also revealing epistatic interactions and cross-resistance with existing inhibitors.

A fitness model based on predicted replication capacity and drug resistance is able to rank in vitro HIV mutant infectivity with significant accuracy. This fitness model is incorporated into a simulation of viral evolution, which correlates with clinically observed mutation prevalence. Simulations also affirm the high level of cross-resistance among protease inhibitors, highlighting the importance of alternative drug targets.

Current drug discovery projects often use computer-based models of protein-ligand interaction for docking and virtual screening. A novel analysis of binding energy results from large-scale virtual screening identifies representative wild-type and mutant protease structures, focusing future efforts. Complementary efforts in the study of APS reductase reveal correlations between the distribution docking results and the underlying energy surface. Cluster analysis is shown to be an empirical measure of docking entropy which can improve the accuracy of binding energy predictions.

Applying these insights in a virtual screen for new inhibitors of HIV protease, a library of 1,585 compounds is narrowed to 36 candidates. Five of these compounds prove to be inhibitors. Modeling indicates that two of them bind outside the protease active site, suggesting potential leads for a new class of protease inhibitor.

Chapter 1

Introduction

The HIV/AIDS epidemic has claimed millions of lives worldwide, while millions more are currently infected. The availability of antiviral drug therapy has been successful in controlling infections and prolonging lives, often suppressing viral replication to undetectable levels. However, current therapies are unable to completely eliminate infection. Over time, even low levels of viral replication, coupled with the selective pressure of therapy, can drive the evolution of drug-resistant virus. The continued development of drug resistance limits the efficacy of antiviral therapy, eventually leading to death.

To combat HIV drug resistance, several major strategies have been laid out. With large-scale drug development efforts, it may be possible to continue developing new inhibitors indefinitely. However, this results in an evolutionary “arms race” where the HIV’s high mutation rate confers an advantage – new resistant strains arise quickly. Furthermore, many antiviral drugs work using a similar mechanism, e.g. HIV protease inhibitors, and certain mutations can confer resistance against multiple drugs of the same class. As a consequence, newer and more potent inhibitors are often limited by pre-existing resistance mutations.

Alternative strategies have been proposed which attempt to address the development of resistance more directly. With a detailed understanding of the HIV’s fitness landscape, it may be possible to guide the virus into an evolutionary dead-end, where its replication capacity is extremely limited. Another strategy involves applying multiple treatment strategies to “ping-pong” the virus between different resistant states. In either case, these approaches require a detailed understanding of viral evolution, including the anticipation of prospective mutations. In addition, targeting inhibitors against specific mutants will benefit from enhancements in the drug development process.

The current work seeks to address both the modeling of drug resistance evolution and improved methods for drug development. In addressing viral evolution, there is a wealth of information that spans multiple scientific disciplines. A major challenge lies in building coherent models of viral evolution with data from multiple sources, including clinical, biochemical, and structural studies. Integrating these data sources into a comprehensive fitness model is important in understanding the underlying basis for the evolution of drug resistance. With this detailed fitness model, computational simulations incorporating different drug therapies would yield meaningful estimates of viral evolution. These simulations would also be useful in testing new treatment strategies or predicting likely mutations against novel inhibitors.

To combat resistant forms of HIV, it is also clear that accelerating the development of new inhibitors is vital. One major technique used in modern drug discovery is virtual screening, which makes use of molecular docking to predict interactions between a ligand and macromolecule. Using large-scale computing resources can accelerate the process of screening large libraries of compounds and panels of mutant proteins, but requires new techniques for analysis. In finding new inhibitors, the current level of accuracy in virtual screening is generally sufficient to narrow the pool of inhibitor candidates for further experimental testing. However, often virtual screening suffers from too many false positives for the results to be used without complementary experiments. Improvements in the underlying docking process would boost accuracy in screening, allowing new inhibitors to be found more quickly.

Chapter 3 details different methods for estimating viral replication capacity. Statistical and machine learning approaches for predicting replication capacity from an amino acid sequence are described. Estimates based on sequence homology are also tested. Lastly, the basis for protease specificity is investigated using the physicochemical properties of substrate sequences and decision trees.

In Chapter 4, methods for predicting drug resistance are presented. This section begins by noting the high levels of cross-resistance in existing protease inhibitors. A structure-based method involving docking is presented, which is able to account for most of the major mutations in protease resistance. This model is employed in resistance mutations for a novel protease inhibitor, AB2, with encouraging results.

Combining the work in predicting replication capacity and drug resistance to explore mutation pathways and evolution is documented in Chapter 5. Covariation analysis of resistant sequences is shown to be useful in reconstructing mutation pathways. A model of viral fitness that accounts for both drug resistance, replication capacity, and epistasis is used to simulate drug resistance evolution.

Chapter 6 describes work related to protein-ligand docking and drug discovery. It begins with an analysis of a large set of docking experiments between wild-type and mutant HIV proteases and a diverse ligand library. The following section addresses the role of entropy in protein-ligand docking and the use of entropy estimates in predicting interactions between APS reductase and several ligands. These results are incorporated into a virtual screen of potential protease inhibitors, leading to the discovery of a set of novel inhibitors.

This work continues in Chapter 2 with a review of concepts relevant to HIV drug resistance and computational methods in drug discovery. Discussion of research related to this dissertation is also included.

The massive increase in biological data in recent years has also affected scientific publishing. In Chapter 7, a specialized program for searching scientific literature, HIVLink, is presented. By focusing on HIV-related literature, this program is able to provide several advantages over traditional methods.

Finally, in Chapter 8, the findings are summarized, along with a discussion of future research directions.

Chapter 2

Background

This dissertation has been influenced by work spanning a number of fields, from the evolution of HIV to molecular docking. The following material provides a brief introduction to the main concepts applied in later chapters. An overview of the HIV replication cycle is first discussed, describing the key steps targeted by current anti-HIV drugs. One such target is HIV protease, which is a main topic of this work. Following a description of HIV protease structure, features of drug resistance are considered, with emphasis on the prediction of the consequences of mutations in viral proteins. Drug resistance is a major factor in viral fitness and evolution, and such predictions can be used to anticipate viral evolution in computer simulations. Previous simulation efforts have addressed a variety of HIV behaviors and are described below. Finally, as the discovery of new inhibitors is of major importance in addressing drug resistance, the use of molecular docking in drug discovery is discussed, with an emphasis on the AutoDock program.

2.1 The HIV Replication Cycle

HIV is a retrovirus with a genome consisting of 9.7 kb of RNA. Two species of HIV are known to infect humans, HIV-1 and HIV-2. HIV-1, which is thought to have originally been transmitted from chimpanzees to humans, is the dominant form of the virus, and responsible for the majority of HIV infections.[1] Since HIV-1 is also more virulent than its counterpart,[2] it is the focus of this work, and all references to HIV indicate the HIV-1 species.

The genome of HIV contains 9 genes and encodes 15 proteins,[3] as shown in Figure 2.1. Of most relevance to current drug therapies are the protease, reverse transcriptase (RT), and integrase enzymes, which catalyze crucial steps in viral replication. The envelope glycoproteins

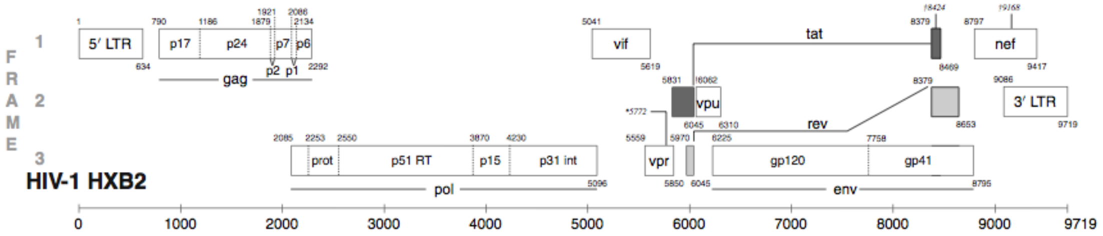


Figure 2.1: Overview of the HIV genome. Reproduced from the HIV Sequence Compendium.[4]

gp41 and gp120 are also of note as targets for both vaccine- and inhibitor-based therapies.

An overview of the HIV replication cycle is shown in Figure 2.2. Viral entry begins when an HIV virion binds to the CD4 receptor of a cell. Entry is mediated by a co-receptor, either CCR5 or CXCR4, and the contents of the virion are transferred into the cell as the membrane and envelope fuse. Following entry, the RT enzyme reverse transcribes the viral RNA genome to DNA. A key feature of the reverse transcription process is its high error rate, which causes HIV to mutate rapidly. Another interesting aspect of HIV replication arises from the two copies of the viral genome present in each virion, which may be heterozygous. During reverse transcription, the RT enzyme alternates between the two RNA strands, and is capable of producing a recombinant genome. Viral DNA is incorporated into the host genome by integrase. The integrated DNA is transcribed in large quantities for use as genomic RNA and mRNA. The viral genes are expressed as large polyproteins, which assemble with viral RNA to bud from the cell to form immature virions. As the multiple sites in the polyproteins are cleaved by HIV protease, the virion matures and becomes infectious, continuing the cycle. Substantial efforts have gone toward disrupting the function of viral proteins, with the greatest successes against protease and RT. Since structure-based techniques have been so useful in the design of protease inhibitors,[5] the protease enzyme is a focus of this work.

2.2 HIV Protease

The HIV protease enzyme is responsible for cleaving multiple sites in the polyproteins within an immature virion. If this process is blocked, the virions are non-infectious. The pivotal role of this enzyme has made it a major target for drug therapy, and especially for structure-based drug design.[5] From a structural standpoint, it is one of the most well-studied proteins in history, with hundreds of structures available in the Protein Data Bank (PDB).

The enzyme itself is a homodimeric aspartic protease, with each subunit consisting of

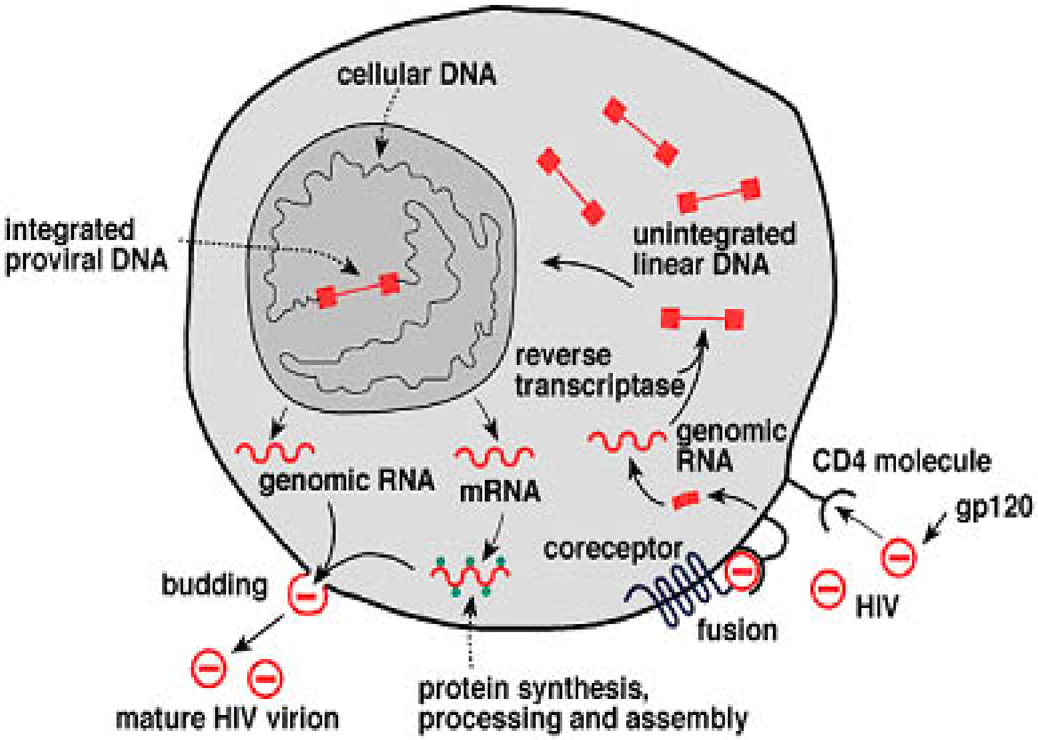


Figure 2.2: Overview of the HIV replication cycle. Reproduced from the NIAID fact sheet on HIV/AIDS.[6]

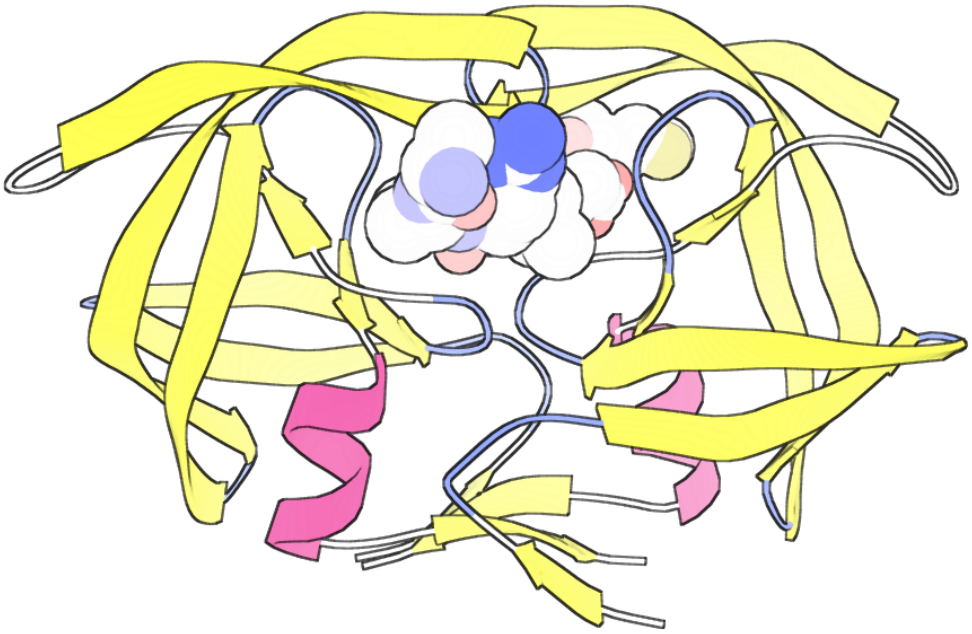


Figure 2.3: HIV protease with a bound polypeptide substrate. PDB structure 1F7A.

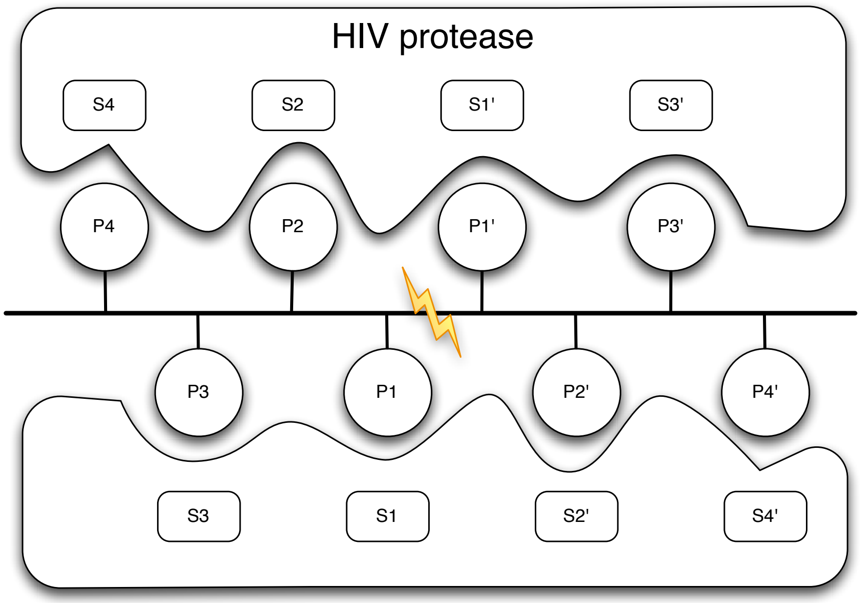
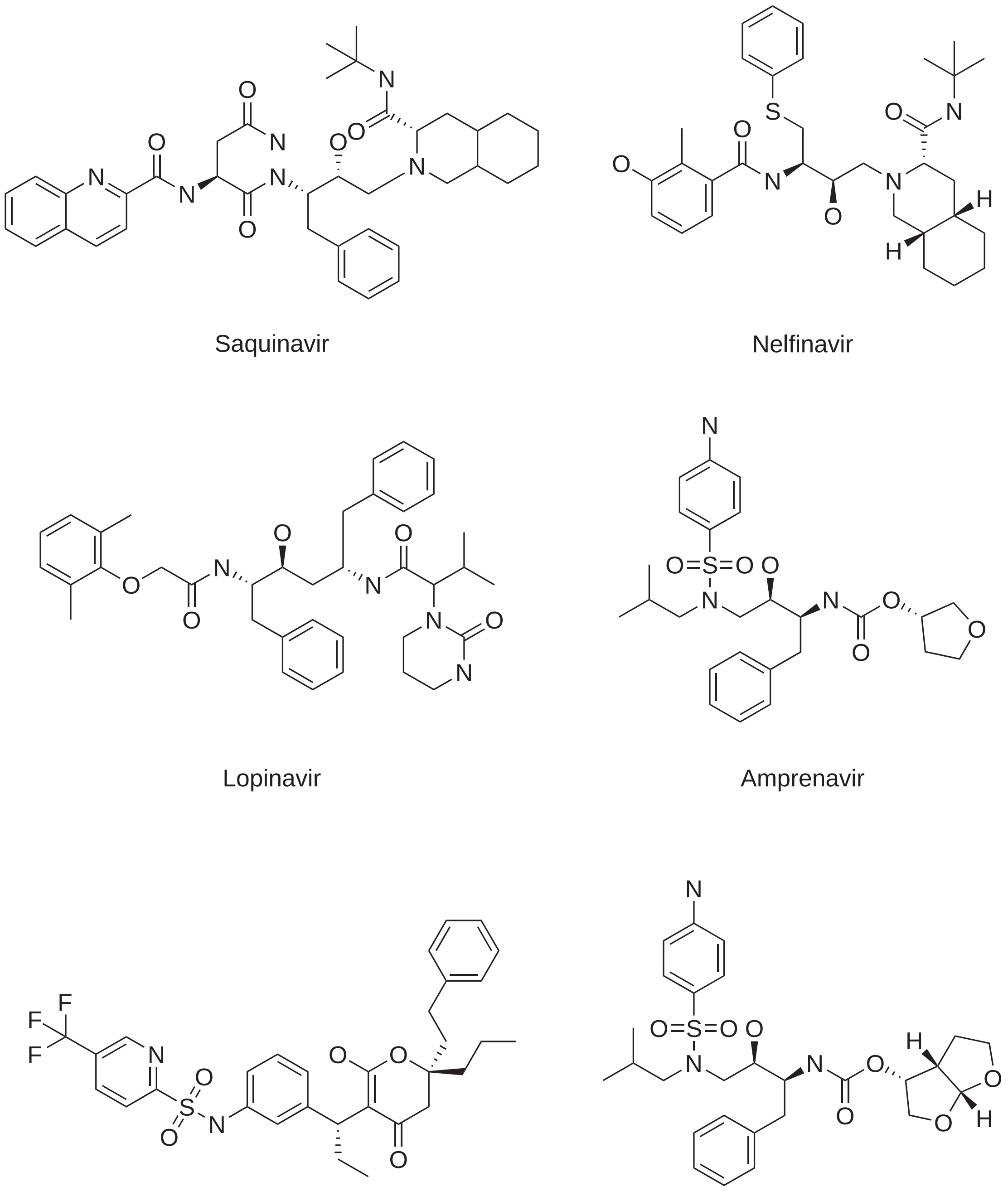


Figure 2.4: Interaction between HIV protease and a polypeptide substrate. The substrate is represented as an octapeptide, with residues P1-P4 and P1'-P4', which interact with subsites in the protease active site. Recognition of the substrate leads to cleavage of the scissile peptide bond, indicated with a lightning bolt.

99 amino acids. As shown in Figure 2.3, the HIV protease active site consists of a large cavity which can accommodate certain polypeptide substrates. Frequently, the protease active site is represented by a lock-and-key model,[7] where the active site is broken into 8 subsites, S1-S4 and S1'-S4', with corresponding substrate residues P1-P4 and P1'-P4', depicted in Figure 2.4. The cleavage event is catalyzed by the aspartic acid residues at position 25, which hydrolyze the peptide bond in the substrate between P1 and P1'. The twelve cleavage sites in HIV polyproteins comprise a diverse group of sequences,[8] and determining the specificity of protease with respect to cleaved sequences is addressed in Section 3.3.

The specificity of HIV protease in recognizing particular amino acid sequences has been exploited in rational drug design. Most current FDA-approved HIV protease inhibitors are peptidomimetic – they function by occupying the active site as a substrate would, but are not susceptible to cleavage. A similar mechanism is exhibited by non-peptidomimetic inhibitors, such as tipranavir and DMP 450. The structures of several inhibitors are shown in Figures 2.5 and 2.6. To inhibit effectively block protease function, the inhibitors must bind with greater affinity than the natural substrates, and so their K_i must fall in the low nanomolar range. In comparison, the K_m for the highest affinity protease substrate is roughly 30 μM.



Saquinavir

Nelfinavir

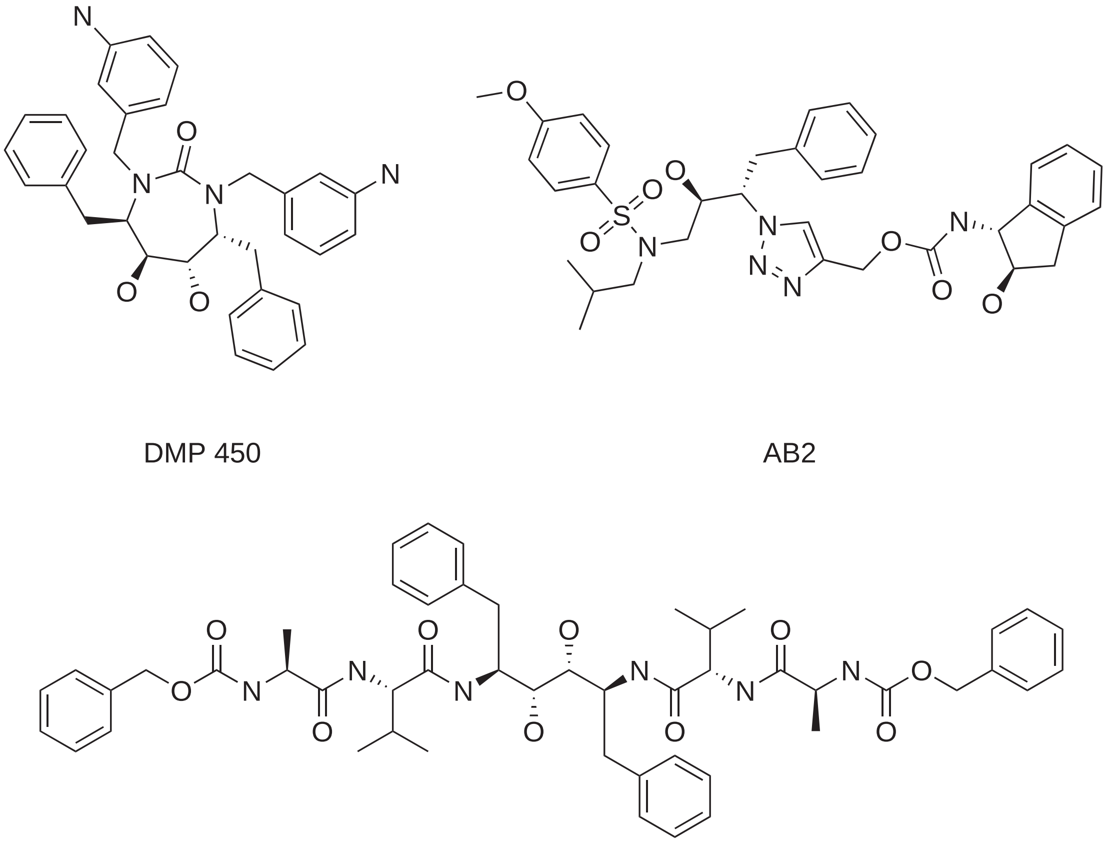
Lopinavir

Amprenavir

Tipranavir

Darunavir

Figure 2.5: FDA approved protease inhibitors.



DMP 450

AB2

TL-3

Figure 2.6: Additional protease inhibitors, not FDA-approved.

Some recent work has sought to identify alternative inhibitor binding sites outside of the enzyme's active site. Located above the active site (as shown in Figure 2.3) are mobile pairs of anti-parallel β-sheets, commonly known as "flaps.[9]" The movement of these flaps is thought to be critical in the function of protease, and a possible means of allosteric inhibition.[10] Molecular dynamics experiments have shown that restricting the motion of specific residues distant from the active site can affect flap motion,[11] indicating that a small molecule may function as an inhibitor in an area distant from the protease active site. Efforts to discover such an inhibitor are described in Section 6.3.

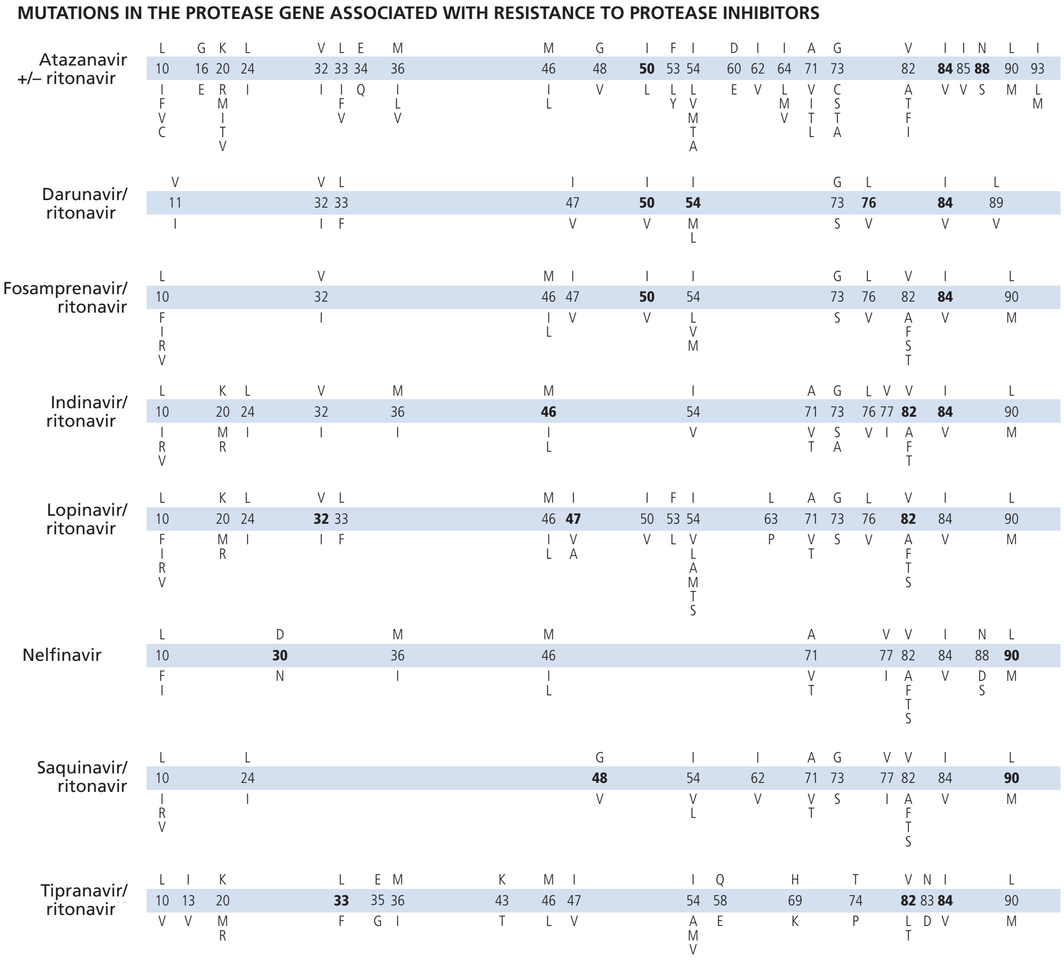
2.3 HIV Drug Resistance and Replication Capacity

During antiviral therapy, frequently specific mutations arise that render HIV less susceptible to the drugs used. Patterns of mutations vary across different drugs, but inhibitors of the same class are often thwarted by similar mutations. Mutations involved in resistance have been studied in detail for protease and RT inhibitors, which are the most widely used antiviral drugs. The most common drug resistance mutations for protease and RT are shown in Figures 2.7 and 2.8. Note that among drugs of the same class, multiple drugs may be affected by particular mutations, a phenomenon known as cross-resistance.

The resistance mutations shown in Figure 2.7 are divided into major and minor mutations. Major mutations, also referred to as primary mutations, typically arise with the onset of antiviral therapy in a drug-naive patient, and lead to reduced inhibitor binding and frequently impaired protease function as well. Minor, mutations are compensatory in nature, typically arise after primary mutations and restore some protease function. They also generally have a lesser effect on inhibitor binding or are in some cases dependent on the presence of a primary mutation.

In contrast, mutations in RT are usually classified based on their tendency to associate with certain other mutations. For example, the thymidine analogue-associated mutation (TAM) pathways shown in Figure 2.9 confer resistance against multiple nucleoside (or nucleotide) RT inhibitors (NRTI). The TAM1 and TAM2 pathways represent ordered accumulations of mutations that are often mutually exclusive. Another class of RT drug, non-nucleoside RT inhibitors (NNRTI), cause allosteric inhibition. As the mechanisms for NRTIs and NNRTIs are different, they are not affected by the same mutations.

The degree of drug resistance conferred by one or more mutations is often summarized with an IC_{50} value, which denotes the concentration of inhibitor needed to reduce the activity of an enzyme by 50%. Figure 2.10 illustrates the use of dose-response curves in calculating



MUTATIONS IN THE PROTEASE GENE ASSOCIATED WITH RESISTANCE TO PROTEASE INHIBITORS

	L	G	K	L	V	L	E	M	M	G	I	F	I	D	I	I	A	G	V	I	I	N	L	I
Atazanavir +/– ritonavir	10	16	20	24	32	33	34	36	46	48	**50**	53	54	60	62	64	71	73	82	**84**	85	**88**	90	93
	I F V C	E	R M I T V	I	I	I F V	Q	I L V	I L	V	L	L Y	L V M T A	E	V	L M V	V I T L	C S T A	A T F I	V	V	S	M	L M

	V	V	L	I	I	I	G	L	I	L
Darunavir/ ritonavir	11	32	33	47	**50**	**54**	73	**76**	**84**	89
	I	I	F	V	V	M L	S	V	V	V

	L	V	M	I	I	I	G	L	V	I	L
Fosamprenavir/ ritonavir	10	32	46	47	**50**	54	73	76	82	**84**	90
	F I R V	I	I L	V	V	L V M	S	V	A F S T	V	M

	L	K	L	V	M	M	I	A	G	L	V	V	I	L
Indinavir/ ritonavir	10	20	24	32	36	**46**	54	71	73	76	77	**82**	**84**	90
	I R V	M R	I	I	I	I L	V	V T	S A	V	I	A F T	V	M

	L	K	L	V	L	M	I	I	F	I	L	A	G	L	V	I	L
Lopinavir/ ritonavir	10	20	24	**32**	33	46	**47**	50	53	54	63	71	73	76	**82**	84	90
	F I R V	M R	I	I	F	I L	V A	V	L	V L A M T S	P	V T	S	V	A F T S	V	M

	L	D	M	M	A	V	V	I	N	L
Nelfinavir	10	**30**	36	46	71	77	82	84	88	**90**
	F I	N	I	I L	V T	I	A F T S	V	D S	M

	L	L	G	I	I	A	G	V	V	I	L
Saquinavir/ ritonavir	10	24	**48**	54	62	71	73	77	82	84	**90**
	I R V	I	V	V L	V	V T	S	I	A F T S	V	M

	L	I	K	L	E	M	K	M	I	I	Q	H	T	V	N	I	L
Tipranavir/ ritonavir	10	13	20	**33**	35	36	43	46	47	54	58	69	74	**82**	83	**84**	90
	V	V	M R	F	G	I	T	L	V	A M V	E	K	P	L T	D	V	M

Figure 2.7: Drug resistance mutations in protease. Major mutations are shown in bold. Ritonavir is generally combined with protease inhibitors in order to boost their bioavailability, not for antiviral effect. Reprinted with permission from the International AIDS Society–USA. Johnson VA, Brun-Vézinet F, Clotet B, et al. Update of the drug resistance mutations in HIV-1: Spring 2008. *Topics in HIV Medicine*. 2008;16(1):62-68. ©2008, IAS-USA. Updated information and thorough explanatory notes are available at www.iasusa.org.

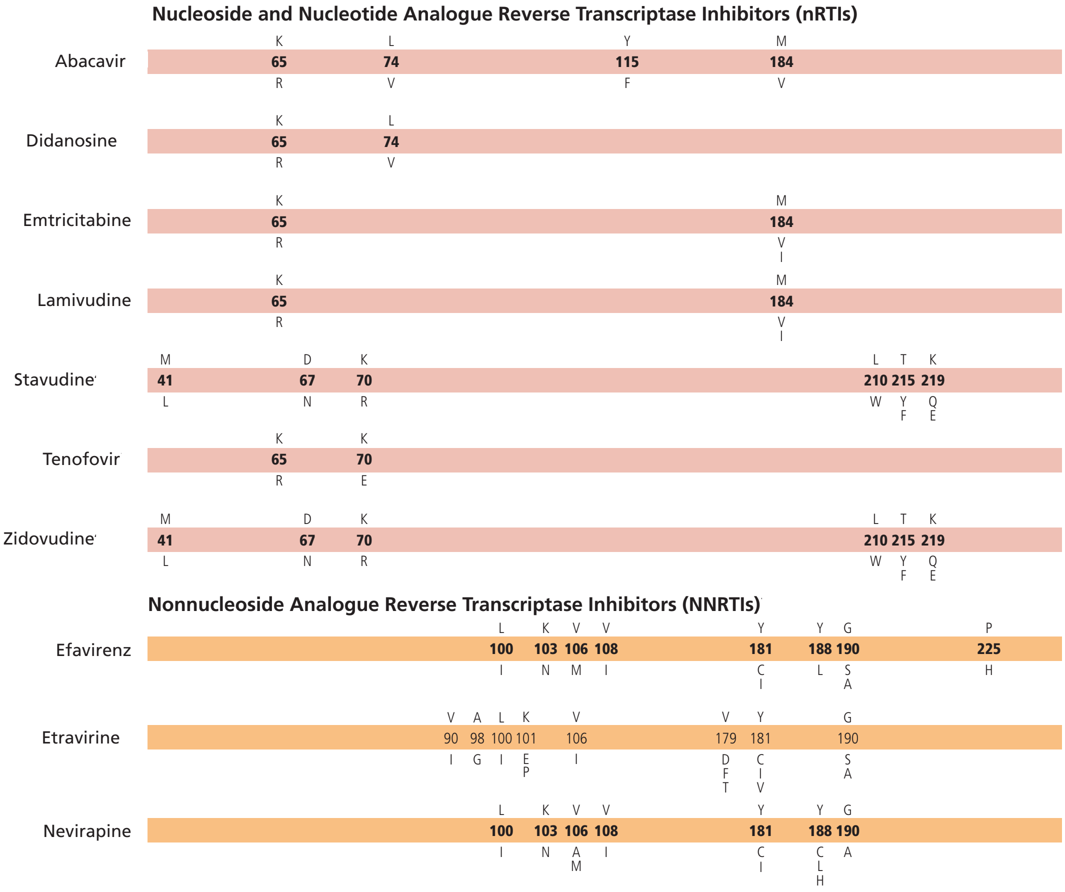


Figure 2.8: Drug resistance mutations in reverse transcriptase. Reprinted with permission from the International AIDS Society–USA. Johnson VA, Brun-Vézinet F, Clotet B, et al. Update of the drug resistance mutations in HIV-1: Spring 2008. *Topics in HIV Medicine*. 2008;16(1):62-68. ©2008, IAS-USA. Updated information and thorough explanatory notes are available at www.iasusa.org.

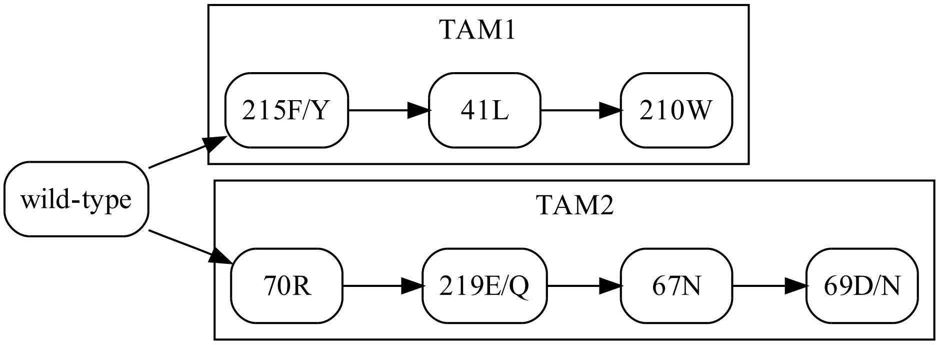


Figure 2.9: Thymidine analogue-associated mutations for drug resistance in reverse transcriptase.[12]

IC_{50} values. The information needed for these calculations can be obtained through a variety of cell-based or biochemical assays, but one of the dominant methods is a luciferase-based system developed by ViroLogic Inc. (now Monogram Biosciences).[13] The ViroLogic method involves the amplification of protease and RT genes from a viral isolate, and inserting these genes into a replication-deficient vector. The test vector contains the luciferase gene and is only able to produce infectious virus when supplied with envelope proteins from another source. In this way, replication can be reduced to a single cycle, following which luciferase activity can be measured to assess drug resistance. Since the ViroLogic assay is relatively high-throughput and consistent, it has been useful for clinicians in performing phenotypic resistance testing on their patients.

The ratio of IC_{50} values between a mutant and wild-type enzyme is referred to as the fold-change, i.e.

$$fold\,change \quad = \quad \frac{mutant\,IC_{50}}{wild-type\,IC_{50}}$$

The fold-change is also often referred to as the resistance factor (RF). In Figure 2.10, the there is a 30-fold change in IC_{50} between wild-type and mutant protease activity in the presence of the inhibitor.

Mutations that confer resistance often have consequences for viral replication capacity (RC). Often ignored in experiments concerning drug resistance, RC represents the ability of the virus to replicate in the absence of inhibitors. In some cases, HIV mutants may exhibit high levels of resistance with low RC, and are able to replicate faster than wild-type HIV in the

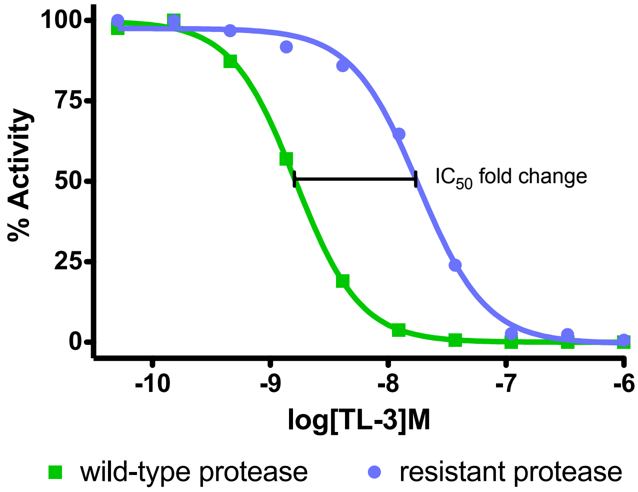


Figure 2.10: Dose-response curve demonstrating TL-3 resistance in a mutant protease. The resistant protease contains six mutations and displays more than a 10-fold increase in IC_{50} over the wild-type protease. Data courtesy of Dr. Michael Giffin (private communication).

presence of drug, but more slowly without it.[14–16] When describing the overall phenotype of the virus, some have noted that both RF and RC should be taken into account, along with the particular environment.[15] Efforts to combine RF and RC in predicting viral fitness are reported in Chapters 4 and 5.

2.3.1 Phenotype prediction

Large sets of phenotypic data have been studied in order to determine genotype-phenotype relationships. Resistance data has been researched more extensively, often focusing on a large set of resistance data with corresponding viral sequences available through the Stanford HIV Drug Resistance Database (HIVDB).[17] Various statistical and machine learning approaches have been used, including linear regression, decision trees, and support vector machines.[18–20] In these studies, the viral genotype is used to predict the logarithm of the corresponding RF values. Interestingly, for most inhibitors, linear regression is able to achieve high accuracy, comparable with more sophisticated nonlinear approaches, like support vector machines.

Alternatively, structure-based approaches have also been use to determine resistance to HIV protease inhibitors. These methods rely on molecular dynamics simulations, protein-ligand docking programs, and known protein-inhibitor structures. Beginning with a wild-type structure, specific amino acid substitutions are modeled, followed by energy minimization via molecular dynamics simulation, allowing large-scale structural shifts to accommodate the protease mutations. The free energy of binding is then evaluated by docking, which may attempt further optimization of the ligand in relation to the protease. The predictions correlate significantly with experimentally-derived free energies of protease-inhibitor complexes,[21] as well as RF values from single-cycle infection assays.[22,23]

Though far different approaches, structure- and sequence-based methods share underlying principles. The sequence-based methods seek to predict a fold-change in IC$_{50}$, which is related to the inhibition constant, K_i, by the Cheng-Prusoff equation:[24]

$$IC_{50} = K_i(1+\frac{[S]}{K_m})$$

Further, the K_i is related to the free energy, ΔG:[25]

$$\Delta G = RT\,ln(K_i)$$

which can be estimated through structure-based methods. So despite radically different methodology, the results of these approaches are essentially interchangeable. This means that structure-based predictions of changes in inhibitor binding, such as the work described above and in Section 4.2, are proportional to changes in IC$_{50}$.

A contrast becomes more apparent when considering cases with novel inhibitors or mutations. Sequence-based methods rely on finding patterns in existing data. For a new inhibitor, resistance data may not exist and are difficult to obtain in large quantities. Similarly, analyses based on *in vivo* or *in vitro* experience with approved inhibitors are obviously not able to predict the effects of mutations which have not been previously observed and measured. However, in cases where resistance information is unavailable for inhibitors or mutations, structure-based methods remain applicable. Prior structure-based studies have focused on validation of their models against existing data,[21–23] rather than attempt to predict resistance against novel inhibitors or the effects of previously unseen mutants. Chapter 4 describes the use of structural information to predict resistance mutations, culminating in validation against the novel HIV protease inhibitor AB2.

2.4 Simulations

A large class of HIV-related simulations center on the use of differential equations.[26–29] One of the seminal works in this field was published by Perelson et al.,[27] who used a system of three equations to model HIV infection *in vivo* following ritonavir therapy. Although conceptually simple, the model was able to predict viral load in patients for several days after therapy. More recent extensions to the differential equation models have incorporated more complex behaviors, such as effects of multiple inhibitor classes.[30] As a whole, differential equations models are attractive in their simplicity, and generally require few parameters. However, these models represent homogeneous populations, making it difficult to represent the mutant strains which would be expected to arise during drug therapy.

A different approach was employed in a coevolutionary study between peptidomimetic inhibitors and HIV protease.[31] This work did not explicitly model a viral population, but used a minimax approach to find inhibitors effective against a broad range of mutant proteases. A key feature of this model was its sophisticated notion of fitness, based on Michaelis-Menten kinetics. Fitness of a mutant protease was evaluated by predicting enzyme velocity:

$$v(m,i) \quad = \quad V_{max} \frac{[S]}{[S] + K_m(m) + \frac{[I]K_m(m)}{K_I(m,i)}} \tag{2.1}$$

where V_{max}, $[S]$, and $[I]$ were constant. The remaining values, $K_m(m)$ and $K_i(m,i)$, were calculated using a volume-based estimate of binding energy. $K_m(m)$ related to the affinity between the protease and nine substrate cleavage sites, while $K_i(m,i)$ described protease-inhibitor affinity. However, a major simplification in this simulation was in limiting the inhibitor molecules to octapeptides. Also, many of the resistance mutations in the study are not found in clinically-observed resistance patterns. Variants of Equation 2.1 are discussed in Section 5.2, with parameters based on analyses in Chapter 3 and 4, permitting simulations involving clinically-approved inhibitors that select for known resistance mutations.

Increases in computing power have allowed more fine-grained simulations that model the life cycle of individual virions and allow for a large number of mutations.[32,33] One group investigated the effect of multiplicity of infection on sequence diversity and the evolution of drug resistance.[32] Though the simulation modeled a each individual in a viral population, the viral genomes were represented as bit strings and the fitness values were arbitrary, unrelated to known resistance patterns. A highly sophisticated fitness function was used in another study, which sought to predict the development of resistance against a hypothetical therapy based on RNA interference.[33]

2.5 Docking in Drug Discovery

In modern drug discovery, one of the main methods for finding new inhibitors is the screening of large compound libraries against a target of interest. Since the number of compounds that will inhibit a target is generally very low compared to the number of compounds tested, huge libraries must be screened in order to find a reasonable number of potential inhibitors, or hits. For example, a screen of 400,000 compounds against protein tyrosine phosphatase-1B found 85 inhibitors with IC_{50} less than 100 μM, a hit rate of 0.021%.[34] Only 6 compounds, or 0.0015%, were able to meet a more stringent cutoff of 10 μM. A smaller-scale screen of compounds against HIV protease used a library of 2,000 compounds.[35] The authors claimed a single hit from this screen, a molecule known as sanguinarine, which is highly toxic to humans. From these studies, two major difficulties are evident: the low probability of finding a hit and finding hits without undesirable properties (e.g. toxicity).

As a complementary approach, computation is often employed in a process known as virtual screening.[36–38] In virtual screening, the binding affinity estimates are calculated between large compound libraries and a target of interest. One of the primary methods for performing such calculations is the use of protein-ligand docking programs, such as AutoDock.[39,40] AutoDock functions by optimizing a flexible ligand molecule in relation to a rigid macromolecule, typically a protein. The docking process can be viewed as consisting of two main parts: (1) evaluating the binding energy of a given conformation and (2) searching possible conformations. The binding energy calculation is based on the AMBER force field[41] and takes into account 5 terms:

1. van der Waals forces
2. electrostatic interaction
3. hydrogen-bond formation
4. desolvation
5. a torsional free energy penalty

Evaluating the binding energy of all ligand conformations is difficult, since the number of conformations grow exponentially with the number of rotatable bonds in the ligand. Performing an exhaustive search of all possible ligand conformations is generally computationally intractable, so AutoDock employs a genetic algorithm, which samples many conformations, but biases its search toward regions near more favorable solutions. There is no guarantee of finding the lowest energy conformation, so many independent docking runs are generally used. Consistent docking results, as determined through cluster analysis, have been correlated with more favorable binding energies.[42]

Ideally, the predicted ligand conformation with lowest binding energy should be the preferred binding mode, and the predicted binding energy should match experimental results. In practice, this is often not the case. Under cross-validation, AutoDock 4 has a standard error of roughly 2.5 kcal/mol on a set of various protein-ligand complexes.[40] Potential improvements to the energy calculation have focused on the role of entropy. The torsional free energy penalty is a rudimentary estimate of the configurational entropy of the ligand, and is simply proportional to the number of rotatable bonds in the ligand. This term does not take into account the ligand conformation or any constraints imposed by the proximity of the protein.

Recent work has theorized a correspondence between configurational entropy and AutoDock clustering.[43–45] The authors reasoned that clusters of similar conformations represented a mixture of states that could be used to estimate configurational entropy. As suggested by Figure 2.11, isolated conformations may exhibit highly favorable binding energies without being close to the observed conformation. Essentially, groupings of similar conformations indicate entropic favorability, which can be used to augment the AutoDock binding energy prediction. Applying this rescoring to a test set of protein-ligand structures significantly improved the chances of finding the experimentally observed conformation.[43] Use of the configurational entropy improving the correlation between predicted and observed binding energies, however, was not investigated. Extending AutoDock clustering to provide an empirical estimate of entropy for use in binding energy calculations is described in Section 6.2. Ultimately, a major goal of these entropy estimates is the improvement of accuracy in virtual screening, supporting the discovery of new HIV drugs. This is documented in Chapter 6, with discovery of several HIV protease inhibitors.

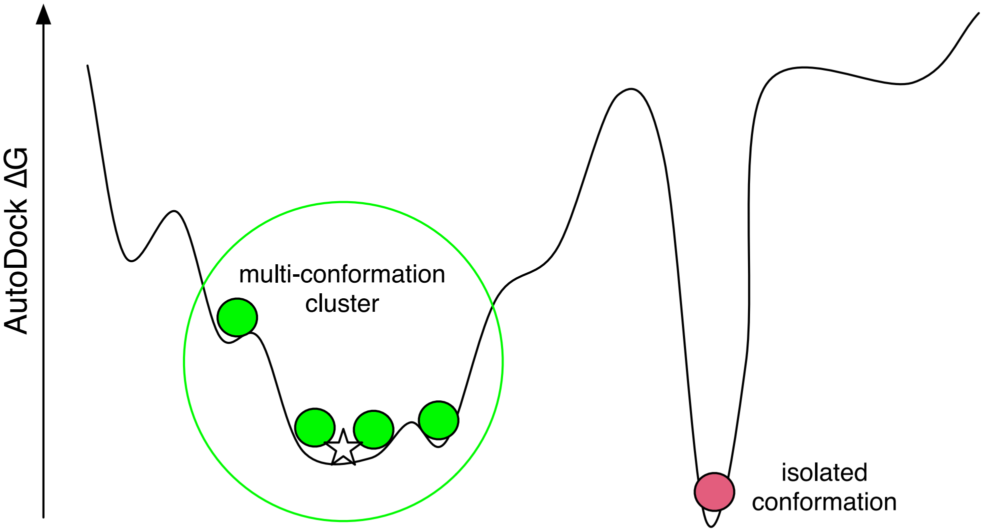


Figure 2.11: An example of a protein-ligand energy landscape. Circles correspond to AutoDock results and the star represents the experimentally observed conformation. AutoDock will often find isolated low-energy minima (red) that inhabit a “narrow” energy well. Seemingly less favorable conformations (green) may actually be more representative when entropic contributions are taken into account.

Chapter 3

Determinants of Viral Replication Capacity

The ability of HIV to replicate quickly within a host is an important factor during infection and therapy. Mutations arising during treatment are often noted for their effects against inhibitors, but the presence of these mutations can also affect replication capacity. The most straightforward approach takes advantage of data obtained via *in vitro* single-cycle infection assays[13] and applying various statistical or machine learning methods.[46,47] However, as shown in Section 3.1, this type of approach is imprecise. Another weakness stems from the limited diversity of the viral protein sequences used. Relying on sequences derived from clinical sources biases the observed viral sequences towards solutions which contribute to drug resistance for current inhibitors. A purely sequence-based statistical method would be unable to make predictions for any mutations which had not been previously observed, limiting its usefulness.

Ideally, a predictor for viral replication capacity would be able to predict the effect of any possible mutation. Generalizing a protein sequence-based analysis to incorporate more general structural properties of residues allows for a greater range of predictions. The Multivariate Analysis of Protein Polymorphism (MAPP) program was designed to perform this task.[48] Its authors previously analyzed both HIV protease and reverse transcriptase, finding that MAPP was able to qualitatively discriminate between neutral and deleterious mutations. This work is extended here with additional sequence data. A correlation between MAPP scores and protease catalytic efficiency is demonstrated in Section 3.2. The predicted variability in various regions of protease is also examined, with implications for drug design.

Mutations not directly involved in enzyme function can also affect viral replication, as is the case with HIV protease cleavage sites.[49–54] Through alterations in the amino acid sequence of these cleavage sites, the rate of substrate processing can be modified, allowing the virus to replicate effectively despite the presence of inhibitors. Understanding the substrate specificity of HIV protease and the range of allowable substitutions in the cleavage sites will aid in anticipating the emergence of these mutations during drug therapy. An interpretable model of protease cleavage based on the physicochemical properties of the residues at each cleavage site generated using machine learning software is presented in Section 3.3.

3.1 Sequence-based regression

Viral replication capacity (RC) measures the ability of a virus to replicate in the absence of inhibitors. The use of drugs to treat HIV infection leads to the evolution of resistance mutations, which often have consequences for replication capacity. Given viral protein sequences and the corresponding RC values, it is possible to apply machine learning methods to predict RC from a protein sequence.[46,47] A dataset from ViroLogic Inc. contains RC measurements for 317 HIV isolates obtained via single-cycle infection assays.[46] Amino acid sequences are available for positions 4-99 for protease and 40-233 for reverse transcriptase. RC values range from 0.26 to 151, with a value of 100 indicating wild-type.

As the replication capacity of the virus depends on both the protease and reverse transcriptase enzymes, the dimensionality of this dataset is large. When considering the mutations at all positions available, there are 807 attributes. In protease, variability was observed at 90 positions, with an average of 2.7 mutations per position. 183 positions showed variability in reverse transcriptase, and there were an average of 3.1 mutations at each position. The high dimensionality suggests the use of a support vector machine (SVM), which is a machine learning technique suitable for high dimensional data.[55] However, since the number of attributes still exceeds the 317 instances available, choosing a subset of the features is necessary to obtain accurate predictions. To narrow the mutations considered, a set of treatment-selected mutations (TSM) were considered, which were previously used in predictions of drug resistance.[20] These mutations, which occur significantly more frequently in treated than untreated populations, reduce the number of attributes to 61. This smaller dataset contained 30 protease positions with an average of 1.3 mutations at each position and 23 reverse transcriptase positions with an average of 1.3 mutations per position.

3.1.1 Machine learning details

The Weka machine learning library was used to carry out the predictions.[56] The SMOreg (Sequential Minimal Optimization Regression) module was used to perform SVM regression with a radial basis function (RBF) kernel. In addition, two other methods were included for comparison, linear regression and ZeroR. The linear regression implementation used a ridge regression parameter of 10^{-8}. ZeroR is a naive method which predicts the mean RC value in each case, completely ignoring the mutations present, providing a baseline level of error.

Each clinical isolate in the dataset was represented as a series of binary values indicating the presence or absence of specific mutations, and linked to the associated RC value. Experiments were also carried out using the natural log of the RC, as some studies have used this transformation to reduce the effect of outliers.[47] In all cases, results were obtained using 10 rounds of 10-fold cross-validation.

3.1.2 Results and discussion

The most naive type of regression predictor, ZeroR, uses the mean of the training data RC values, ignoring the mutations present. This strategy obtained root mean square error (RMSE) of 25.2, with an insignificant correlation coefficient. When the RC values were log-transformed, the RMSE obtained was 0.953. These errors indicate a baseline value that the other learning methods should surpass. The performance of linear regression and SVM regression shown in Tables 3.1 and 3.2. In the case of linear regression, use of the full set of mutations resulted in extremely poor accuracy. When restricting features only to the TSMs, accuracy improved, but was still comparable to the naive learner when the natural logarithm was not applied. The SVM results showed improvement over linear regression, though the error remained high. When using the full set of mutations, the standard deviation was larger than with the TSM subset, possibly indicating over-fitting during cross-validation.

Table 3.1: Replication capacity prediction using linear regression.

Data set	Correlation coefficient	Root mean square error
RC	0.093 ± 0.045	121 ± 10.5
log(RC)	0.110 ± 0.030	4.91 ± 0.336
RC-TSM	0.367 ± 0.015	24.2 ± 0.363
log(RC)-TSM	0.529 ± 0.010	0.834 ± 0.009

The performance of the SVM was similar to that reported in previous studies. The best RMSE value obtained by Segal et al. with a random forest approach was 24.0, using untransformed RC values.[46] Birkner et al. used log-transformed RC values and a data-adaptive

Table 3.2: Replication capacity prediction using a support vector machine.

Data set	Correlation coefficient	Root mean square error
RC	0.547 ± 0.167	20.8 ± 2.83
log(RC)	0.699 ± 0.107	0.679 ± 0.095
RC-TSM	0.431 ± 0.041	22.9 ± 0.433
log(RC)-TSM	0.582 ± 0.045	0.789 ± 0.030

regression, obtaining RMSE 0.688.[47] In absolute terms, however, the accuracy of all methods on un-transformed RC values was dubious, with RMSE values only slightly better than guessing the mean RC.

3.2 Predicting impairment in HIV protease

As an alternative to directly estimating viral replication capacity the activity of individual genes can be investigated. In general, substitutions from the wild-type protein sequence will not have a positive effect, so the goal lies in predicting the level of impairment caused by particular mutations. A simple approach to this problem would apply a amino acid substitution scoring matrix, such as BLOSUM62, to classify changes in a given sequence.[57] However, applying a general scoring matrix ignores any characteristics specific to the protein or position of interest.

More sophisticated techniques generally rely on a multiple sequence alignment of related sequences and are able to make more detailed predictions. The SIFT program, written by Ng and Henikoff, used a set of orthologous sequences to generate probabilities for all possible substitutions at each position in a sequence.[58] Improvements over more general scoring schemes, such as the BLOSUM62 matrix, were demonstrated on three different systems, including HIV protease. More recent work by Stone and Sidow extended this approach with the Multivariate Analysis of Protein Polymorphism (MAPP) program, which was shown to have even higher accuracy than SIFT.[48] High MAPP scores indicate that the physicochemical properties of a putative mutation are markedly different than expected based on an alignment of orthologous sequences. An examination of the MAPP scores for known drug resistant mutants revealed that several major mutations with high scores were present at low frequency in untreated patients, but much higher frequency in treated patients, indicating that MAPP scores capture some aspects of protease fitness. Both programs are able to realize gains over the generalized BLOSUM62 matrix due to system- and position-specific scoring.

The following three sections first describe an effort to augment the original MAPP

Table 3.3: Comparison of MAPP prediction accuracy based on experimental data. Sensitivity and specificity, respectively, are shown in brackets.

	Positive vs. Deleterious	Intermediate vs. Negative
original MAPP paper (n=39)	80.4% [64.8% 87.7%]	76.7% [61.7% 89.5%]
PSI-BLAST set (n=74)	74.1% [79.8% 71.4%]	78.3% [85.3% 70.4%]

analysis of HIV protease by including a larger set of related sequences, which shows improvement in discriminating among clinically-relevant protease mutations. Next, the predictions are also shown to correlate with experimentally-observed measures of protease catalytic efficiency. Finally, following the work of Wang and Kollman,[59] along with predicting functional impairment, the degree of variability at each position are used to provide insights into drug design.

3.2.1 Sequence variability

A diverse set of protease homologs was obtained by querying the Swissprot database using PSI-BLAST with the HXB2 HIV protease sequence, yielding 103 sequences[*]. After generating a multiple alignment using ClustalW, 29 of these sequences were eliminated due to poor alignment, leaving 74 sequences. This refined set of sequences was used to generate a phylogenetic tree using the Semphy program[60][†]. The sequences and tree were input to the MAPP program to generate estimates of all possible substitutions on protein function.

In comparison, the original research by Stone and Sidow used only 39 sequences related to HIV protease. To validate their predictions, the authors attempted to classify 330 sequences with experimentally determined phenotypes.[48,61] Since the experimental data used a tripartite classification scheme differentiating between wild-type, intermediate, and negative phenotypes, Stone and Sidow performed two binary classification experiments. They reported 80.4% accuracy in differentiating between positive and deleterious mutations (i.e. wild-type vs. intermediate or negative phenotypes) in HIV protease, and 76.7% for discrimination between intermediate and negative mutations (wild-type or intermediate vs. negative phenotypes). In both cases, MAPP score thresholds at values of 5 and 7 were used to indicate positive, intermediate, and negative predictions, respectively.

With the 74 protease sequences harvested from PSI-BLAST and a phylogenetic tree built using SEMPHY, accuracy for differentiating positive and deleterious mutations was 74.1% (Table 3.3), using the same thresholds. This accuracy is significantly lower than what was obtained using the smaller set of sequences, however, classification between intermediate and neg-

[*] Default PSI-BLAST parameters were used. Search was terminated after 2 iterations and used an E-value cutoff of 1e-8.

[†] Homogeneous ASRV, “classic” joining, and JTT model were specified.

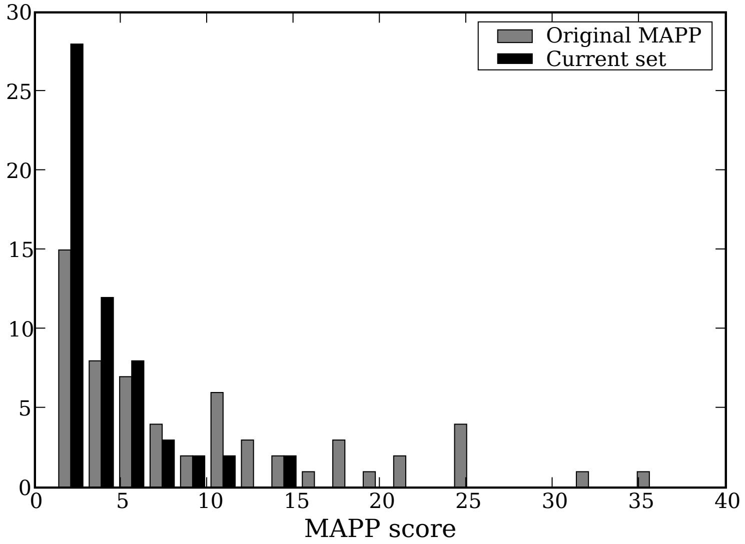


Figure 3.1: MAPP score distributions for clinically-observed mutations associated with protease inhibitor drug resistance. Higher MAPP scores predict deleterious mutations. The gray histogram shows the distribution of scores for the sequences from the original MAPP paper, and the black histogram reflects sequences gathered using PSI-BLAST.

ative mutations improved slightly, to 78.3%. Some of this discrepancy can be accounted for by considering trade-offs between specificity and sensitivity. By examining a larger set of sequences, the model tolerates greater variability, manifested as lower MAPP scores. The average score from Stone and Sidow's model was 13.7, while the PSI-BLAST set's average was 9.48. The median score dropped from 10.8 to 5.23 as well. Since the same thresholds were used, a larger number of tolerated mutations were likely to be predicted, leading to an increase in false positives and decrease in false negatives. This was supported by the improved sensitivity and reduced specificity of the new set of MAPP scores relative to the originals.

Differences between the models become more obvious when clinical data is considered using a set of treatment-selected mutations (TSM), which are mutations likely to arise during drug therapy.[20,62] As the TSMs were discovered through analysis of clinical samples, it follows that their protease function is not severely impaired, and that the MAPP scores should be relatively low. The score distributions for TSMs associated with protease inhibitors are shown in Figure 3.1. The distribution of original MAPP scores contained a larger number of high values, predicting functional impairment when the TSMs indicated otherwise. In this regard, our larger set of protease sequences was noticeably superior.

3.2.2 Processing rates

Moving beyond qualitative comparisons, the most potentially useful aspect of MAPP is its ability to make quantitative predictions on the level of protein function. Stone and Sidow did find a significant negative correlation between MAPP score and enzymatic activity in HIV reverse transcriptase ($r = -0.56$), but lacked the data for a similar analysis of protease activity. For protease, activity can be defined in terms of substrate cleavage rates, which can be determined from enzymatic k_{cat} and K_m values. K_m relates to the affinity of an enzyme for substrate, while k_{cat} indicates the number of substrate molecules processed per unit time. Relative processing capability is considered as the k_{cat}/K_m ratio of a particular mutant against wild-type protease. Previous work has demonstrated that protease is able to tolerate a large degree of impairment while still allowing viral replication, however.[63] Even when the k_{cat} of protease is decreased to one-fourth of normal, as with a T26S mutant, only slight decreases in infectivity are observed, although decreasing activity 50-fold is sufficient to halt viral replication, as demonstrated with the A28S mutation.

Using data from several biochemical analyses of protease activity,[63–67] the correspondence between MAPP scores and relative processing rates are shown in Figure 3.2. As expected, there is a negative correlation, with $r = -0.54$ ($P = 0.019$), indicating that MAPP scores are useful for making quantitative predictions of protease function. Using Stone and Sidow's MAPP predictions, a weaker, statistically-insignificant results is obtained, with $r = -0.15$.

3.2.3 Implications for drug design

Related work performed by Wang and Kollman examined the molecular basis of drug resistance in HIV protease by studying both sequence variability and the results of molecular dynamics simulations with substrates and inhibitors.[59] In this model, sequence variability, determined via multiple sequence alignment of HIV protease orthologs, was used to estimate changes that could affect replication capacity. Low sequence variability implied that the residue was important for catalytic ability or structural stability.

The distribution of MAPP scores has implications for drug design, since inhibitor interaction with more conserved residues may hinder the development of resistance. The per-position distribution of Wang and Kollman's "variability" measure and MAPP scores can be seen in Figure 3.3. In general, positions in the substrate cleft are highly conserved, as evidenced by the magnitude of their average scores. The catalytic triad and neighboring residues represent a cluster of especially important residues. However, in some cases, the average MAPP score can

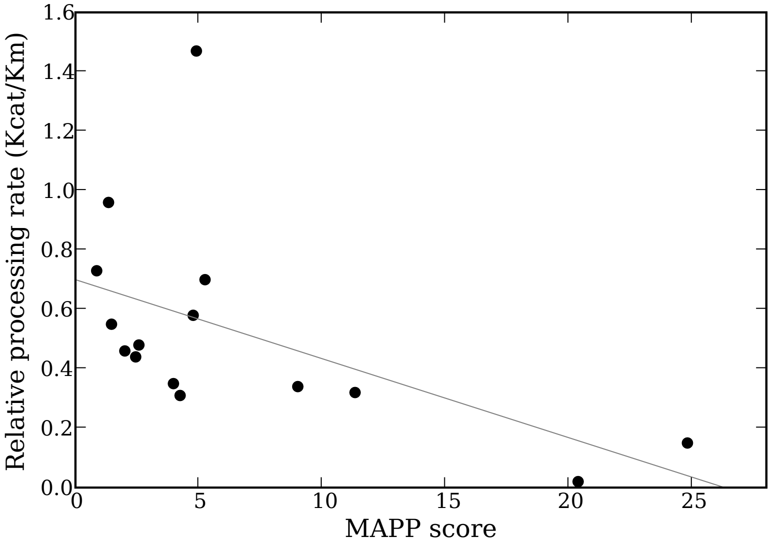


Figure 3.2: Relationship between MAPP scores and experimentally-derived protease processing rates. The gray line indicates the linear least-squares fit ($r = -0.54$).

be misleading, for instance at position 90. Though the average score indicated a highly conserved residue, the L90M mutation is a very common drug resistance mutations that affects nearly all inhibitors. Further, with a score of 4.8, the specific mutation to methionine was predicted to have little or no effect on protease function. Instead, looking at the most favorable MAPP score of each mutation is a more stringent standard for determining per-position variability which can avoid this problem.

Based solely on this sequence analysis, it appears that positions 9, 23, 25-29, 49, 52, 86-87, and 97 would be the best positions to focus the binding of new protease inhibitors. However, only a subset of these positions are accessible so as to be able to interact directly with ligands. Further insights can be found in structure-based approaches that involve protein-ligand binding predictions.

3.3 Protease cleavage prediction

Current HIV protease inhibitors are designed to occupy the enzyme's active site, preventing its natural substrates from being processed. If protease does not cleave at several specific points in HIV polyproteins, non-infectious particles are produced.[7] Due to its vital function, HIV protease has become an important drug target, and to further understanding of the protease active site, experimental and computational studies have been undertaken to find "lock and key"

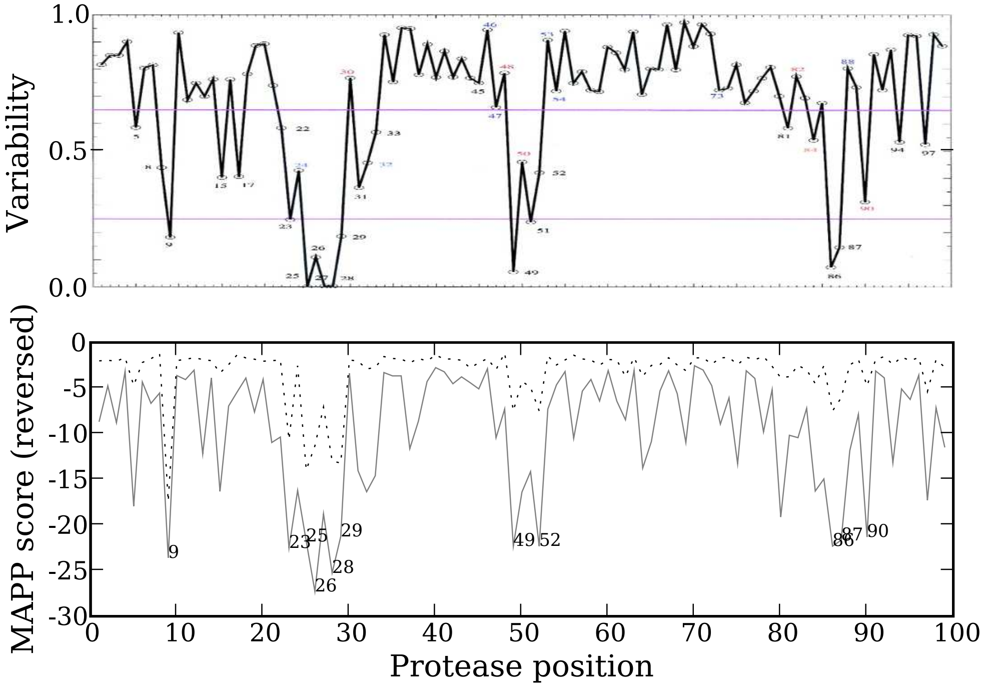


Figure 3.3: Sequence variability as calculated by Wang and Kollman (top) and average/minimum MAPP scores for each protease position using the expanded sequence set (bottom; solid gray and dotted black, respectively). The minimum score corresponds to the mutant residue with the lowest MAPP score (i.e. the most favorable change). [Reproduced with permission, Copyright PNAS 2001]

relationships between a sequence of amino acids as they fit into the active site of the protease. It is hoped that identifying important features of cleaved sequences will aid drug design.

As shown in Figure 2.4, the protease cleavage mechanism is typically represented by an 8-residue substrate (P4, P3, P2, P1, P1', P2', P3', and P4') that interacts with 8 sites in HIV protease (S4, S3, ... S4'). A wide range of machine learning studies have attempted to find patterns in protease cleavage data.[68–74] In all cases, the data instances are based on the 8-residue substrates, encoded as either a sequence of letters or a 160-dimensional vector (20 possible amino acid residues at 8 positions).

In all prior work, training data is taken from the set published by Cai *et al.* (1998), which contains 114 cleaved octapeptide instances and 248 non-cleaved instances. Experiments reported here also incorporate data available from experimental studies,[75,76] which encompass an additional 82 cleaved and 16 non-cleaved instances, for a total of 460 instances.

Previous learning methods used to find patterns in these data include support vector machines (SVM), artificial neural networks (ANN), and decision trees (DT). SVMs and ANNs are able to classify approximately 90% of instances correctly, but interpreting support vectors and multi-layer neural networks is difficult. The DT-based rules produced by Narayanan et al. are easily interpreted, but only about 85% accurate.[70] More recently, Rognvaldsson and You applied linear classifiers to generate models as accurate as the more sophisticated non-linear classifiers,[73] achieving greater than 90% accuracy under 10-fold cross-validation.

However, a common thread between all previous studies was their limited feature set, which only encoded sequence information. In seeking insights for drug design, it is important to take into account the physical properties of the peptide sequences. This work focuses on improving the accuracy and interpretability of predictions by incorporating additional features based on the physical properties of the substrate sequences.

3.3.1 Methods

Representing Physical Properties

When classifier required numerical input (as with the SVM and perceptron), sequences were encoded as 160-dimensional vectors. This process and its consequences have been described in detail previously.[73] DT generation allows discrete values, so sequences were represented as 8 separate amino acids. In these experiments, the octapeptide sequence information associated with each cleaved/uncleaved instance was augmented with two types of physical properties: amino acid residue properties and whole-peptide properties. The amino acid residue properties were:

- the residue mass (Daltons),
- volume (Å^3),
- surface area (Å^2),
- polarity (polar or nonpolar),
- charge (+, -, or neutral),
- aromaticity (aromatic or non-aromatic),
- aliphaticity (aliphatic or non-aliphatic),
- length (defined as the longest chain of heavy atoms), and
- length divided by mass

The whole-peptide features are determined by constructing a SMILES representation of the peptide chain and processing via JOELib[‡]. The 32 whole-peptide descriptors include such values as inventories of particular molecules, number of hydrogen bond donors and acceptors, and total mass. Some geometry-based JOELib features were unreliable, as the SMILES representation is only 2-dimensional.

Classifiers

Because previous work showed that simple linear classifiers could perform competitively over sequence features,[73] the MATLAB Neural Network Toolbox implementation of a perceptron network was included in this study. As gradient learning methods like the perceptron are known to be subject to convergence to local minima, a type of iterated hill-climbing was used so that the network did not remain trapped in local minima. 10 "iterated restarts" at random locations were used to provide a more robust estimate of classifier accuracy. This resampling occurred during each of 10 folds of cross-validation, keeping only the most favorable result. The cross-validation accuracy was then averaged over 10 rounds.

The SVM implementation was supplied via LIBSVM,[77] which includes facilities for scaling data, cross-validation, and parameter selection. To determine which features were most

[‡] http://www-ra.informatik.uni-tuebingen.de/software/joelib/index.html

Table 3.4: Protease cleavage prediction accuracy.

Learning Method	Feature Set	Mean Accuracy (%)
SVM	sequence	92.9 ± 0.82
	sequence + physical	90.2 ± 0.53
	sequence + whole	92.6 ± 0.75
	sequence + physical + whole	89.7 ± 0.84
	physical	88.2 ± 0.75
	physical+ whole	88.6 ± 0.80
	whole	86.6 ± 0.74
DT	sequence	87.1 ± 1.1
	physical	87.5 ± 1.4
Perceptron	sequence	94.8 ± 0.70
	physical	88.5 ± 0.89

informative for an SVM, the accuracy of various feature sets was compared using cross-validation. Each feature set was scaled using the LIBSVM's svm-scale program. For each set of features, results were calculated using a linear kernel and 10 rounds of 10-fold cross-validation.

C4.5 was used for decision tree generation.[78] For sequence-based trees, discrete value subsets were enabled, as the large number of discrete values can quickly fragment the data, greatly limiting branching. Otherwise, default parameters were used for tree construction in our experiments. However, for the sample trees shown, branch pruning was increased to allow a more compact display. The bundled xval script was used to perform 10 separate rounds of 10-fold cross-validation.

3.3.2 Results

Support Vector Machine

Table 3.4 summarizes the results of SVM training using various feature subsets. "Sequence" implies only the eight amino acid features were used, "physical" refers to site-specific physical features of individual amino acids, and "whole" refers to physical features of the full peptide. Surprisingly, with SVMs using sequence features alone gives the highest mean accuracy; physical features do not provide the SVM any additional benefit. Experiments with other kernels showed similar results (data not shown).

Decision trees

One key advantage of SVMs is their robustness in the face of "the curse of dimensionality": because the search for separating hyperplanes occurs in a dual space, they can find regularities among training instances even when these are represented in the primal space using large numbers of features. However, for inductive methods such as DTs which search through primal space directly, the inclusion of increasing numbers of features without exponentially more training data makes their search much more difficult. For this reason, to maintain approximately constant feature spaces, in DT experiments whole-peptide features were dropped, and comparison was made between sequence-only and site-specific physical features only.

As shown in Table 3.4, trees based on physical features had comparable cross-validation accuracy to sequence-based representations discovered by our experiments. These experiments are also consistent with the DT-based accuracies reported previously.[70] As shown in Figures 3.4 and 3.5, however, the type of DT formed over the two feature sets is much different. The tree based on physical features makes use of only P2, P1, P1', and P2', ignoring the residues farther from the cleavage site. Rognvaldsson and You found similar behavior with their perceptron model, noting that some residues distant from the cleavage site could be ignored with no penalty.[73] Both trees contain similar branching at the root, depending on P1.

Perceptron

The previous perceptron classifier experiments showed that this system reliably converged to a zero-error solution, demonstrating (via the Perceptron Convergence theorem) that this data was linearly separable.[73] The present study included the additional instances culled from experimental studies,[75,76] but the perceptron classifier over sequence features again converged to a zero-error solution, indicating that the data set remains linearly separable with this new data. Cross-validation accuracy of the perceptron using sequence features is similar to that of the SVM, with an average of 94.8% correct.

Using physical features, the training error did not converge to 0, even when given additional training time, suggesting that the data set is not linearly separable when site features alone are used. Mean accuracy using the perceptron classifier was 88.5%, comparable to the decision tree. The slightly higher accuracy is likely due to the iterated restart procedure used, which effectively allows more trials to be performed.

The matrix-capsid site

In order to produce active virions, protease must first cleave itself out of the Gag-Pol polyprotein and then successfully accomplish cleavage to produce the matrix, capsid, nucleocapsid, reverse transcriptase, and integrase proteins. Each of these natural substrates presents slightly different constraints on a protease capable of effectively cleaving them all, and yet pro-

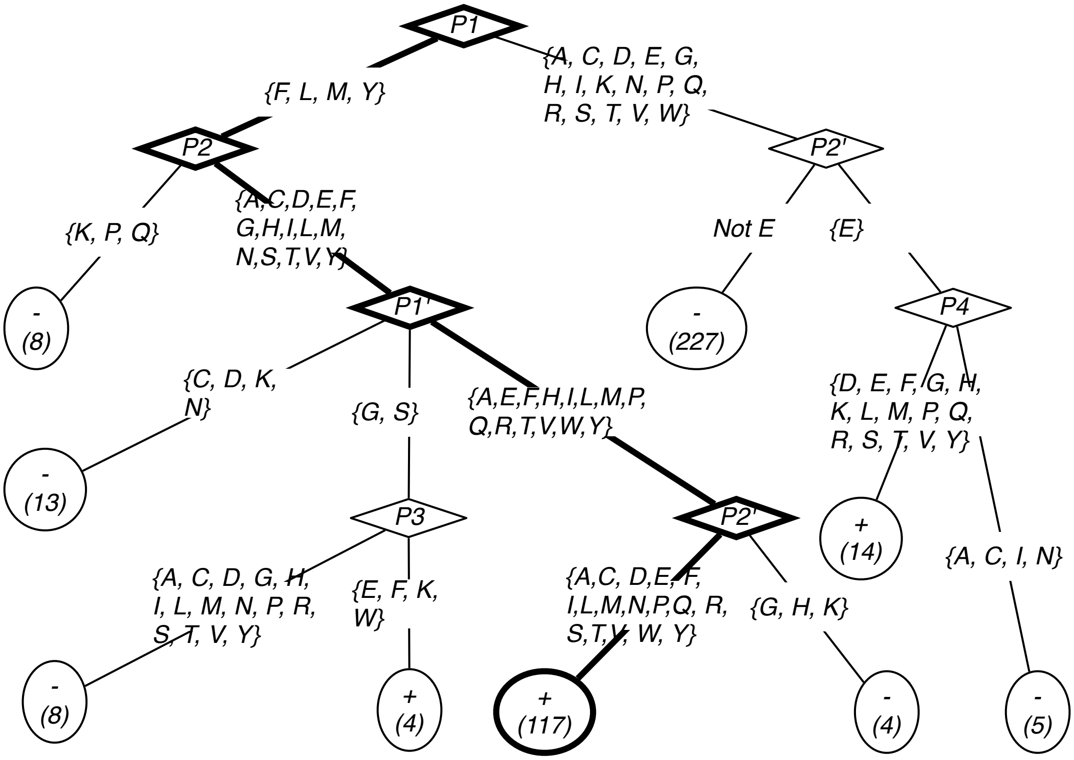

P1
{F, L, M, Y}
{A, C, D, E, G, H, I, K, N, P, Q, R, S, T, V, W}
P2
P2'
{K, P, Q}
{A,C,D,E,F, G,H,I,L,M, N,S,T,V,Y}
Not E
{E}
- (8)
P1'
- (227)
P4
{C, D, K, N}
{G, S}
{A,E,F,H,I,L,M,P, Q,R,T,V,W,Y}
{D, E, F, G, H, K, L, M, P, Q, R, S, T, V, Y}
- (13)
P3
P2'
+ (14)
{A, C, I, N}
{A, C, D, G, H, I, L, M, N, P, R, S, T, V, Y}
{E, F, K, W}
{A,C, D,E, F, I,L,M,N,P,Q, R, S,T,V, W, Y}
{G, H, K}
- (8)
+ (4)
+ (117)
- (4)
- (5)

Figure 3.4: Sequence-based substrate cleavage decision tree.

(+) and (-) inside round nodes represent cleaved and uncleaved instances, respectively. The numbers in parentheses indicate the total number of instances classified by a node. Branches and nodes shown in bold show the path which classifies most of the cleaved instances.

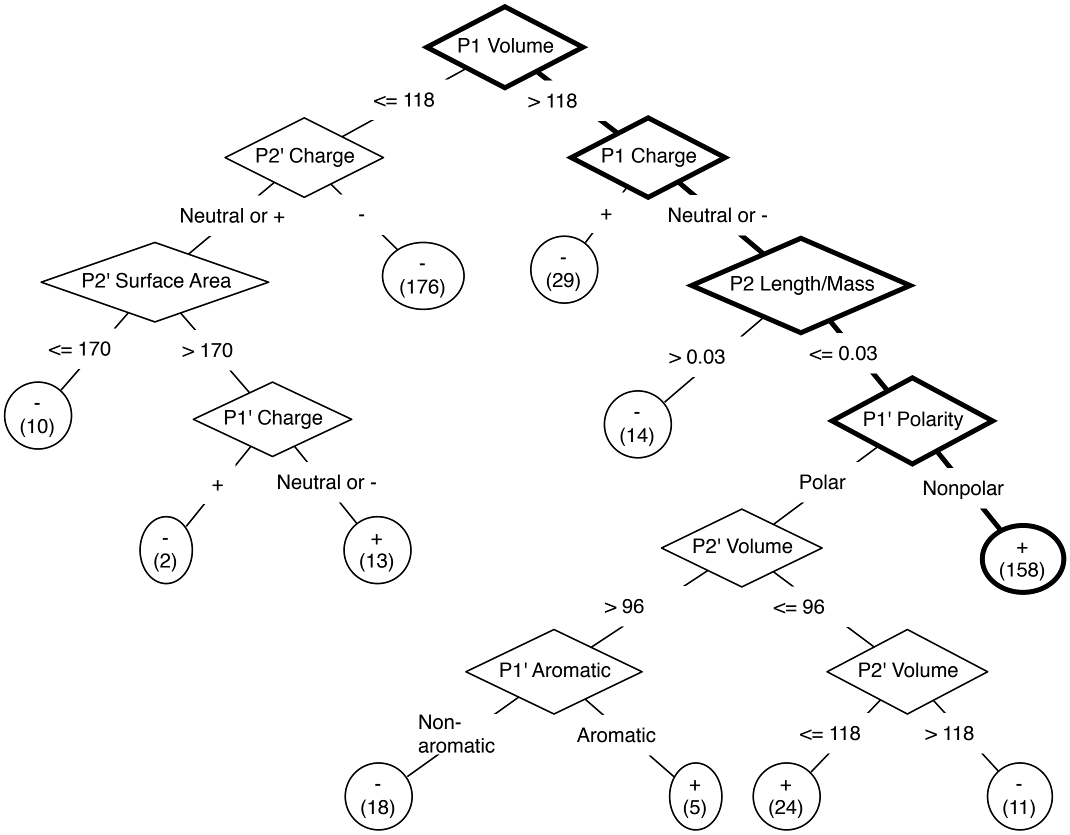


Figure 3.5: Substrate cleavage decision tree based on physical features. Units are as follows, volume: Å^3 , surface area: Å^2, and mass: Daltons.

tease does not cleave indiscriminately.[75] For these reasons, this study investigates the potential structure among various subsets of the cleaved/uncleaved training data, as perhaps indicating features especially relevant to cleavage at *individual* sites.

On closer inspection, it was found that the matrix-capsid cleavage site – SQNYPIVQ – and sequences very similar to it, were particularly well-represented within the existing training data. Table 3.5 enumerates the 27 cleaved and 9 uncleaved sequences with significant "overlap" with the natural substrate; i.e., simply the number of shared amino acids with the naturally occurring substrate.[§] Interestingly, there were nine uncleaved sequences which differed from the natural (cleaved) substrate by only a single base. It was hypothesized that attempting to build classifiers discriminating cleaved from non-cleaved instances based on sequence information alone would be particularly difficult, and that providing physical feature information might provide especially insightful distinctions for the matrix-capsid cleavage site.

The results for various features are shown in Table 3.6. Here, only a SVM was used since it was previously shown to perform well with various feature sets. Using only sequence information, the SVM built a classifier that was 66.7% accurate. Using physical information instead, the accuracy increased slightly. Both types of features in concert appeared to perform significantly better than sequence information alone. However, given that the data set consists of 75% cleaved instances, even a naive classifier that never predicts uncleaved instances would do roughly as well as a SVM with the full feature set. Additionally, the number of support vectors differs widely depending on the feature set used. Sets including sequence information require nearly all of the training instances be used as support vectors. When only physical features are used, the model contains fewer support vectors. This behavior indicates that the decision boundary is especially difficult to find in terms of sequence features.

3.3.3 Discussion

The classification performance of the most simple, linear classifiers described by Rognvaldsson and You[73] continues to provide a competitive standard of performance for much more sophisticated learning technologies, at least over sequence information alone. This study provides the first results extending to classification over richer sets of physical features of the cleaved and uncleaved instances.

A SVM exploiting these additional physical features did not perform substantially bet-

[§] More elaborate notions of sequence similarity, involving edit distance, PAM-style amino acid similarities, etc. might also be possible, but even this simple measure successfully separates training instances near the natural subsite from others and so other variants were not explored.

Table 3.5: Training instances similar to the matrix-capsid cleavage site. The matrix-capsid cleavage site is shown in bold.

Cleaved	Uncleaved
Sequence	Sequence
SQNYPIVQ	RQNYPIVQ
LQNYPIVQ	SQKYPIVQ
MQNYPIVQ	SQNPPIVQ
SKNYPIVQ	SQNSPIVQ
SNNYPIVQ	SQNYAIVQ
SQAYPIVQ	SQNYDIVQ
SQCYPIVQ	SQNYKIVQ
SQIYPIVQ	SQNYPKVQ
SQLYPIVQ	SQQYPIVQ
SQNFPIVQ	
SQNMPIVQ	
SQNYLIVQ	
SQNYPAVQ	
SQNYPIEQ	
SQNYPIFQ	
SQNYPIIQ	
SQNYPILQ	
SQNYPIVE	
SQNYPIVL	
SQNYPIVP	
SQNYPLVQ	
SQNYPNVQ	
SQNYPVVQ	
SQNYTIVQ	
SQTYPIVQ	
SQVYPIVQ	
TQNYPIVQ	

Table 3.6: Protease cleavage prediction for sequences similar to the matrix-capsid cleavage site.

Feature Set	Mean Accuracy	Standard Deviation
sequence	66.7%	2.62%
physical	71.4%	1.34%
sequence+physical	76.1%	2.68%

ter, probably due to several factors. Features based on site-specific physical properties, such as volume, are possibly redundant when sequence information is given. These features correspond exactly to amino acid residues, indicating that an SVM may already be able to generate an internal representation where the relationships between different residues are exploited, so additional physical information is unnecessary. The whole-peptide features used are likely too coarse to be of any additional benefit. While the sequence representation is extremely sparse, it appears to provide sufficient information. Representations based on physical features are more dense, but by compacting and transforming the representation, some information is lost.

However, the trees based on physical features are more readily interpretable in terms that can be used more directly for drug design. The groupings of amino acids found in the sequence-based decision tree lack any obvious commonality. Given physical features instead, the trees become much more comprehensible from a human point of view, even though trees from either perspective heavily favor classification based on the same amino acid positions. When focused on a single cleavage site where differences in cleaved and uncleaved instances are small in terms of sequence, the physical features provide increased classification ability.

In summary, the results suggest that the relatively small set of training instances currently available has been exhausted, certainly with respect to sequence-only features. Also, physical features of individual residues can be useful in providing more refined, biologically relevant characterizations of protease specificity. More experimental data, investigating other cleaved and uncleaved sequences, particularly sampling “near” other cleavage sites (analogous to the set near SQNYPIVQ), will be necessary before we can discover how protease is able to efficiently cleave across all these sites while retaining such high specificity.

Acknowledgements

Chapter 3.2 contains material that will be submitted to the journal *Retrovirology* as *Combining molecular docking and sequence analysis to predict resistance mutations for novel inhibitors of HIV protease*. Max W. Chang, Michael J. Giffin, Ying-Chuan Lin, John H. Elder, Arthur J. Olson, Bruce E. Torbett, and Richard K. Belew. I was the primary investigator and author of this work.

Chapter 4

Structural Analysis of Drug Resistance

Sequencing of resistant HIV strains, combined with genotypic and phenotypic analyses of related samples, have produced large volumes of data that have proven useful in monitoring disease progression in clinical settings.[79] For all FDA-approved reverse transcriptase and protease inhibitors, these analyses have yielded detailed lists of drug resistance mutations and their effects.[17,80] Related machine learning experiments involving the prediction of resistant HIV sequences have also proven useful.[19,20,81] These post hoc approaches have shortcomings, though, in that the results are based on historical data – drawn from extended clinical experience with approved inhibitors – and cannot be used when sufficient data is unavailable, as with newly approved or experimental drugs. A deeper understanding of the basis for drug resistance may allow for predictions of drug resistance mutations before clinical data is available.

For example, the recently synthesized protease inhibitor AB2[82] is a potent compound which can inhibit HIV replication in the low nanomolar range *in vitro*. However, it is far from approval for use in a clinical setting, nor has it been subjected to *in vitro* selection experiments, as has been done with other experimental inhibitors, such as TL-3.[16] For a drug candidate to progress to clinical trials takes years, leaving *in vitro* selection experiments via serial passage as the more tractable method for determining resistance pathways for a new inhibitor. Unfortunately, these experiments are demanding in terms of resources and time, and can take months to years to complete. A quick *in silico* method to "preview" drug resistance mutations would be useful to complement these more time-consuming techniques.

In examining drug resistance, it is important to note differences in the effects of drug resistance mutations. For HIV protease, mutations are generally divided into primary and compensatory mutations.[80] *Primary* mutations arise with the onset of antiviral therapy in a drug-naive patient, and lead to reduced inhibitor binding and frequently impaired protease function as well. *Compensatory* mutations typically arise after primary mutations and restore some protease function. They also generally have a lesser effect on inhibitor binding or are in some cases dependent on the presence of a primary mutation. Mutations are also classified as major or minor, depending on the degree of resistance conferred.

These observations regarding differences in drug resistance mutations provide a glimpse into more general notions of viral fitness with respect to protease. To replicate effectively, the viral protease must cleave specific targets with some minimal level of efficiency.[63] The presence of competitive inhibitors reduces enzymatic activity below the threshold required for viral propagation. Even when not fully inhibited, protease activity may become a bottleneck for viral replication capacity. However, some mutants are less susceptible to particular inhibitors, and will remain viable, despite lower catalytic efficiency (and overall replication capacity) than the more vulnerable wild-type HIV.[83] Therefore, in understanding the basis of drug resistance and viral fitness, it is important to examine the interplay between replication capacity and resistance to inhibition.

Previous research performed by Wang and Kollman examined the molecular basis of drug resistance in HIV protease by studying both sequence variability and the results of molecular dynamics simulations with substrates and inhibitors.[59] This work involved studying the sequence variability among protease orthologs, with low sequence variability implying that a position was important for catalytic ability or structural stability. Protein-ligand interactions captured by the molecular dynamics simulations revealed differences in the binding modes of inhibitors relative to a substrate molecule. Combining the sequence information with structural insights led to success in predicting some major drug resistance mutation sites. However, while protease *positions* were implicated as likely mutation sites, particular amino acid substitutions were not identified.

In this chapter, Section 4.1 first focuses on predicting drug resistance mutations for clinically-approved drugs. It demonstrates that highly accurate results can be obtained with predictions based directly on clinically-observed mutations found in inhibitors of the same class. However, this approach lacks generality, as it relies directly on the use of existing mutation patterns. Section 4.2 presents a more broadly applicable strategy extending the framework established by Wang and Kollman, using sequence homology and protease-ligand interactions. When tested on protease inhibitors in clinical use, the predictions were less accurate than the predictions made on the basis of observed mutations, but over half of the major protease re-

Table 4.1: Drug resistance predictions based on cross-resistance. An asterisk (*) indicates a 0.01 significance level. Each additional asterisk represents an additional 10^{-1} decrease in significance level, to a minimum level of 10^{-5}.

inhibitor	accuracy	precision	recall	p-value
APV	0.98	34/77	34/34	****
ATV	0.98	47/64	47/60	****
DRV	0.98	29/75	29/31	****
IDV	0.99	48/77	48/48	****
LPV	0.98	42/75	42/44	****
NFV	0.98	45/75	45/47	****
SQV	0.98	43/77	43/43	****
TPV	0.98	29/70	29/36	****
average	0.98	0.54	0.93	–

sistance mutations were identified. Finally, in Section 4.3 a combination of clinically-observed mutations and predicted binding interactions was used to predict resistance mutations against the novel inhibitor AB2. Biochemical testing of the protease mutants, involving the 47V, 53L, and 84V mutations, confirmed increased resistance to AB2, but also revealed unanticipated nonlinear effects.

4.1 Cross-resistance in HIV protease

When examining existing patterns of drug resistance, it is evident that there are many mutations which confer resistance against multiple inhibitors (see Figures 2.7 and 2.8). These commonalities among resistance patterns can be exploited to generate drug resistance predictions. With a leave-one-out cross-validation scheme, the drug resistance mutations of one drug were predicted using the union of drug resistance mutations for all other protease inhibitors. For example, predictions for amprenavir resistance were made by examining the resistance profiles for all other drugs, and assuming that any mutation mentioned would confer amprenavir resistance. As shown in Table 4.1, this method had high accuracy and recall, and was able to capture the majority of resistance mutations. The high level of recall likely relates to the shared mode of action of all approved protease inhibitors.[5] On the other hand, this approach would not be applicable for a novel target.

Drug resistance mutation data Comprehensive information on drug resistance mutations is available from two major sources: the Stanford HIV Database (HIVDB) and the International AIDS Society–USA (IAS-USA).[79,80] To reconcile these two sources, HIVDB score thresholds were matched to IAS-USA classifications. Any mutation with HIVDB score >= 20 or mutation designated as major by IAS-USA was identified as a “major” drug resistance mutation for our experiments. Remaining mutations with an HIVDB score >= 5 or listed by IAS-USA were used to form a broader set of all drug resistance mutations we will refer to as “minor” mutations. For this work, the IAS-USA data from Fall 2005[84] was used when possible, as ritonavir-boosting of protease inhibitors has become standard more recently, which could have some effect on mutation profiles.

In evaluating the accuracy of resistance predictions, accuracy alone may be misleading, as resistance mutations make up only a small fraction of possible mutations. There are, for example, 34 resistance mutations for amprenavir, out of a possible 1,881 mutations (19 substitutions at each of 99 positions). A trivial predictor could achieve average classification accuracy of 99.5% on major mutations and 97.7% on all mutations by simply assuming that no substitutions confer drug resistance. Therefore, examining the precision and recall of the predictors provides additional information to determine the utility of the results. Precision was determined as the proportion of predicted drug resistance mutations that were correct (i.e., part of either the major or minor mutation sets defined above). Recall (or sensitivity) was calculated as the number of drug resistance mutations correctly predicted, compared against the total number of known drug resistance mutations.

The level of statistical significance was assessed by comparing results to a hypergeometric distribution, $p(Y = k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$ where N is the total number of possible mutations, r the number of drug resistance mutations, n the number of predictions made, and Y the number of correct predictions. The significance values reported correspond to the probability of correctly identifying at least Y resistance mutations.

4.2 Structure-based prediction of resistance

A combination of structural and sequence factors was used to predict resistance mutations in HIV protease, excluding any prior knowledge directly related to resistance. Clinical data was used only for validation, in contrast to the previous section. All possible mutations were subjected to filtering based on changes in protease-inhibitor and protease-substrate binding energies, in addition to a sequence-based threshold via MAPP.

Table 4.2: Protease-substrate complex structures.

PDB ID	Substrate peptide sequence	Cleavage site
1F7A	`KARVL/AEAMS`	CA/P2
1KJ4	`VSQNY/PIVQN`	MA/CA
1KJ7	`PATIM/MQRGN`	p2/NC
1KJF	`RPGNF/LQSRP`	p1/p6
1KJG	`GAETF/YVDGA`	RT/RTp66
1KJH	`IRKIL/FLDGI`	RTp66/INT
2FNS	`RQANF/LGKIN`	NC/p1

Table 4.3: Protease-inhibitor complex structures.

PDB ID	Inhibitor
1HPV	amprenavir (APV)
2O4K	atazanavir (ATV)
1T3R	darunavir (DRV)
1SDT	indinavir (IDV)
2O4S	lopinavir (LPV)
1OHR	nelfinavir (NFV)
2NMW	saquinavir (SQV)
2O4P	tipranavir (TPV)

4.2.1 Protease-ligand binding energy prediction

Protease crystal structures in complex with several polypeptide substrate segments are available at the Protein Data Bank (PDB),[85] as shown in the Table 4.2. Structures for several protease inhibitors (including all FDA-approved drugs) were also obtained from protease-inhibitor complexes in the PDB (Table 4.3). In all cases, proteases with the fewest mutations relative to the "wild-type" HXB2 sequence and with high resolution (almost always < 2 Å) were selected. Protein-ligand binding energies were estimated using a Python implementation of the AutoDock4 forcefield.[40] This program is able to report per-residue contributions to binding energy. Energies from each of the protease chains were merged, such that a position's energy represented the sum of the residues on both chains.

Implications for drug design

Evaluating the intermolecular binding between each protease-substrate pair revealed consistent interactions across nearly all of the positions in protease (Figure 4.1a). A majority of the total binding energy was contributed by roughly a dozen positions in close proximity to the substrate. Similar consistency in per-position binding energies was observed across protease inhibitors (Figure 4.1b). Overall, the inhibitor binding energies appeared weaker than the substrates near the catalytic triad, especially at position 27. Stronger binding at positions 50, 82,

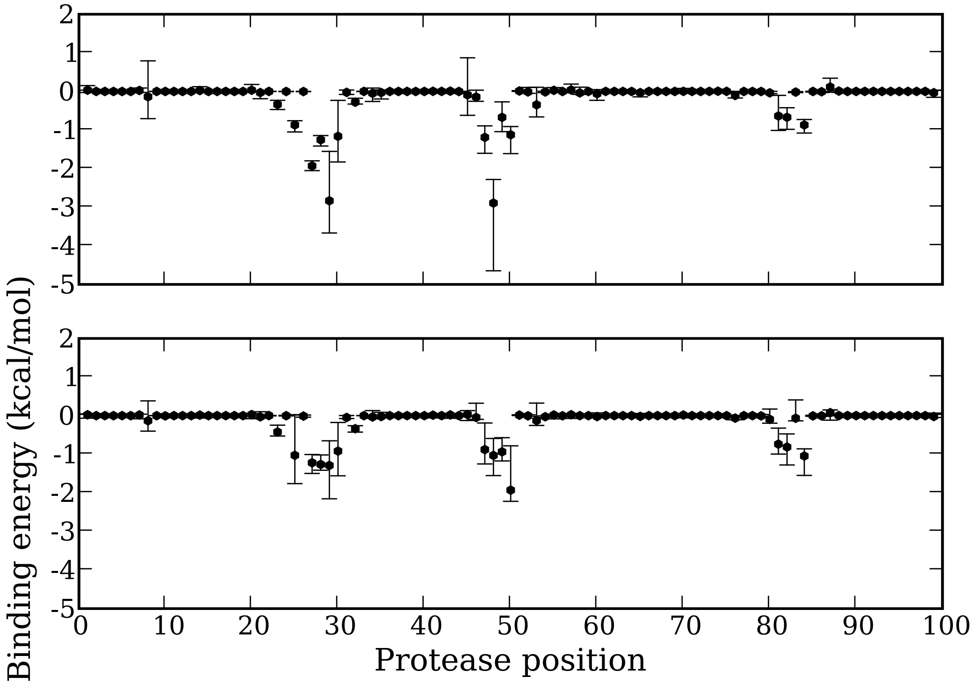


Figure 4.1: Per-position binding energies for 7 protease-substrate complexes (top) and 10 protease-inhibitor complexes (bottom). The mean energy is shown with a point. Error bars indicate the minimum and maximum across all complexes.

and 84 corresponded to the location of some of the most prevalent drug resistance mutations.

4.2.2 Predicting mutations with position-specific binding energy

These binding energies can be applied to predict positions where drug resistance is likely to occur, based on differences between substrate and inhibitor binding interactions. Here, $\Delta\Delta G$ is defined as the difference between the mean substrate and inhibitor binding energy at a specific position. Resistance predictions are made where $\Delta\Delta G$ is less than -0.05 kcal/mol and any non-wild-type residue has a MAPP score less than 12, which was chosen based on the analysis in Chapter 3. Positions not meeting these criteria were assumed not to harbor drug resistance mutations. Known resistance positions are specified from the IAS-USA and HIVDB data detailed above, and included positions where any major or minor mutation was present. Prediction accuracy, with precision and recall, is shown in Table 4.4. Predictions were best for amprenavir and nelfinavir, where key mutations lie near the active site and protease-inhibitor interaction is strong. For saquinavir, on the other hand, resistance mutations are prevalent at sites more distant from the active site, and therefore, more difficult to predict through evaluation

Table 4.4: Position-only mutation predictions based on sequence variation and position-specific binding energy.

	major resistance positions				all resistance positions			
inhibitor	accuracy	precision	recall	p-value	accuracy	precision	recall	p-value
APV	0.96	3/5	3/5	**	0.88	3/5	3/13	-
ATV	0.93	3/8	3/5	*	0.78	4/8	4/22	-
DRV	0.94	2/6	2/4	-	0.87	3/6	3/13	-
IDV	0.86	2/14	2/4	-	0.75	3/14	3/17	-
LPV	0.93	3/9	3/4	*	0.81	4/9	4/18	-
NFV	0.91	3/8	3/7	-	0.86	6/8	6/18	**
SQV	0.92	1/7	1/3	-	0.84	2/7	2/13	-
TPV	0.94	2/7	2/3	-	0.79	2/7	2/18	-
average	0.92	0.32	0.55	–	0.82	0.45	0.21	–

Table 4.5: Substitution predictions based on sequence variation and position-specific binding energy.

	major mutations				all mutations			
inhibitor	accuracy	precision	recall	p-value	accuracy	precision	recall	p-value
APV	0.99	5/22	5/8	****	0.98	5/22	5/34	***
ATV	0.98	7/46	7/9	****	0.96	12/46	12/60	****
DRV	0.97	2/53	2/5	*	0.96	5/53	5/31	*
IDV	0.92	6/147	6/9	***	0.91	11/147	11/48	**
LPV	0.97	6/56	6/8	****	0.96	12/56	12/44	****
NFV	0.97	8/56	8/18	****	0.96	13/56	13/47	****
SQV	0.97	3/48	3/9	*	0.96	10/48	10/43	****
TPV	0.98	5/46	5/6	****	0.97	10/46	10/36	****
average	0.97	0.11	0.60	–	0.96	0.19	0.22	–

of position-specific binding energy. Indinavir predictions have the opposite problem, with too many contacts causing spurious predictions of drug resistance.

Turning to the prediction of specific substitutions, similar criteria were used to make predictions. Any mutation with MAPP score less than 12 and located at a position with $\Delta\Delta G$ less than -0.05 kcal/mol was predicted to be a resistance mutation. Due to the large number of mutations considered, accuracy increased greatly (Table 4.5). However, precision decreased dramatically, with a huge increase in the number of falsely predicted mutations. Given the coarse-grained structural information, this result is unsurprising, the $\Delta\Delta G$ provides no information that could be useful in distinguishing between possible mutations at a single locus. Despite poor precision, recall for the major mutations averaged approximately 60%, indicating that MAPP and $\Delta\Delta G$ were sufficient to capture the majority of the most severe drug resistance mutations. Further, when considering only major mutations, a number of the false positives were actually minor drug resistance mutations. In general, the statistical significance of these results was quite high, indicating success far beyond random choices.

Table 4.6: Substitution prediction for known drug resistance mutations from IAS-USA and HIVDB.

	major mutations				all mutations			
inhibitor	accuracy	precision	recall	p-value	accuracy	precision	recall	p-value
APV	0.99	5/22	5/8	****	0.98	7/22	7/34	****
ATV	0.98	6/39	6/9	****	0.96	11/39	11/60	****
DRV	0.99	2/24	2/5	*	0.98	7/24	7/31	****
IDV	0.97	5/55	5/9	****	0.96	12/55	12/48	****
LPV	0.99	6/31	6/8	****	0.97	10/31	10/44	****
NFV	0.98	7/43	7/18	****	0.96	11/43	11/47	****
SQV	0.98	3/27	3/9	**	0.97	8/27	8/43	****
TPV	0.98	3/40	3/6	**	0.97	6/40	6/36	***
average	0.98	0.14	0.53	–	0.97	0.27	0.21	–

4.2.3 Estimating mutation-induced changes in protease-ligand affinity

A more detailed structural model was considered, using a set of protease structures with modified side chains. The protease structures included the complexes described above, with all possible single-mutants at 29 positions chosen due to close contact with at least one substrate or inhibitor in a crystal structure complex. No conformational search or energy minimization was performed. Mutant protease structures were generated using SCWRL3,[86] with ligands specified to avoid steric clashes.

For every ligand, the effect of each single mutation was determined relative to the original structure, yielding a $\Delta\Delta G$ measure that represents the change in protein-ligand binding between a wild-type protease and a particular mutation. Predicted drug resistance mutations satisfied two criteria related to predicted binding energies: (1) the largest change in substrate binding was less than +3.0 kcal/mol and (2) the change in inhibitor binding exceeded +0.25 kcal/mol. An increase of 3.0 kcal/mol represents a roughly 100-fold decrease in binding affinity, which is assumed to negatively affect protease function. Since the observed changes in predicted inhibitor binding energy were less dramatic, rarely exceeding +1.0 kcal/mol, a lower threshold of +0.25 kcal/mol was used.

The results shown in Table 4.6 indicate accuracy comparable to the leave-one-out scheme, but with lower precision and recall. On average, the recall rate for major mutations still exceeded 50%, indicating that this method was able to account for over half of the major mutations in HIV protease. The results also remained statistically significant, with most p-values far below 0.01. However, when considering all mutations, the average recall was approximately 25%, revealing a high false positive rate.

Despite lower performance than the leave-one-out approach, the combination of pre-

Table 4.7: All possible mutations at position 84 in relation to amprenavir resistance. Red cells indicate rules violations (MAPP score > 12, max. substrate $\Delta\Delta G$ > 3.0, Amprenavir $\Delta\Delta G$ < 0.25). Green rows indicate mutations that satisfy all rules.

Mutation	MAPP score	Max. substrate $\Delta\Delta G$	Amprenavir $\Delta\Delta G$
84A	11.47	1.77	0.93
84C	8.97	1.48	0.62
84D	37.2	1.07	0.64
84E	32.09	3.39	0.63
84F	7.34	374.21	62.55
84G	15.96	1.89	1.05
84H	17.52	45.32	17.97
84K	27.75	86.2	16.19
84L	4.5	0.76	-0.15
84M	11.03	1.3	-0.44
84N	16.5	1.43	0.57
84P	21.61	1.62	0.77
84Q	14.81	4.28	0.46
84R	33.3	5545.27	4590.03
84S	14.69	1.83	1.02
84T	15.48	1.36	0.8
84V	5.25	1.1	0.51
84W	15.68	7701.87	1302.7
84Y	13.85	1856.03	272.24

dicted binding energies and MAPP scores offers a deeper look into the basis of drug resistance. For example, this approach worked well for determining amprenavir resistance at position 84, which is located near the active site. Clinically, the wild-type isoleucine at position 84 is found to mutate to valine, alanine, and cysteine, with the 84V mutation being the most prevalent by far.[17] Looking at all possible mutations at position 84, only the alanine, cysteine, phenylalanine, methionine, and valine residues had MAPP scores less than 12 (see Table 4.7). The phenylalanine residue was predicted to have a major decrease in substrate binding, so it was rejected. Leucine and methionine residues did not not decrease inhibitor binding, leaving residues alanine, cysteine, and valine as candidates. Interestingly, the valine mutation, which is most commonly found, had the lowest MAPP score, indicating that a relatively high level of catalytic efficiency may explain its prevalence over the alanine and cysteine mutations.

4.3 Predicting resistance to a novel protease inhibitor

For a novel protease inhibitor, both existing resistance profiles and structure-based predictions have potential utility in anticipating drug resistance mutations. Assuming the inhibitor binds in the protease active site, existing resistance patterns should remain useful. These previ-

Table 4.8: Predicted AB2 drug resistance mutations.

Mutation	IAS-USA frequency	Predicted AB2 $\Delta\Delta G$ (kcal/mol)	Predicted APV $\Delta\Delta G$
47V	4/8 [4/9]	0.34	0.17
53L	2/8 [1/9]	0.21	0.00
84V	8/8 [9/9]	0.53	0.52

ously observed mutations indicate viable changes that should be tolerated by the enzyme, and can be used in place of the MAPP scores and predicted protease-substrate binding energy. The protease-inhibitor interaction remains relevant, though, because the clinical and sequence data does not provide any information specific to a particular inhibitor.

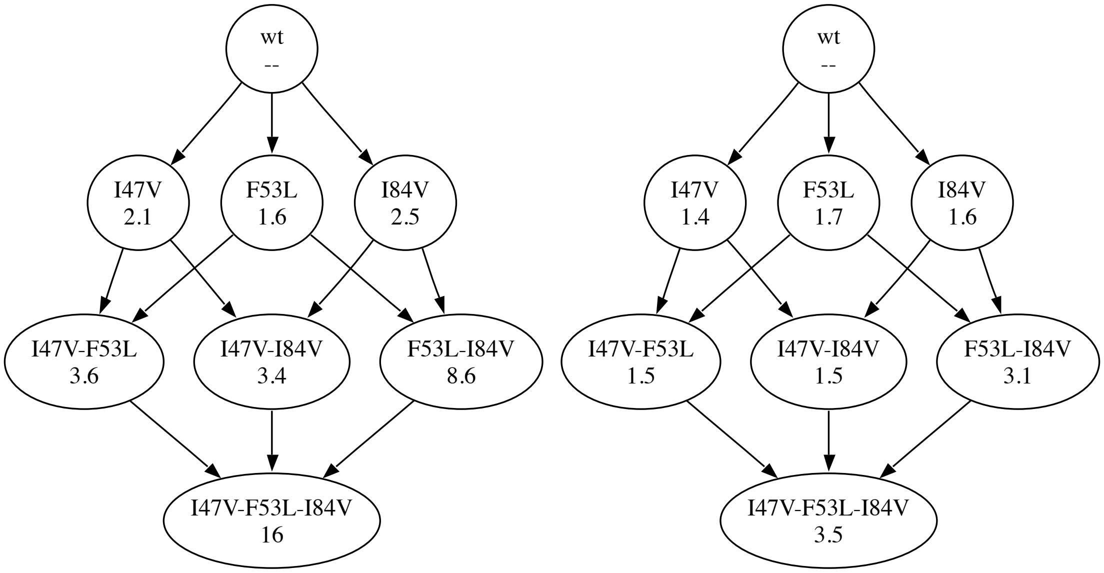
To validate the use of structure-based predictions combined with existing resistance profiles, mutant proteases were synthesized and tested biochemically against the novel protease inhibitor AB2, which targets the protease active site.[82] Three point mutations were selected, and to avoid having the synthesized mutants be dominated by the most common previously observed mutations, mutations were selected based on having different frequencies in the IAS-USA data set, as well as reduced binding affinity to AB2 (Table 4.8). The 84V mutation is common to all clinically-approved inhibitors, while 47V is noted in roughly half, and 53L only for the lopinavir / ritonavir combination. This range of frequencies was chosen to avoid selecting mutations where major resistance would be observed for any inhibitor. Additionally, these mutations were all predicted to cause an unfavorable change in binding with AB2. As a control, the effects of these same mutations were tested with amprenavir, where the overall effect was predicted to be much smaller.

Following the synthesis of the protease mutants, the degree of drug resistance for each mutation and every combination of mutations was determined through IC_{50} measurements. As shown in Table 4.9 and Figure 4.2, there was an increase in AB2 IC_{50} for each of the single mutants, going up to a 2.5-fold increase for the I84V mutation over wild-type protease. This trend continued for the double mutants, where all proteases demonstrated greater than 3-fold increases in IC_{50}. The F53L-I84V mutant showed an especially large increase of roughly 8.5-fold beyond wild-type protease. Finally, the triple mutant had an IC_{50} increase of over 15-fold relatively to wild-type. Overall, these changes showed a clear trend of large IC_{50} increases as the number of mutations grew.

In contrast, the predictions for amprenavir indicated that there should be more modest IC_{50} increases. For the single mutants, the amprenavir results resembled those for AB2, each of the mutants showed an IC_{50} increase over wild-type, but a smaller increase on average than shown for AB2 (Table 4.9 and Figure 4.2). Interestingly, the double mutants 47V-53L and 47V-

Table 4.9: IC_{50} and fold changes in protease inhibition for AB2 and amprenavir.

Mutation	AB2 IC_{50} (nM)	IC_{50} fold change	APV IC_{50}	IC_{50} fold change
wild-type	14.1	-	6.90	-
47V	30.2	2.1	9.37	1.4
53L	22.7	1.6	11.6	1.7
84V	35.0	2.5	10.9	1.6
47V-53L	50.2	3.6	10.4	1.5
47V-84V	48.4	3.4	10.4	1.5
53L-84V	121	8.6	21.5	3.1
47V-53L-84V	226	16	24.0	3.5



(a) AB2

(b) Amprenavir

Figure 4.2: Landscapes for mutations conferring resistance in AB2 and amprenavir. Labels indicate the fold-change in IC_{50} relative to wild-type HIV protease.

84V showed no increase in IC_{50} when compared to single mutants, in contrast to what was seen with AB2. The highest fold increase seen with amprenavir was on par with some of the AB2 double-mutant results, and far short of a 15-fold increase.

Effects of the mutations on protease extended beyond IC_{50}, however. All of these changes had some impact on substrate processing, as shown in Table 4.10. With the accumulation of mutations that impede inhibitor binding, substrate affinity decreased as well. Even the single mutations 53L and 84V were sufficient to boost K_m more than 5-fold. Two of the double mutants had K_m that increased even further, while the 53L-84V and 47V-53L-84V mutants had K_m far beyond the detectable range. According to changes in IC_{50}, these same mutants had

Table 4.10: Effect of AB2 resistance mutations on K_m. * indicates a K_m estimate that exceeds the maximum substrate concentration used. ** denotes K_m estimates that do not converge, i.e. far outside the substrate range.

Mutation	K_m (μM)
wild-type	34.4
47V	31.5
53L	207
84V	219
47V-53L	307*
47V-84V	424*
53L-84V	**
47V-53L-84V	**

major effects on the binding of AB2, though not amprenavir, indicating that the mutations that caused changes in substrate affinity affect some inhibitors, but not all of them. The large, non-linear increases in K_m also indicated that there were significant epistatic effects not evident when examining isolated mutations.

4.3.1 Protease expression and activity assays

Site-directed mutagenesis was carried out using the QuickChange protocol. After plasmid purification, the mutations were verified by DNA sequencing. Protein purification was carried out as described by Heaslet et al.,[87] following the wild-type protocol. The concentration of each protease was determined through active site titration with amprenavir.

HIV protease activity was measured with a fluorogenic hexapeptide substrate (Abz-Thr-Ile-Nle-p-nitro-Phe-Gln-Arg-NH2) using an FLX-800 Microplate Fluorescence Reader (Bio-Tek Instruments, Inc., Winooski, VT). Changes in fluorescence were measured over 15 minutes at 37°C with 340/30 nm excitation and 420/50 nm emission filters. Initial reaction rates were determined by linear regression of the initial 2 minutes of the reaction using KC4 (Bio-Tek). IC_{50} values were determined by non-linear regression using the initial reaction rates versus inhibitor concentration with Prism 4.0c for Macintosh (Graphpad Software, San Diego, CA).

All reactions were run in 100 μl total volume in 96-well microtiter plates, with buffer containing 50 mM MES (pH = 5.5), 200 mM NaCl, 1 mM DTT, 0.0002% Triton X-100, and 5% glycerol. For IC_{50} determination, protease concentration was 25 nM with initial substrate concentration at 30 μM. To determine K_m, conditions remained the same, except that substrate concentrations varied from 0 to 200 μM.

Acknowledgements

This chapter contains material that will be sumitted to the journal *Retrovirology* as *Combining molecular docking and sequence analysis to predict resistance mutations for novel inhibitors of HIV protease*. Max W. Chang, Michael J. Giffin, Ying-Chuan Lin, John H. Elder, Arthur J. Olson, Bruce E. Torbett, and Richard K. Belew. I was the primary investigator and author of this work.

Chapter 5

Viral Fitness and Evolution

The previous chapters have shown that viral replication capacity and drug resistance can be predicted with various computational methods. These quantities represent major factors in determining viral fitness under drug selection. However, as the work in Chapter 4 has shown, the effects of mutations can combine nonlinearly, and in modeling viral fitness, relationships between mutations should be taken into account. These relationships can be studied by examining patterns of covariation among mutations arising during drug therapy. Linking these covarying mutations can elucidate drug resistance pathways, which is demonstrated below by reconstructing the TAM pathways for NRTI resistance.

These covariation patterns are also of use in more directly calculating viral fitness, where estimates based on the RC and RF contributions of individual mutations are modified based on the degree of covariance between mutations. In this way, a model of viral fitness is able to account for nonlinear effects, and is shown to make accurate predictions of relative fitness among mutants selected under drug therapy. Adapting this notion of fitness to a simulation of drug resistance evolution provides the means to test novel treatment strategies and inhibitors. This simulation was applied to test the effects of combination therapy based on multiple protease inhibitors, focusing on a comparison of clinically-approved inhibitors against a putative allosteric inhibitor. Combinations of a clinically-approved inhibitor and allosteric inhibitor were found to be highly effective, even under assumptions that the allosteric inhibitor is weak and highly vulnerable to drug resistance.

5.1 Mapping resistance pathways

During drug therapy, the emergence of resistance mutations follows certain patterns, which have been widely studied in a clinical context.[17,80] The effects of some of these mutations are often conditional upon the presence of other mutations, such as the dependency of the 88D protease mutation on 30N during nelfinavir treatment.[88,89] A study performed by Rhee et al. reported a large number of significantly covarying pairs in HIV protease and reverse transcriptase from clinical records.[90] Notably, the study found that conditional probabilities between pairs had some ability to predict the order in which mutations developed.

Several other statistical analyses have focused on discovering covariation patterns in protease and reverse transcriptase sequences from clinical isolates. The mutagenic tree models described by Beerenwinkel et al. have been useful in visualizing mutation pathways, including the concept of a genetic barrier.[91] Similar pathway representations have been explored by multiple groups using Bayesian networks,[89,92,93] which are a sophisticated method for analyzing the topology of resistance pathways.

For the broader goal of simulating the evolution of drug resistance, techniques to recreate mutational pathways play an important role. The analyses of replication capacity and drug resistance in Chapters 3 and 4 focused on the effects of individual mutations, without taking into account their interactions. Estimates of viral fitness based solely on individual mutations ignore the higher order information that a covariation analysis can provide. Below, the utility of covariation analysis in determining mutational pathways is demonstrated by reconstructing the well-characterized TAM pathways associated with NRTI resistance. Graphical representations of the covariation patterns can also serve as a useful global view of drug resistance evolution in specific contexts.

5.1.1 Single-drug covariation analysis

The previous covariation study performed by Rhee et al. examined viral sequences from patients treated with any type of protease or RT inhibitor.[90] This allowed a large number of sequences to be examined, which exposed the relationships between many mutations at a high level of statistical significance. However, combining various treatments in a single analysis obfuscates the effects of individual inhibitors. For example, certain favored dependencies, such as the previously mentioned 30N and 88D combination, will have higher probabilities when considering only nelfinavir-treated patients rather than patients receiving any protease inhibitor.

To carry out a covariation analysis based on the effects of individual inhibitors, amino acid sequences from patients treated with single drugs were retrieved from the Stanford HIV Drug Resistance Database.[17] For each patient, only the latest isolate was included. The protocol described by Rhee et al. was used[62] for covariation analysis. Multiple hypothesis testing was not performed, and a z-score threshold of ± 2 applied for all covariation pairs. For analysis of reverse transcriptase results, only the first 240 amino acids in the protein were considered.

The results shown in the following sections focus on drugs for which a large number of records were available. For drugs with a small number of records, the number of covariation pairs was too small to reveal higher order structures beyond pairwise interactions. Because the significance of interactions found in the covariation analysis depends on the distribution of mutations specific to each set of records, a rough empirical threshold was used rather than a statistically-based cutoff. Focusing on the largest treatment sets for which at least 250 sequences were available, there were 676 indinavir-, 753 nelfinavir-, and 337 AZT-treated records.

5.1.2 Thymidine-analog mutation pathways

Resistance to AZT (aka zidovudine) often follows well-characterized paths, known as thymidine-analog mutation (TAM) pathways.[94] A previous analysis using mutagenic trees was able to reconstruct the TAM1 and TAM2 pathways using clinical data (Figure 2.9).[91] Our analysis of the AZT covariation analysis to TAM-associated mutations revealed a similar pattern, shown in Figure 5.1. The 215Y and 70R mutations appear to initiate their respective pathways. Also, the TAM1 and TAM2 pathways are shown as distinct, with several negative associations between mutations in either group. In contrast to the mutagenic tree model, the covariation pathways associate the 215F mutation with the TAM2 pathway, which is supported by a mutagenesis study.[95] The covariation pathways also exhibit a more branched structure, while the mutagenic trees depict a strictly ordered progression.

5.1.3 Nelfinavir and indinavir mutation pathways

In protease drug resistance, there are no canonical pathways that are equivalent to the reverse transcriptase TAMs. The general model is that primary mutations confer drug resistance, with negative effects on enzyme function, and are followed by compensatory mutations which restore function.[14, 15] In the indinavir and nelfinavir pathways shown in Figures 5.2 and 5.3, one can imagine initial resistance mutations favoring the diamond nodes, then spreading out along the edges. Also important are the negatively correlated interactions. In the evolution of resistance, certain combinations of mutations are likely to be ill-matched, such as the combination of

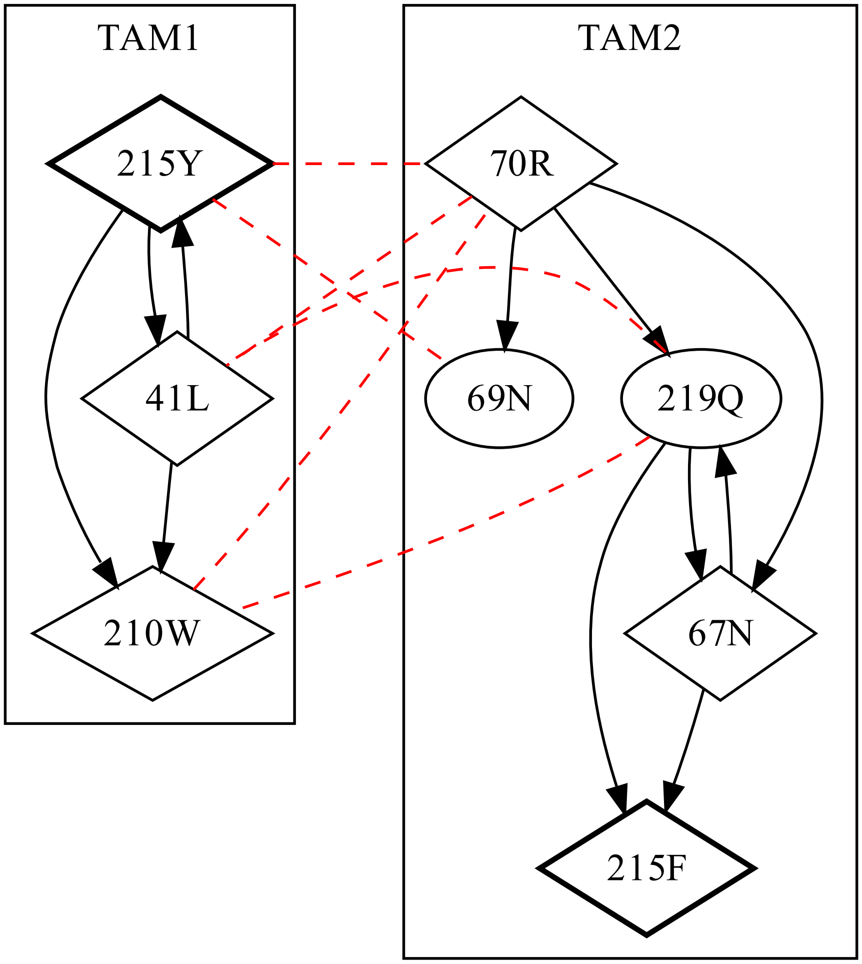


Figure 5.1: TAM pathways revealed by covariation analysis. Diamond mutation nodes indicate a significant drug resistance, bolded nodes have an even larger effect. Directed edges indicate conditional probabilities greater than 0.5. Dashed edges indicate negative covariance.

53L and 84V in protease.

5.2 Viral fitness model

In the development of resistance, one of the main mechanisms is reduced affinity between the viral targets and their inhibitors, caused by specific mutations. Generally, these mutations reduce inhibitor binding while not crippling the viral protein's function, such as HIV protease's cleavage of specific amino acid sequences. In attempting to model drug resistance evolution using a computer simulation, the interplay between the function of viral enzymes and their resistance to inhibitors must play a central role. Models for predicting RF and RC independently were described in Chapters 3 and 4. Combining these quantities to predict overall viral fitness is addressed in this section.

5.2.1 Fitness function

A kinetic model of protease function, which accounted for protease activity and drug resistance was presented by Tang and Hartsuck.[96] The processing activity a_{MI} was described with the function:

$$a_{MI} = \sigma \frac{k_{cat}/K_m}{1 + \frac{[I]}{K_i}} \tag{5.1}$$

where σ is a scaling value dependent on initial substrate and protease concentrations. K_m is the Michaelis-Menten constant and k_{cat} measures the number of reactions catalyzed by the enzyme. The ratio of k_{cat}/K_m is an indicator of overall enzyme efficiency. Impairments in protease function reduce the number of infectious particles produced,[63] i.e. RC. [I] and K_i represent inhibitor concentration and the inhibition constant, respectively. [I]/K_i determines the level of inhibition. For simplicity, and because mutation-induced changes in inhibitor binding are generally described by IC_{50} fold changes, the IC_{50} is substituted for K_i:

$$\frac{[I]}{K_i} \quad \rightarrow \quad \frac{[I]}{IC_{50}}$$

Further, the [I] and IC_{50} terms can be described relative to wild-type, rather than as absolute values. So [I] becomes the inhibitor concentration relative to the wild-type IC_{50}, while the IC_{50} term in the denominator becomes the change in IC_{50} relative to wild-type, or RF. Also, in systems where protease alone is being examined, viral fitness depends on processing activity, so

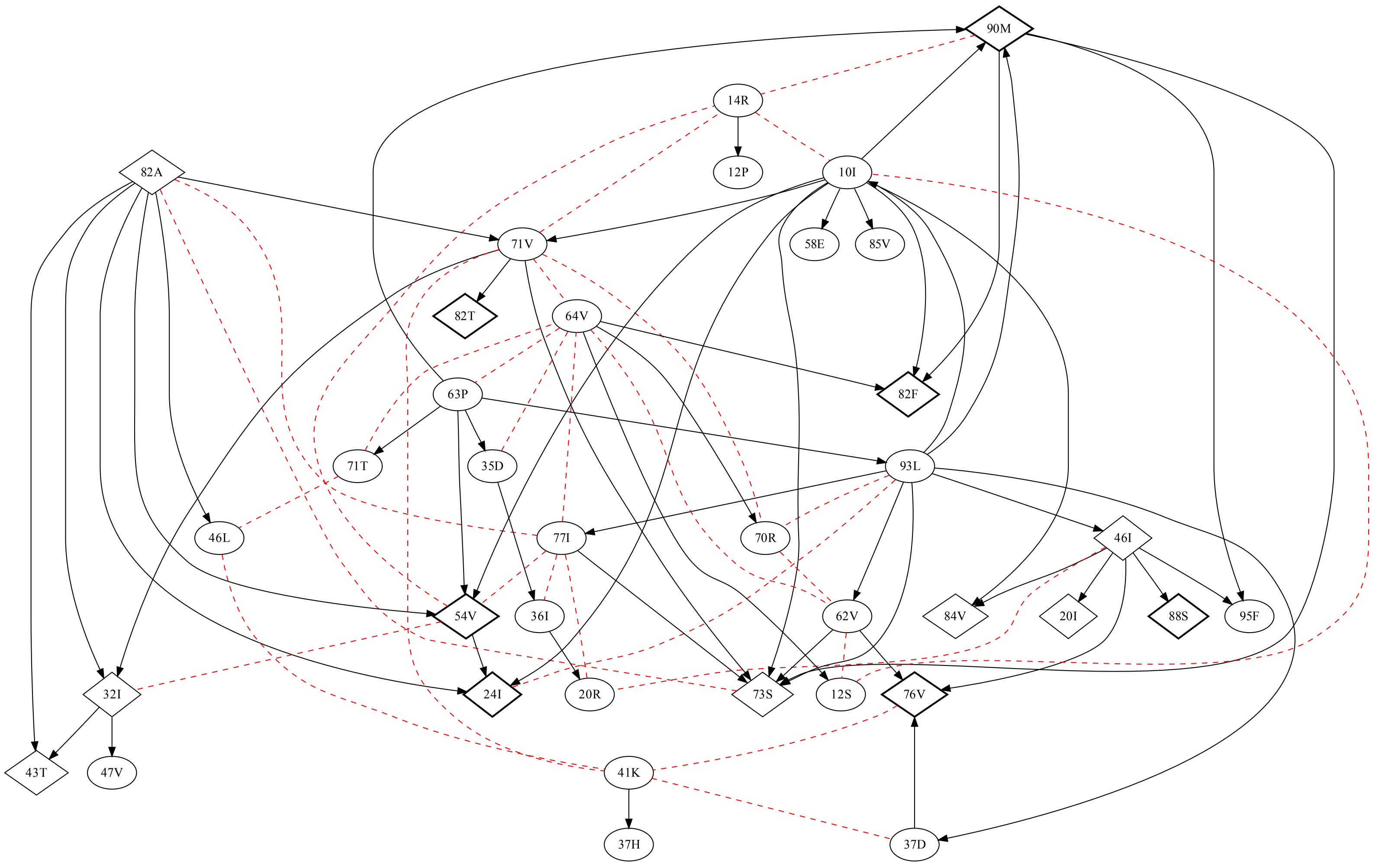


Figure 5.2: Indinavir resistance pathways from covariation analysis. See Figure 5.1 for description.

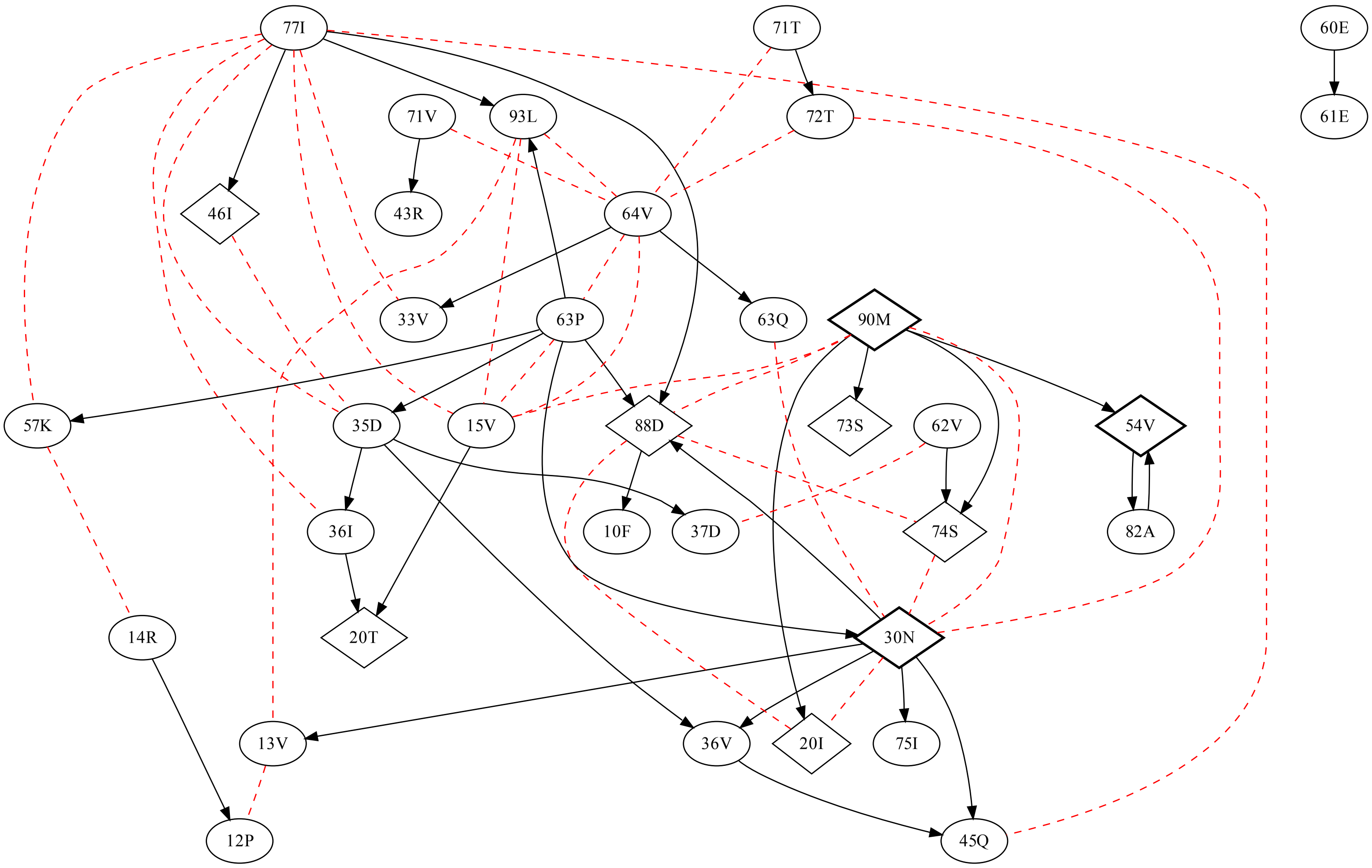


Figure 5.3: Nelfinavir resistance pathways from covariation analysis. See Figure 5.1 for description.

$a_{MI} = fitness_{protease}$. Adapting Equation 5.1 to use RF and RC, ignoring the σ scaling value, yields:

$$fitness_{protease} = \frac{RC}{1+\frac{[I]}{RF}}$$

In some clinically-relevant situations, multiple protease inhibitors may be of use.[97] As all current protease inhibitors are competitive inhibitors targeting a common site, their effects are mutually exclusive.[98] So for situations where multiple protease inhibitors are used, the function becomes:

$$fitness_{protease} = RC/\left(1+\sum_{i=1}^{n}\frac{[I_i]}{RF_i}\right)$$

The determination of RF and RC values were based upon analyses conducted in Chapter 3 and other works. The RF values were taken from a linear regression experiment performed by Rhee et al.[20] Based on a set of fold-change data, the authors applied regression to weight the contribution of individual mutations in resistance toward several protease and RT inhibitors. The RF value for a particular mutant is simply the sum of the individual RF contributions per mutation.

RC values were based on the MAPP results discussed in Chapter 3. The MAPP scores were correlated to the level of enzyme activity for several protease mutants, so the effect of a single mutation on protease was assumed to follow a sigmoidal relationship with the MAPP score:

$$RC = 1/\left(1+e^{a(MAPP_i-b)}\right)$$

for a single mutation i. A sigmoidal response was chosen to allow a range of low MAPP scores to have small effects on RC. For the constants $a = 0.25$ and $b = 0.8$, based on work in Chapter 3. Previously, a MAPP score of 12 was used to represent the threshold where the effects of mutations became significantly deleterious. As virus containing a protease with 25% normal activity was found to retain near-normal infectivity,[63] the constants were chosen so that a MAPP score of 12 would yield an RC value near 0.25. In the absence of epistasis, the effect of multiple mutations on RC is assumed to be multiplicative,[99, 100] such that:

$$RC = \prod_{i=1}^{n} 1/\left(1+e^{a(MAPP_i-b)}\right)$$

The material in Section 5.1 demonstrated that covariance measures could be used in visualizing the interaction between mutations in drug resistance. Incorporating these covariance measures into the fitness function is a way of accounting for the nonlinear, or epistatic, interactions that arise during the evolution of drug resistance.

For a simple model of epistasis, it is assumed that positively covarying mutation pairs will increase the fitness above what is expected from the RC and RF components, which are each based on contributions of individual mutations. Similarly, negatively covarying pairs will decrease fitness below this expected value. The covariation analysis yields Jaccard similarity coefficients and conditional probabilities that may be relevant in determining the strength of association between each pair. However, these values are not immediately applicable in calculating fitness. The Jaccard coefficient alone is biased by the number of occurrences of mutations. The conditional probabilities, on the other hand, represent a more complex and asymmetrical relationship. These are difficult to incorporate into the current model, as they indicate more complex dependencies between mutations that would require re-weighting on the individual RC and RF values. In the current study, the RC and RF values are modified proportionally to the number of positively and negatively covarying mutation pairs, represented by the variables c and d. These values have associated weights α and β, such that the presence of positively covarying pairs will increase fitness, and negatively covarying pairs will decrease it. The fitness with respect to protease becomes:

$$fitness_{protease} = \frac{RC + \alpha c + \beta d}{1 + \sum_{i=1}^{n} \frac{[I_i]}{RF_i}} \tag{5.2}$$

Applications of this fitness function on selected mutant proteases over a range of nelfinavir concentrations are shown in Figure 5.4. The graph shows that the fitness of wild-type protease is higher than the mutants at low drug concentrations, but declines quickly. 30N, a common nelfinavir resistance mutation, is shown to retain fitness near wild-type and to maintain a high level of function as the drug concentration increases. The combination 30N and 90M mutations is predicted to greatly impair viral fitness, which coincides with an experimental result.[101] 84A is rarely seen in clinical isolates,[17] but substantial resistance to nelfinavir and a decrease in replication capacity have been reported as a consequence of this mutation.[102] Both of these properties are also evident in the predicted fitness curve for 84A. These examples illustrate the basic usage of this type of fitness function, further validation is discussed in the following sections.

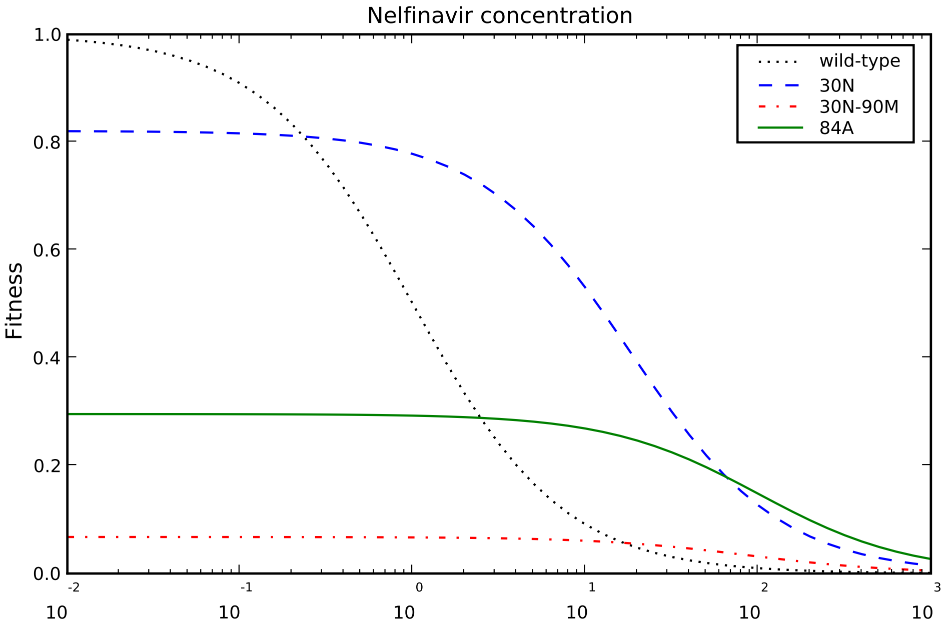


Figure 5.4: Predicted fitness for nelfinavir-resistant protease mutants. Nelfinavir concentrations are given relative to wild-type IC_{50}.

Although Equation 5.1 was developed for the activity of protease inhibitors, the generality of this form makes it appealing for other enzymes as well, such as reverse transcriptase. The MAPP scores, regression coefficients, and covariation measures are all available for reverse transcriptase, so the same approach can be used.

One key difference is in the calculation of RF when considering multiple reverse transcriptase inhibitors. In protease, inhibitors all target a single site and are assumed to be mutually exclusive. Since reverse transcriptase inhibitors fall into two major classes with different modes of action, this assumption does not hold. Complex pharmacokinetics are not modeled, so only one NRTI and one NNRTI are currently taken into account in the function. Also, for simplicity, the inhibitors are both assumed to be competitive in nature.

$$fitness_{RT} = \frac{RC + \alpha c + \beta d}{1 + \frac{[I_{NRTI}]}{RF_{NRTI}} + \frac{[I_{NNRTI}]}{RF_{NNRTI}} + \frac{[I_{NRTI}][I_{NNRTI}]}{RF_{NRTI}RF_{NNRTI}}} \tag{5.3}$$

When considering a system where both protease- and reverse transcriptase-based fitness are applicable, the lower fitness level is used, approximating a rate-limiting step, i.e.:

$$fitness_{total} = min(fitness_{protease}, fitness_{RT}) \tag{5.4}$$

Table 5.1: Rank correlation between prediction and actual fitness. One-tailed p-values shown. The full model (Equation 5.2) includes RF, RC, and epistatic components.

Drug	τ (RF-only model)	p-value	τ (RC+RF model)	p-value	τ (full model)	p-value
IDV	0.17	0.23	-0.16	-	0.29	0.11
NFV	0.29	0.11	0.35	0.070	0.71	0.0012
RTV	0.52	0.014	0.078	0.37	0.51	0.015
SQV	0.40	0.036	0.31	0.078	0.52	0.0091

5.2.2 Validation

Ranking resistant protease mutants

To validate the utility of the fitness function with respect to protease (Equation 5.2), predictions were compared against experimental results involving nelfinavir-, ritonavir-, and saquinavir-resistant mutants.[15,103] These studies reported the viral replication capacity of several mutants as a function of protease inhibitor concentration. Comparing the reported replication capacity values, which are measured at a high drug concentration, provides a ranking of the mutants. Calculating the fitnesses of each mutant based on Equation 5.2 yields another ranking, which is compared to the experimental results using the Kendall τ rank correlation. In all cases, the fitnesses are evaluated assuming that the drug concentration is 10× the wild-type IC_{50}.

The results in Table 5.1 show that predictions for nelfinavir, ritonavir, and saquinavir are significantly better than random. The indinavir predictions are the worst, and not significant at a 0.05 significance level. However, the indinavir results are a unique case because the mutants tested were selected for ritonavir resistance and showed low replication capacity when treated with indinavir,[15] thus presenting a more difficult ranking problem. Fitness calculations which incorporated RF, RC, and epistasis performed better overall than calculations based on RF alone. The best correlation was obtained for the nelfinavir mutants, which are shown in Table 5.2. The epistasis values may play an even larger role with data sets not dominated by single and double mutants.

Ranking resistant reverse transcriptase mutants

A smaller set of mutants was available from a set of growth competition experiments involving AZT resistance in reverse transcriptase.[95] The experiments were carried out in with AZT present and absent, reporting the relative fitness between five pairs of mutants. Applying the reverse transcriptase fitness model (Equation 5.3), predicted fitnesses were evaluated based on their correspondence to the relative ordering of mutants in each of the five pairs. In the absence of drug, the epistasis term was ignored, as it is meant to capture nonlinear effects in both RC and

Table 5.2: Ranking nelfinavir mutants. *In vitro* fitness rankings are based on replication capacity in the presence of nelfinavir.[103]

Mutant	*in vitro* fitness	predicted fitness (RF-only)	predicted fitness (complete)
30N, 71V	1	5.5	3
30N, 88D	2	4	1
30N, 71V, 90M	3	2	4
10I, 90M	4	7.5	5
30N, 63P, 90M	5	2	6
30N	6	5.5	2
90M	7	7.5	7
71V	8	10.5	10
wild-type	9	10.5	9
88D	10	9	8
30N, 90M	11	2	11

RF. When applicable, the concentration of AZT was assumed to be 10× the wild-type IC_{50}.

For the five pairs of mutants tested in the absence of AZT, the predicted fitnesses matched the reported ordering in all five cases. In the presence of AZT, four out of five predictions matched the previous study. The single failure involved the comparison between 215F and 215Y mutants. The 215Y mutant was able to out-compete the 215F mutant *in vitro*, regardless of the presence or absence of AZT. In the fitness predictions, however, the 215F mutant had greater AZT resistance than the 215Y mutation, resulting in a higher fitness value, though 215Y did retain a slight advantage in terms of RC (Figure 5.5). Accordingly, the error stems from the RF component of the fitness function, which is based on linear regression analysis.

This exposes a weakness of basing RF on regression coefficients, as the coefficients are set using an assumption of independence. As the 215F mutant rarely occurs alone, the isolates that were the basis of the linear regression analysis do not include even a single instance of 215F occurring without several associated mutations. Consequently, the regression coefficient that characterizes the 215F mutation's contribution to RF is also accounting for interactions with associated mutations, exaggerating its effects when found in isolation. To remedy this flaw would require the evaluation of resistance on additional mutants, or a more sophisticated analysis of the dependencies between mutations used in the regression data set.

Consistency with mutations arising during treatment

Clinical records from patients undergoing drug therapy provide another set of data for validation. These records consist of patient treatment histories (the inhibitors used) and viral sequences. In contrast to *in vitro* studies, fitness information is not directly available for these clinical records. Instead, the use of therapy is assumed to cause mutations to arise. For instance,

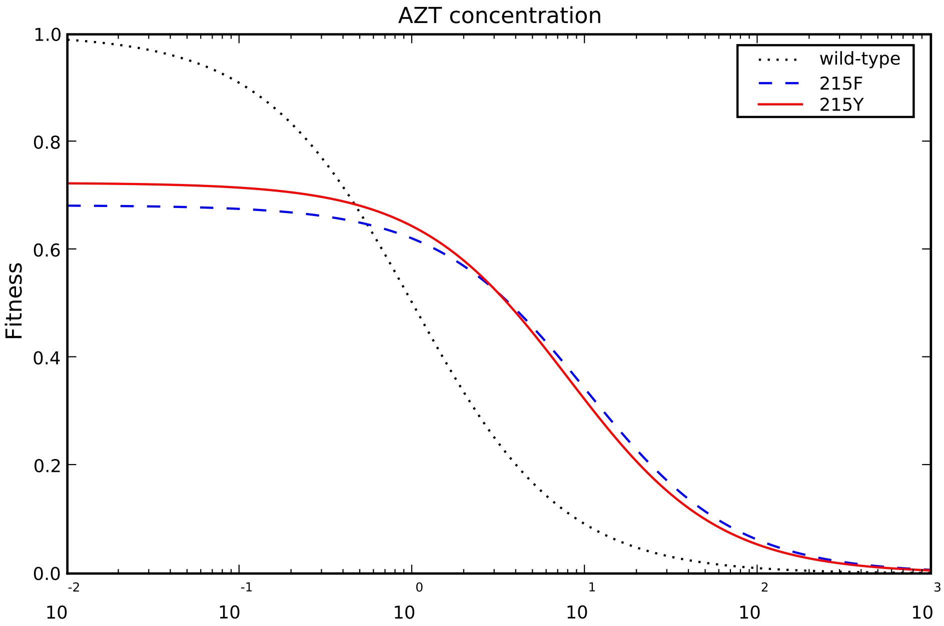


Figure 5.5: Predicted fitness for AZT-resistant RT mutants. AZT concentrations are given relative to wild-type IC_{50}.

consider a patient with mutant virus m_0 before treatment t_1, who is later determined to have mutant virus m_1 following treatment t_1. Under treatment t_1, the fitness of m_1 should be greater than the fitness of m_0. Any successive m_i, where $i > 0$, should also have greater fitness than m_0 with treatment t_i. For convenience, a treatment regimen with associated mutations is defined as a genetic selection episode (GSE).

The records of 701 patients treated with protease inhibitors were obtained from the HIVDB,[17] containing a total of 1,095 GSEs. For patients treated with RT inhibitors, 642 records were obtained, with a total of 930 GSEs. For each GSE, viral fitness was predicted for mutants m_0 and m_i in the presence of treatment t_i. Instances where the fitness of m_i exceeded that of m_0 were considered correct, while predictions where the fitness of m_i is less than that of m_0 are considered incorrect. Equal fitnesses were treated as a separate category. Overall accuracy was determined as the number of correctly assigned GSEs divided by the sum of correct and incorrect predictions. Results are shown in Table 5.3. Interestingly, the accuracy for the protease-treated samples is much higher than for the RT-treated samples. The error level was found to be high with lamivudine (3TC), and removing GSEs containing lamivudine from consideration increased accuracy.

In practice, many of the viral sequences obtained following treatment do not exhibit

Table 5.3: Fitness predictions for clinical isolates.

Protease			
Correct	Equal	Incorrect	Accuracy
601	293	191	76%

Reverse transcriptase				Reverse transcriptase without 3TC			
Correct	Equal	Incorrect	Accuracy	Correct	Equal	Incorrect	Accuracy
386	193	351	52%	185	76	78	70%

Table 5.4: Fitness predictions for clinical isolates, using a 2-fold change threshold.

Protease			
Correct	Equal	Incorrect	Accuracy
488	537	60	89%

Reverse transcriptase				Reverse transcriptase without 3TC			
Correct	Equal	Incorrect	Accuracy	Correct	Equal	Incorrect	Accuracy
289	375	266	52%	141	147	51	73%

drug resistance mutations, and have roughly the same fitness as the pre-treatment mutant m_0. An alternative scheme was used that designated fitness difference of less than 2-fold when comparing m_0 and m_i to be insignificant, and considered as ties. As shown in Table 5.4, this increased the accuracy of predictions in protease, but had little effect on the reverse transcriptase set.

Overall, these results indicate that the predicted fitnesses are more accurate in protease than in RT. Regimens involving lamivudine seemed especially problematic, and this appeared to result from difficulty in assessing the 184V mutation, which is a common lamivudine resistance mutation. The MAPP score for this mutation was high, causing a drop in predicted replication capacity that is unlikely to reflect the actual effect on replication capacity, given the prevalence of 184V. In future studies, alternative methods for replication capacity estimation or refinements in the MAPP analysis could remedy this problem.

It should also be noted that the fitness model developed in this work was designed for a more general purpose than identifying resistant viral sequences. For instance, the model's components are able to detect antagonistic relationships between mutations or identify highly deleterious mutations. However, since clinical isolates are the products of an evolutionary process that selects against these cases, the generality of the fitness model is less applicable. In contrast, when simulating this evolutionary process, a fitness measure must be able to account for individual mutations and combinations that are not clinically observed.

5.3 Simulation of HIV drug resistance evolution

5.3.1 Model

The fitness function presented in the previous section represents one of the key components in HIV evolution. Incorporating this function into a model of the HIV replication cycle allows detailed simulations of HIV drug resistance evolution, emphasizing clinically-relevant mutations more than previous simulations of viral evolution.

The general framework of the simulation mimics a serial passage experiment, where the virus is introduced to a population of uninfected cells, allowed to replicate, and the progeny used to infect a new population of cells (cf.[16,104]). Each round of passage corresponds to a single generation for the virus. The viral population is made up of individual virions, each with its own genotype and phenotype. Consistent with the diploid nature of HIV, the genotype consists of two copies of the viral genome, which may be heterogeneous. However, only the protease and reverse transcriptase genes are modeled, as comprehensive drug resistance information is not yet available for other genes. Each virion's phenotype corresponds to protease and reverse transcriptase as well. However, the genotype is represented as an RNA sequence, while the phenotype is a translated amino acid sequence.

In the simulated replication cycle (Figure 5.6), an experiment begins with infection from a population of wild-type virus. In contrast to previous experiments,[105] the infection process assumes that there is no structure to the cell population, akin to a solution rather than tissue, and the virions infect cells completely at random. Multiple infections per cell are possible, so the distribution of virions infecting a cell follows a Poisson distribution. Following infection, each virion performs reverse transcription, which allows for recombination between the two copies of the viral genome, as well as mutation. This recombination process allows for an average of nine recombination events per replication cycle, which has been reported for HIV infection in T-cells.[106] The mutation rate corresponds to one mutation per full-length HIV genome, a compromise between previously published values, which range from roughly 0.3 to 1.1 mutations per genome.[107,108] Specific nucleotide substitution rates are also taken from a previous study.[109] Both mutation and recombination rate are uniform across the entire genome. At the end of the reverse transcription process, a single nucleotide sequence is produced, and remains associated with the viral phenotype.

The fitness evaluation models degree to which this process is successful and results in integration and expression of new viral genotype, based on a the viral phenotype and presence of

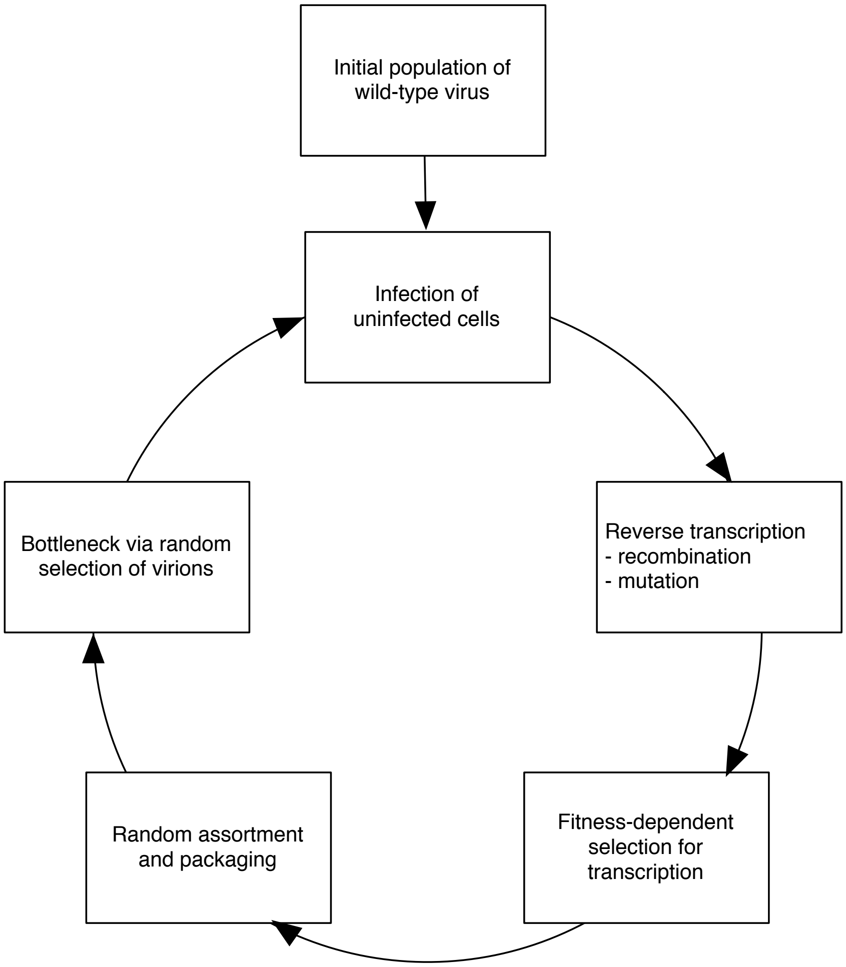

Initial population of wild-type virus
Infection of uninfected cells
Reverse transcription
- recombination
- mutation
Fitness-dependent selection for transcription
Random assortment and packaging
Bottleneck via random selection of virions

Figure 5.6: Overview of the HIV evolutionary simulation.

Table 5.5: Key simulation parameters.

Parameter	Value	Comment
Viral population size	1,000 – 2,000	effective population sizes reported in literature
Number of cells	100 – 200	
Maximum burst size	10	production of infectious particles is low
Genome length	1017 bp	size of full-length protease and partial reverse transcriptase
Mutation rate (per base)	1.01E-4	average of 1 mutation per cycle
Crossover rate (per base)	9.08E-4	average of 9 recombination events per cycle

inhibitors in the current environment. The environment is a shared feature among all cells, which specifies the presence of inhibitors and their concentrations. Fitness values are calculated using the function in Section 5.2, and the contents of new virions are apportioned using stochastic universal selection,[110] a method similar to roulette wheel selection. With this selection process, a new virion's genotype and phenotype are based proportionally on the fitness of virions that have infected the cell. For each cell, this produces up to 10 virions. A portion of the "wheel" corresponds to the production of no virion, so that infection with low-fitness virus will be likely to yield fewer progeny. As each new virion's genotype must contain two nucleotide sequences, the selection process is performed twice for each of the progeny virions. It is assumed that the phenotype corresponds directly to one of these sequences, assuming partial localization during viral packaging.

Producing a maximum of 10 virions per cell may appear low, but studies have shown that the vast majority of virions produced during infection are noninfectious.[111] Finally, a bottleneck is applied at the end of each round via random selection, which limits the total number of virions that go on to infect a new population of cells. Given that the average lifespan of an infected cell is 2.2 days,[27] a simulation run for 200 generations should correspond to roughly 1.2 years.

5.3.2 Related work

The simulation described above falls into the category of "agent-based" simulations, where a key feature is the modeling of individual virions. Generally, these simulations incorporate large populations of virions and stochastic processes, such as mutation and recombination. Individual simulation runs may display distinct behaviors due to this randomness, so the experiments are often repeated many times. Because a large number of replicates combined with a sizable viral population requires a significant amount of computing power, agent-based HIV simulations have become widely used only in recent years.

Previous agent-based HIV simulations have investigated the effect of viral population size and mutation rate on the development of drug resistance,[112] the role of multiplicity of infection and recombination in driving viral diversity,[32] and the use of RNAi to disrupt viral replication.[33] All of these studies incorporated a viral replication cycle similar to the process shown in Figure 5.6. Major distinctions between these studies stem from differences in the representation of viral genotypes and fitness evaluation. Althaus et al. examined a two-locus, two-allele model, with mutations conferring increased fitness and including a variety of assumptions regarding epistatic interactions.[112] The simulation held constant both the number of virions and cells, and limited the multiplicity of infection to two infections per cell. The production of new virions was based deterministically on the average fitness of virions infecting a cell. One of the main variables in the simulation was the population size, which varied from 1,000 to 100,000, far exceeding the population examined in the current work.

A related study by Bocharov et al. incorporated a larger viral genome and focused on smaller populations.[32] Their simulation's viral genotype consisted of a bit string of length 100, which corresponds to two alleles at 100 loci. However, only mutations at 3 of these positions were allowed to positively impact viral fitness, and the relative fitness levels were chosen arbitrarily. In addition, the fitness contribution of each beneficial mutation combined linearly. The populations were fixed at 1,000 virions and 200 cells, roughly comparable to the current work. Production of new virions was either limited by the cell or proportional to the multiplicity of infection. Mutation was limited to a single mutation per genome per replication cycle. Similarly, a single recombination event was allowed, implemented as a two-point crossover.

A much more detailed viral genotype and fitness model was employed by Leonard and Schaffer.[33] Their experiments focused on the use of RNAi to disrupt the TAR region of the HIV genome, and required the use of a nucleotide representation of the viral genome corresponding to the TAR region. In addition, mutations in this region had fitness consequences based on their ability to boost transcription and avoid RNAi. Fitness was also affected by a randomly selected host integration site. Unlike the two studies mentioned above, the work of Leonard and Schaffer allowed variations in the viral population, and the simulated RNAi therapy was able to drive the viral population to extinction. In terms of viral production, a specific burst size was not specified, but the authors did state that the population size would increase by roughly a factor of 3 in the absence of therapy. The level of detail in the viral genotype and fitness model in the Leonard and Schaffer study most closely coincides with the model presented in this work.

On the other hand, ordinary differential equation (ODE) models do not focus on the properties of individual virions, unlike agent-based simulations. Instead, ODE models describe one or more homogeneous viral populations. The most basic of these models include only three equations, corresponding to the number of free virions, infected cells, and uninfected cells.[29] By extending these equations to incorporate drug therapy, accurate predictions of viral load under ritonavir treatment were previously obtained.[27] Further extensions with multiple compartments and inhibitor classes have also been explored.[30] Two distinct viral populations were studied by Suryavanshi and Dixit, examining recombination using ODEs.[113] Capturing diverse viral populations using ODEs is difficult, however, because the equations describe homogeneous populations. This limitation is not a problem for agent-based simulations, but there is a trade-off in complexity and computational effort. While agent-based simulations can currently include hundreds of thousands of virions, ODEs are not hampered by large population sizes, and so are capable of modeling the billions to trillions of virions present during infection *in vivo*.

Other viral evolution models do not belong to either the agent-based simulation or ODE categories. The genetic barrier concept of Beerenwinkel et al.[91] and the fitness landscapes derived by DeForche et al.,[114] for example, focus on the use of probabilistic graphical models to capture the evolution of drug resistance. The work of Beerenwinkel et al. was based on the use of mutagenic trees, which specified pathways of resistance mutations, and the expected time to develop each mutation.[91] DeForche et al. combined a Bayesian network analysis with an evolutionary simulation, producing a large evolutionary graph for nelfinavir resistance.[114] Both of these approaches were able to capture important aspects of HIV fitness and evolution, but rely completely on a large volume of clinical data.

5.3.3 Evolution absent drug selection

In the absence of drugs, a retroviral quasispecies like HIV ensures that significant mutations will occur during replication, with generally negative effects on viral fitness. On average, however, the effect of these mutations is small, and some mutations may become more prevalent due to genetic drift. Also, the high mutation rate of HIV, averaging roughly one mutation per replication cycle, indicates that HIV populations should act as quasispecies.[115] This mean that an initial wild-type population will not remain homogeneous for long, and should quickly develop into a “cloud” of mutants.

Using the simulation, a viral population of 1,000 virions was repeatedly passaged over 200 generations into a population of 100 cells. The progeny virions from each generation were used to infect new cells in the subsequent generation. A summary of results from 1,000 independent simulation runs is shown in Figure 5.7. During the early stages of the simulation,

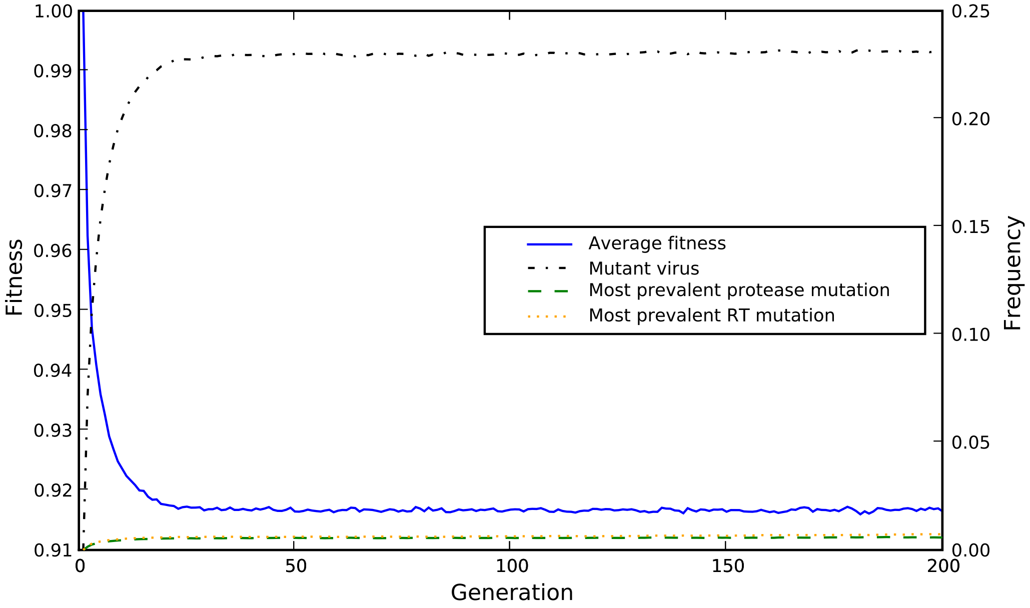


Figure 5.7: Simulated evolution absent drug selection.

the fraction of mutant virus climbed dramatically, causing a drop in average viral fitness. After roughly 20 generations, these quantities plateaued, indicating that an equilibrium had been reached, where mutations from wild-type and the negative impact of these changes were in balance. In later stages, mutant virus made up more than 20% of the viral population, and this mutant population appeared highly heterogeneous, as the most prevalent mutations in both protease and RT were present in less than 1% of the population. The behavior of the viral populations in this simulation demonstrated that diversity arises quickly.

The prevalence of mutations in the simulated population showed a small correlation with viral mutations observed in untreated patients (Figure 5.8), which were obtained from the HIVDB.[17] Though the level of correlation was low, p-values were less than 0.01 even at the 10th generation. By the 200th generation, the end of the simulation, the Kendall τ values had increased to 0.13 and 0.15 for protease and RT, respectively, with p-values less than 10^{-6}. In the absence of drug selection, the frequency of mutations in the simulation is driven by the replication capacity component of the fitness function, which is based on MAPP scores. The statistically significant correlation provides further evidence that these scores are able to predict impairment in viral proteins.

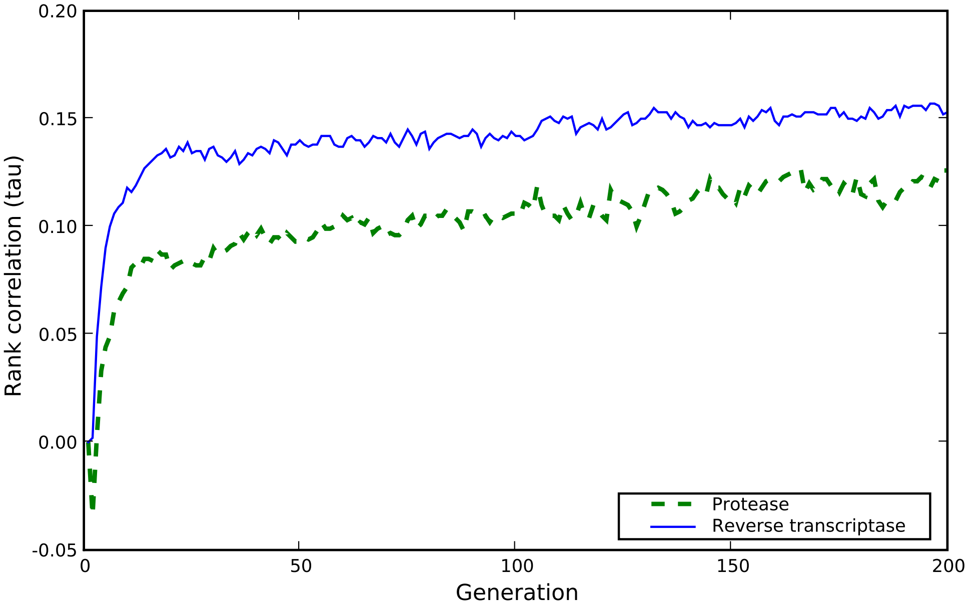


Figure 5.8: Correlation between simulated and observed mutation prevalence in untreated population.

5.3.4 Drug selection and mutation prevalence

The relative prevalence of mutations arising during drug therapy is due to a variety of factors. Presumably, mutations that confer resistance against particular inhibitors will be more prevalent than those that do not, but considerations of replication capacity as well as the context of other mutations also come into play. Analysis of the individual contributions of mutations to drug resistance have been reported through linear regression.[19,20] In that study, each mutation was represented by a regression coefficient that indicated the level of resistance conferred by the presence of the mutation.

The contribution of each mutation to resistance should be a major factor in the evolution of the virus in response to drug therapy, reflected in the prevalence of the mutations in clinical isolates. The fitness function used by the evolutionary simulation incorporates the regression coefficients as measures of drug resistance, but further includes estimates of replication capacity and epistasis. In addition, the simulation carries out an evolutionary process, allowing complex dynamics to arise in the development of resistance. It is expected that the mutations found at the end of simulations should show better correspondence to clinical prevalence than the regression coefficients alone.

Information on clinical prevalence was retrieved from the Stanford HIV Drug Resis-

Table 5.6: Correlation between clinical prevalence and either drug resistance regression coefficients or evolutionary simulation results.

Inhibitor	Regression coefficients (τ)	p	Simulation outcome (τ)	p
AZT	0.19	0.068	0.14	0.20
IDV	0.25	0.004	0.44	5.2e-7
NFV	0.14	0.12	0.37	3.6e-5

tance Database.[17] As in Section 5.1.1, to avoid confusing the effects of multiple inhibitors, only records resulting from AZT, indinavir, or nelfinavir treatment were used. The regression coefficients were obtained from the Rhee et al. study,[20] and are the same ones used in the simulation's fitness function for RF calculation (cf. Section 5.2). To eliminate polymorphic mutations from consideration, only treatment-selected mutations[20] were used when performing comparisons. For the evolutionary simulation, an initial viral population of 1,000 virions was passaged in 100 cells for 10 generations in the absence of inhibitors. After the 10th generation, the concentration of inhibitor was increased by one unit per generation until the end of the simulation at generation 200. The viral population at the end of 1,000 such simulations was used to calculate the prevalence of mutations.

The degree of correlation between clinical prevalence and the drug resistance regression coefficients and evolutionary simulation results was calculated using Kendall's τ rank correlation. For the regression coefficients, correlation was weak and generally insignificant, as shown in Table 5.6. However, the evolutionary simulation showed significant correlation with indinavir- and nelfinavir-treated records. This confirms that the simulation shows better correspondence to clinical prevalence than the regression coefficients, at least for protease. Results were poor for either method when considering mutation prevalence arising during AZT treatment. A similar outcome was noted when considering GSEs in Section 5.2, where predictions for protease were more accurate than for RT. The problems noted in the earlier work, especially difficulties in estimating replication capacity in RT, would also cause problems during the evolutionary simulations. Overall, these results show that the evolutionary simulation is able to reproduce clinically-observed prevalence for protease inhibitors with significant accuracy, while highlighting shortcomings in the modeling of reverse transcriptase.

5.3.5 Protease inhibitor combination therapy

The current guidelines for the clinical treatment of HIV do not recommend the use of multiple protease inhibitors[116] (excepting the addition of ritonavir as a boosting agent). Due to the high degree of cross-resistance between existing protease inhibitors, this appears to be a

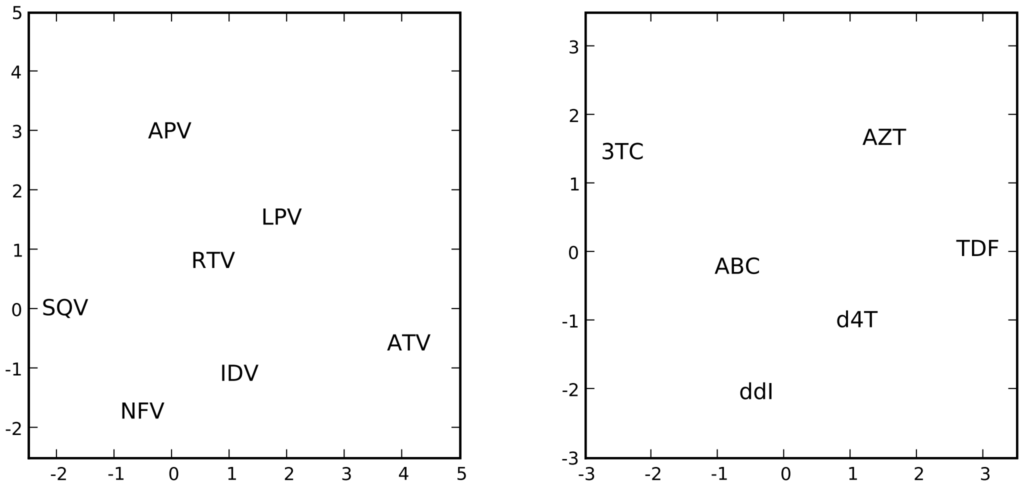


Figure 5.9: Nonlinear maps of drug resistance for protease inhibitors (left) and NRTIs (right). Distances are based on the inhibitors' resistance profiles with respect to specific mutations, and Sammon's nonlinear mapping is used to orient the drugs in two dimensional space.[117]

sensible strategy. However, particular combinations of protease inhibitors may exhibit less cross-resistance. Based on a comparison of drug resistance regression coefficients, the level of cross-resistance between various inhibitors may be depicted graphically, as in Figure 5.9. This figure indicates, for example, that a combination of atazanavir and saquinavir may be more effective than a combination of atazanavir and lopinavir. In comparing the ability of different treatment regimens in suppressing viral replication, it is expected that combinations of protease inhibitors that minimize cross-resistance will be more effective than high levels of single inhibitors.

To test this hypothesis, the ability of different treatments to suppress viral replication was examined through simulation. These treatments included the use of the individual protease inhibitors currently recommended for clinical use: lopinavir, atazanavir, amprenavir, and saquinavir. Three combinations of two protease inhibitors were also used, lopinavir + atazanavir, lopinavir + amprenavir, and atazanavir + saquinavir. Based on the lower level of cross-resistance shown in Figure 5.9, the atazanavir + saquinavir combination seemed likely to be the most effective.

The effects of different protease inhibitor-based regimens were simulated using an initial population of 2,000 wild-type virions, passaged in a population of 200 cells. Ten generations of replication were allowed before the onset of drug therapy, in order to allow mutations to develop. After the start of therapy, the viral population was monitored for extinction, i.e. a viral population of 0. In a clinical setting, antiviral therapy cannot eliminate HIV completely from a

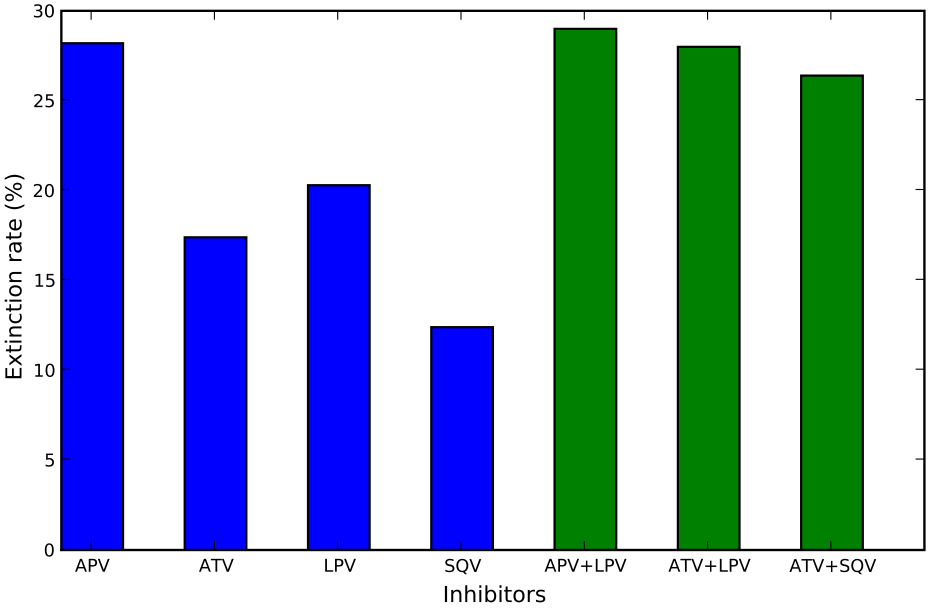


Figure 5.10: Extinction rates with simulated protease inhibitor therapy.

patient, but extinction in the simulation may be considered analogous to strong suppression of the virus. Each treatment was tested in 1,000 independent simulation runs. To better balance the treatment results, single inhibitor therapies were used at a concentration of 20 units, while combinations used 10 units of each inhibitor.

The rates of extinction given the various protease inhibitor therapies are shown in Figure 5.10. Combinations performed better than the average single inhibitor, though amprenavir alone was competitive with the combinations. Among the single inhibitor treatments, effectiveness varied widely, with saquinavir causing extinction less than half as often as amprenavir. While saquinavir and atazanavir were the weakest inhibitors individually, combining them resulted in extinction rates comparable to the other pairs. Overall, combining existing protease inhibitors did appear to increase extinction rates, but not far beyond the most effective single inhibitor.

5.3.6 Allosteric protease inhibition

While cross-resistance may limit the efficacy of combinations of clinically-approved protease inhibitors, novel inhibitors targeting alternative binding sites should exhibit minimal cross-resistance with current protease inhibitors. Chapter 2.2 mentions an alternative binding site in HIV protease that is thought to affect protease mobility. Directly targeting this "exo-site" using a virtual screen is described in Chapter 6.3. This work resulted in the discovery of two protease inhibitors which displayed behavior consistent with allosteric inhibition. Further confirmation of the binding modes will require further experiments, but the significance of an allosteric protease inhibitor can be studied *in silico*.

First, in anticipating drug resistance, since an allosteric inhibitor would target a region of the enzyme away from the active site, the resistance profiles of existing inhibitors would not identify relevant drug resistance mutations. However, structural analysis provides the opportunity to move beyond these prior experiences. Further, joining structure-based predictions of drug resistance with the evolutionary simulation provides an opportunity to gauge the utility of a putative inhibitor, both individually and in combination with existing inhibitors.

Prediction of drug resistance mutations

A major component of the evolutionary simulation is the prediction of RF, which in prior experiments has been based on the properties of known inhibitors *in vitro*. Because there are currently no known protease inhibitors that target the exo-site region of the enzyme, drug resistance predictions cannot be made based on existing experimental data. Structural analysis is a useful alternative, and was previously applied to the protease inhibitor AB2 in Chapter 4.3. The analysis calculates changes in binding energy, which are proportional to the logarithm of RF. The values are, in turn, equivalent to the regression results used in calculating resistance for existing inhibitors.

To perform a structure-based probe of resistance in the exo-site, one of the compounds found using a virtual screen was used as a prototype allosteric inhibitor. This compound, NSC 45621, was predicted to bind the outer surface of protease (Figure 6.9), forming favorable interactions with a large number of residues. The 10 positions contributing most heavily to the binding energy were identified using the AutoDock4 force field[40] and are shown in Table 5.7. For each of these positions, all possible substitutions were performed independently using SCWRL3,[86] generating single-mutant protease structures. The use of all possible substitutions is likely to exaggerate the number of mutations which contribute to drug resistance, but provides a pessimistic bound on the development of resistance to an inhibitor of this type. As with the AB2 predictions in Chapter 4.3, the difference in binding energy caused by each these mutations was evaluated using the AutoDock4 force field.

The binding energy differences were used to fill the role of the drug resistance re-

Table 5.7: Positions in protease with largest contributions to protein-ligand binding energy with NSC 45621.

Position	ΔG
70	-1.76
63	-1.51
14	-1.30
62	-1.08
61	-0.91
41	-0.81
60	-0.65
39	-0.39
43	-0.34
16	-0.27

gression coefficients in RF determination. Restricting the changes in binding energy to be $\geq$ -1 kcal/mol removed a small number of extremely unfavorable values, and also resulted in values with a similar distribution to the existing regression coefficients. Across all protease inhibitors, the regression coefficients averaged 0.31, with a standard deviation of 0.50. The changes in binding energy averaged 0.36, with standard deviation 0.48. The rough similarity between distributions indicates that the level of resistance is not being grossly underestimated against the predicted allosteric inhibitor.

Effects of an allosteric inhibitor

As with Chapter 5.3.5, the rate of extinction resulting from different therapies was measured via simulation, this time with an allosteric inhibitor. Alone, this "EXO" inhibitor was far worse than any of the existing drugs, and was able to drive a viral population to extinction only 2% of the time. Resistance mutations against the allosteric inhibitor were able to develop with relative ease because sequence variability tends to be high outside of protease's active site (see Figure 3.3), allowing mutations to occur with little consequence for enzyme function.

However, when combined with existing drugs, the allosteric protease inhibitor was much more effective (Figure 5.11). These combinations resulted in extinctions in nearly all cases. In contrast, combinations of existing inhibitors were not significantly more effective than single-inhibitor treatments, and resulted in extinctions in less than 30% of cases. In combination with the allosteric inhibitor, the strength of the individual active site inhibitors appeared to have little effect, as saquinavir was found to be weaker than amprenavir, but combinations involving either drug and the allosteric inhibitor were both highly effective. Even when used at a lower concentration, the allosteric inhibitor was still effective when in combination with another inhibitor, as shown in Figure 5.12.

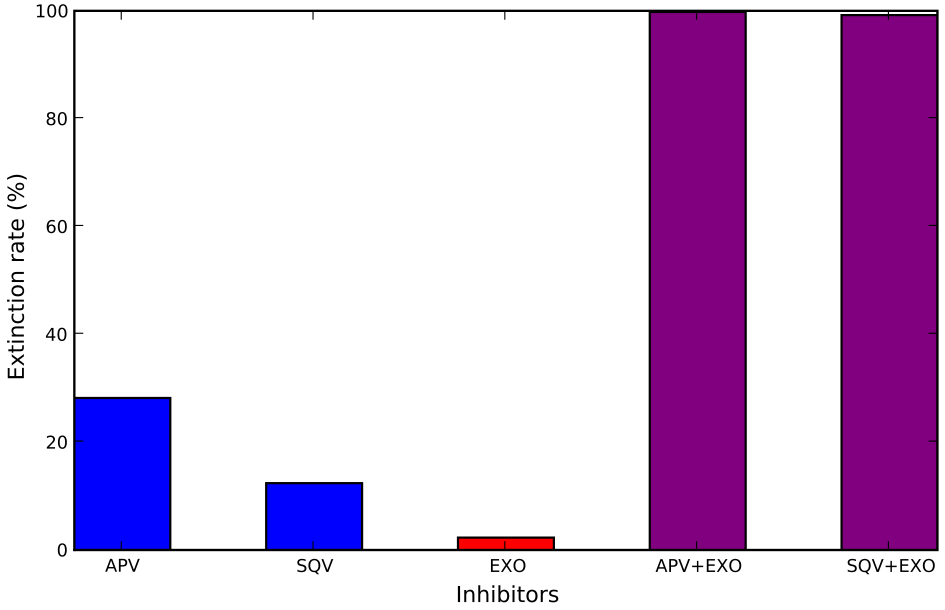


Figure 5.11: Extinction rates with simulated protease inhibitor therapy, including an allosteric inhibitor (EXO).

Aside from the inhibitor concentration, a major factor in the use of inhibitors with different binding modes is their degree of independence in binding. Clinically-approved protease inhibitors all target the enzyme's active site, so the presence of amprenavir, for instance, will exclude saquinavir. An inhibitor targeting a distinct binding site may not exhibit mutually exclusive binding with one of the active site inhibitors, which would improve the overall level of inhibition when used in combination. The above results were based on the assumption that an allosteric inhibitor would be mutually exclusive with amprenavir or saquinavir. If the binding events are instead assumed to be independent (i.e. the presence of one inhibitor has no effect on the other), the combination therapy appears even more effective (Figure 5.12). Any synergism in the binding of these inhibitors would increase the extinction rates even further. On the other hand, antagonism is bounded by the assumption that the allosteric inhibitor is mutually exclusive with an active site inhibitor. In any case, the allosteric inhibitor remained effective at lower concentrations than the inhibitors that it was paired with.

Overall, combinations including the allosteric inhibitor showed substantial advantages over combinations of existing inhibitors, even the pair predicted to exhibit the least cross-resistance. The characteristics of this putative inhibitor reflect a pessimistic outlook – the region that it targets is assumed to be highly variable, allowing many mutations that decrease inhibitor

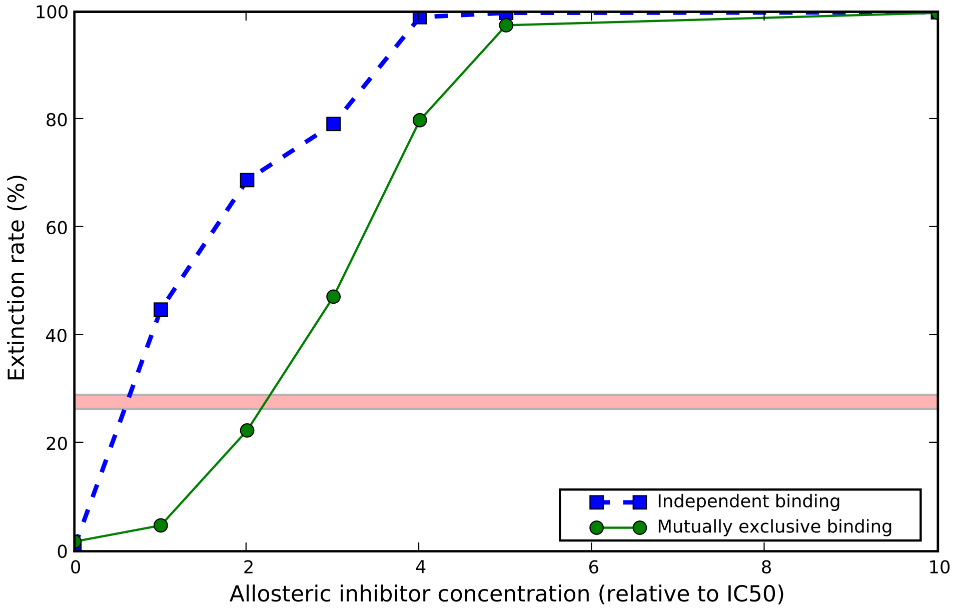


Figure 5.12: Extinction rates with varying concentrations of allosteric inhibitor, assuming either independence or mutual exclusion with amprenavir binding. Each point represents the allosteric inhibitor in combination with 10 units of amprenavir. The red horizontal bar indicates the range of extinction rates observed in previous experiments with combinations of clinically-approved protease inhibitors.

binding with a minimal penalty to enzyme function. When used singly, the weaknesses of this inhibitor were evident in the low rate of extinction in comparison to the clinically approved drugs. However, when used in combination with the clinically approved drugs, the allosteric inhibitor greatly increased extinction rates. As the combination remains effective with relatively low concentrations of the putative inhibitor, even a compound which binds an alternative site with relatively low affinity may still be useful. The current prototype for an allosteric inhibitor is effective against HIV protease at roughly 100 times the concentration of existing inhibitors, but optimizations could boost the affinity further. Even an increase of 10-20 fold would provide allow protease inhibitor combination therapy far more effective than current treatments.

Chapter 6

Docking and Drug Discovery

In recent years, virtual screening has become a useful tool in drug discovery, widely used to replace or support high-throughput screening (HTS) efforts.[37,38,118–120] A key procedure in virtual screening is protein-ligand docking, which models interactions between a small molecule and protein, seeking to find the level of binding affinity and optimal binding mode. AutoDock[39,40] is a widely-used docking program that has been used with many drug targets, including HIV protease.[21,23,121] Recently, AutoDock has been deployed as part of a distributed computing project for HIV drug discovery called FightAIDS@Home. While traditional virtual screening efforts may focus on screening large libraries of ligand compounds against a single protein structure, or a single ligand against multiple related (e.g., mutant) structures, the computing resources provided by FightAIDS@Home allow considerations of large number of structures. This capacity was used to dock a chemical library against multiple HIV protease mutant structures. Section 6.1 provides an analysis of these results, including insights into the distribution of binding energies expected from a random library screen and the relationship between protease structures based on their interaction with these compounds.

For virtual screening, accurate determinations of binding energy are key in selecting the most likely inhibitors. Generally, binding energy estimates in protein-ligand docking programs have not included detailed calculations of entropic forces, such as configurational entropy. In Section 6.2, several methods for approximating configurational entropy in docking are evaluated using experimental information from APS reductase and its binding affinity with a set of ligands.

In the last section of this chapter, insights from the FightAIDS@Home analysis and study of entropy are applied toward protease inhibitor screening, culminating in an experimental screen of the best candidates. Generally, experimental screens use very large chemical libraries, as perhaps less than 1 in 1000 of the compounds tested will show significant inhibition of the target.[34] To address this problem, the HIV protease screen focuses on the use of a library which has proven useful in a cell-based anti-HIV assay. In addition, to reduce the high false positive rate common in virtual screening,[119] a novel approach comparing ligand binding at multiple HIV protease sites was used. This comparison was also helpful in targeting a possible allosteric inhibition site in protease. Biochemical assays of 38 selected compounds showed that five were able to inhibit HIV protease at low micromolar concentrations, including two compounds that are predicted to bind in the putative allosteric site.

6.1 Analysis of HIV wild-type and mutant structures via docking

The FightAIDS@Home* project (FAAH) utilizes the World Community Grid distributed computing network to conduct virtual screens for new inhibitors against HIV protease. The project is built around AutoDock,[39] which uses a Lamarckian genetic algorithm (a hybrid of evolutionary algorithm sampling with local search methods) to search for the optimal conformation of a given ligand in relation to a target receptor structure. Currently, FAAH is installed on approximately 450,000 clients and is capable of screening almost 10,000 ligands per day. As part of a larger screening process seeking new protease inhibitors effective against both wild-type HIV and emerging drug resistant mutants, FAAH completed an initial screen of approximately 1,800 ligands against 268 HIV protease structures, totaling almost 500,000 different dockings and more than 10^{15} separate energy evaluations.

Though the scale of these *in silico* experiments are huge, growing databases of proteins and chemical structures have the potential to surpass these resources. As FAAH moves forward, techniques to judiciously choose informative structures and ligands remains important. The set of protease structures considered in FAAH includes a large number of modeled structures. This analysis focuses on structures taken from the Protein Data Bank (PDB),[85] consisting of 71 wild-type and mutant proteases. More specifically, these structures include 26 wild-type HIV-1, 33 mutant HIV-1, and 12 HIV-2. The ligand library used in the current FAAH experiment consists of 11 known protease inhibitors and compounds from the National Cancer Institute (NCI) Diversity Set †. The NCI Diversity Set is chosen specifically to represent a broad sampling of pharmacophores, providing a characterization of protease docking modes "outside the box,"

*http://fightaidsathome.scripps.edu/

†http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html

Table 6.1: Overview of FAAH ligands and protease structures.

Proteases	wild-type	mutant	HIV-2
number of structures	26	33	12
unique (by sequence)	~1	10	2

Ligands	known inhibitors	NCI Diversity Set
number of compounds	11	1,760

beyond those represented by currently approved protease inhibitors. An overview of the current FAAH dataset is shown in Table 6.1.

One of the long-term goals of this research is to discover compounds that can inhibit a broad range of mutant proteases, so a variety of HIV protease mutant structures are considered. A similar method has previously proven successful in developing an inhibitor effective against FIV, SIV, and HIV.[122] Due to the rapid evolution of drug resistance, such approaches are vital in the design of new inhibitors. Cross-resistance involving current FDA-approved drugs is also a continuing problem.[123]

Other research has addressed docking against an ensemble of protein structures.[124, 125] However, these studies focus on molecular dynamics-based "snapshots" of the protein in motion. A more recent paper by Fernandes et al. addresses the problem of docking to multiple structures in an effort to develop inhibitors effective against a set of targets.[126] Their work includes the comparison of docking results from several HIV protease structures, showing that ligands adopt similar binding modes across various proteases.

Our work incorporates a much larger number of ligands and proteases. Hayashi et al. use such methods to generate ligand profiles by docking small molecules against a panel of various proteins.[127] A key feature of their work is the use of vectors of binding energies as a way of describing particular ligands. Complementary methods that focus on proteins rather than ligands are discussed in this study, with applications toward finding consensus and representative proteases. The consensus protease structure that best captures the central tendency of the larger set of structures would prove useful in more focused virtual screening experiments. Such a structure was constructed by Vinkers et al., using averaged 3D coordinates from a set of crystallographic data.[128] We describe an approach based on binding energy profiles below.

6.1.1 Methods

Dataset

The ligand library used in FAAH consists of 11 known protease inhibitors and 1,990 compounds from the NCI Diversity Set. Known inhibitors include 8 FDA-approved compounds: amprenavir, atazanavir, indinavir, lopinavir, nelfinavir, ritonavir, saquinavir, and tipranavir. The remaining 3 known inhibitors are TL-3, KNI-272, and JE-2147. Structures for all of the known inhibitors can be found in the PDB. Of the compounds from NCI, 153 could not be processed correctly for AutoDock, due to the presence of metal atoms or multiple fragments, and so were not included in this study. An additional 77 were removed due to extremely poor binding, leaving a total of 1,760.

Characterizing "wild-type" for a quasi-species like HIV is a notoriously difficult problem. Two common characterizations of subtype B of the HIV-1 virus are the "consensus B" sequence and "HXB2".[4] Since protease for these two differ only at positions 3 and 37, sequences matching either are considered wild-type HIV-1 structures. The protease structures analyzed include 26 wild-type HIV-1, 33 mutant HIV-1, and 12 HIV-2.

Docking Protocol

Atomic coordinates for the HIV proteases were obtained from the PDB. The ligand and crystallographic waters were removed with the exception of the water bridging the flaps. When absent from the crystal structure, a water molecule was placed with hydrogen atoms oriented to facilitate the hydrogen bonding pattern commonly observed in HIV protease.[129] Polar hydrogens were added and Kollman charges were assigned to all atoms. Affinity grids centered on and encompassing the active site were calculated with 0.375 Angstrom spacing using AutoGrid 4.

The NCI Diversity Set was processed for input to AutoDock 4. Gasteiger charges were assigned to all atoms and rotatable bonds were assigned using AutoDockTools.

AutoDock 4 was used to evaluate ligand binding energies over the conformational search space using the Lamarckian genetic algorithm. Default docking parameters were used with the following exceptions: ga_pop_size, 200; ga_num_evals, 10000000; ga_run, 100. For this study, only the minimum energy found is considered.

6.1.2 Results

Discrimination between specific and non-specific interactions

AutoDock seeks the best interaction energy between a flexible ligand and the protein surface. Computed energies are typically favorable since the docking procedure searches widely. Since FAAH is ultimately focused on lead discovery, accurate discrimination between weak and strong binding is of vital importance. Toward this end, the differences in binding energy between NCI diversity compounds and known inhibitors can be used to determine a threshold at which interactions become significant.

Nearly all of the diversity compounds exhibit weak or moderate binding energies when compared to the known inhibitors. In order to determine the threshold at which specific binding is expected, the distribution of binding energies for the NCI Diversity Set is compared against the energies from the known inhibitors. At least with respect to HIV protease, derivation of a specific-interaction threshold represents an especially appropriate system due to the availability of a large number of positive controls (known inhibitors). The distribution of binding energies for both known inhibitors and the NCI Diversity Set is shown in Figure 6.1a.

The Receiver Operating Characteristic (ROC) curve in Figure 6.1b demonstrates the effect of several threshold values that attempt to separate the known inhibitors and diversity compounds. For the purposes of the plot, the positive class contains only known protease inhibitors and the negative class contains all ligands from the NCI Diversity Set. A -7.0 kcal/mol threshold was selected as the significance cutoff in future experiments. At this level, only a small fraction (5.3%) of all dockings are considered "specific" interactions, which includes 97.7% of known inhibitor dockings and 4.7% of NCI Diversity Set dockings.

Figure 6.2 shows the degree to which predicted ligand-protease interaction energies exceed the -7.0 kcal/mol threshold. These are organized as portrayed in Table 6.1, with known inhibitors along the top and wild-type protease to the left. In addition, ligands have been sorted by average interaction energy, with most favorable (i.e., most negative energies) near the top. As shown in the figure, there are wide variations in binding energy for single ligands docked against multiple proteases. Note that there are noticeable differences even among relatively homogeneous sets, such as the wild-type structures (columns 1-31). This variation underscores the importance of judicious protein structure selection in order to obtain the best binding energy estimates.

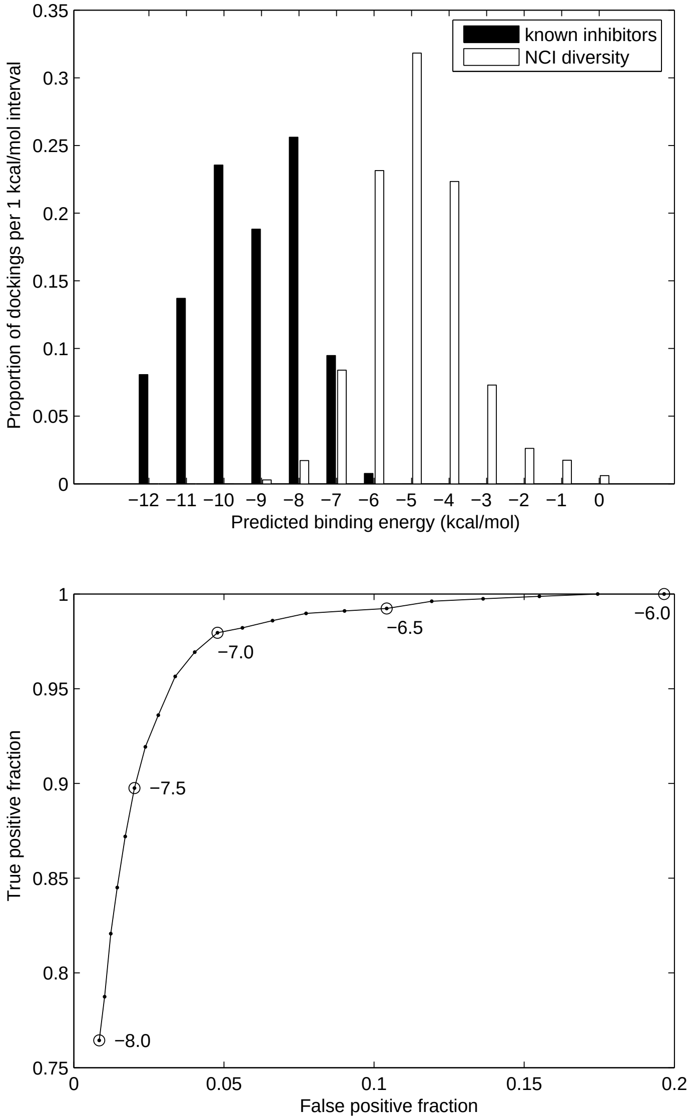


Figure 6.1: (a) Comparison of the distribution of binding energies for known inhibitors and NCI Diversity Set compounds. (b) ROC curve showing a sensitivity/specificity trade-off for threshold values from -8 to -6 kcal/mol.

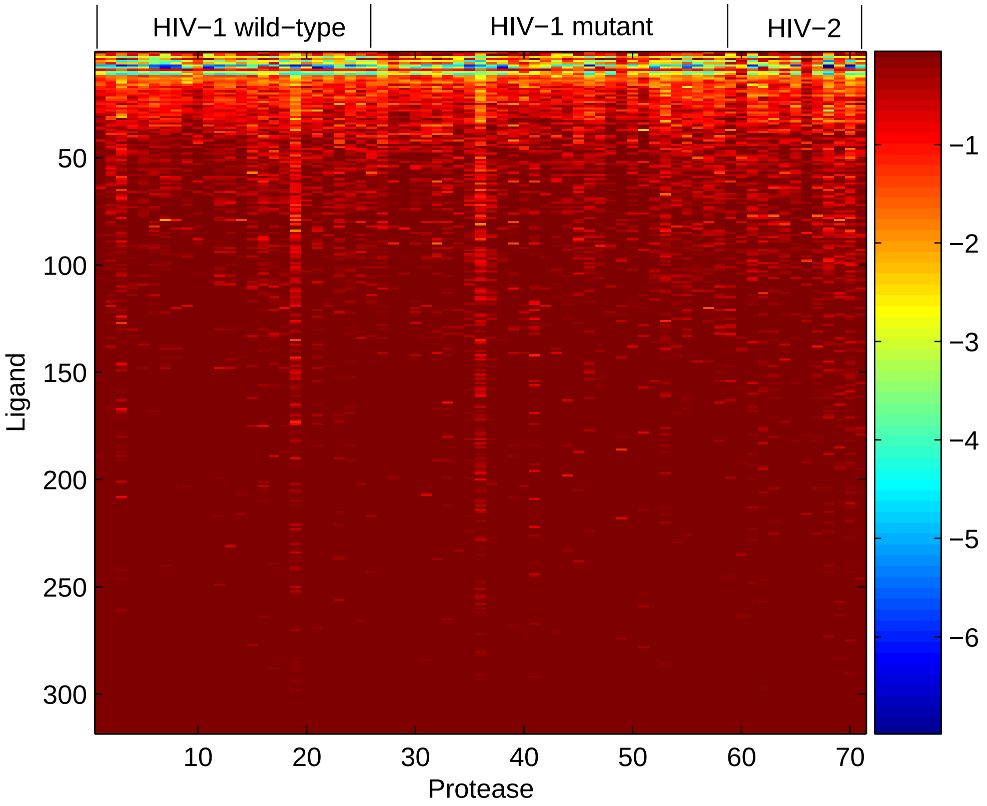


Figure 6.2: Specific energy interaction map. Each energy value indicates the level of binding beyond -7.0 kcal/mol. Ligands are sorted by ascending average binding energy.

Determining a consensus HIV protease structure

The large number of both ligands and protease structures tested in FAAH Stage 1A represents an opportunity to analyze the similarity of ligand/protein interactions across both dimensions of variability. In general, if a ligand binds poorly with one protease structure, it is not expected to bind strongly to others. If a single protease structure is found to capture the central tendency of the entire set, this single probe can be used as an initial probe against large libraries of ligands.

Representing protease structures as vectors of binding energies allows a direct mathematical characterization of a consensus protease structure as their *centroid*. That is, each of the proteases corresponds to a point in a high-dimensional space. The centroid is found by taking an average across all 1,760-element vectors of binding energy values.[‡] Using a Euclidean distance measure, 2BPW is the closest structure to the centroid. This remains true, whether the average is taken across only wild-type protease, or across all proteases.

Representative protease structures

While the centroid provides a convenient characterization of the central tendency across all proteases, identification of a larger set of "spanning" protease structures is also useful, to allow efficient screening of large libraries that capture the full breadth of observed results. To generate such a set of representative protease structures, principal component analysis (PCA) was used on the matrix of protease-ligand binding energies. By convention, columns of the matrix correspond to proteases and rows represent ligands.[§]

In brief, PCA identifies a small set of principal components (orthogonal basis vectors) that capture most of the variance within high dimensional data sets. Because the principal components are linear combinations of the observed data, they cannot be interpreted directly. Since we seek a small set of spanning, nearly orthogonal protease structures, we therefore consider those proteases which *load most heavily* along each principal component.

We consider a 10-dimensional PCA. The first principal component serves as a scaling factor, accounting for approximately 90% of the variance in the data set, with all protease loading coefficients very close in value. The sum of the second through tenth eigenvalues account for approximately 70% of the remaining variance. Any protease's loading coefficient on a principal

[‡] The specific interaction threshold deduced earlier could be used to eliminate less favorable docking results from this analysis. However, since the analysis gains information from the full set of both specific and non-specific binding energies, instances of weak and non-specific binding are retained.

[§] As above and for the same reasons, weak and non-specific interaction energies are included in this analysis.

Table 6.2: Representative protease structures. The coefficient for each structure is at least 2 standard deviations from the mean for at least one principal component. An * indicates proteases that are maximally loaded across at least one principal component.

PDB ID	description
1HII*	HIV-2
1GNM	HIV-1 with V82D mutation
1BDL*	HIV-1 with heavily mutated 30-loop
2BPZ*	HIV-1 wild-type
7UPJ	HIV-1 wild-type
1AJX	HIV-1 wild-type
5UPJ	HIV-2
1HVI	HIV-1 wild-type
1HVJ	HIV-1 wild-type
1HVK	HIV-1 wild-type
1HSI*	HIV-2 apo (no ligand bound in crystal structure)
1AID*	HIV-1 with minor drug resistance mutations
3AID	HIV-1 with minor drug resistance mutations
1BDQ*	HIV-1 with heavily mutated 30-loop and drug resistance mutations
1MEU*	HIV-1 with major drug resistance mutations

component greater than two standard deviations from the mean is deemed significant, and the corresponding protease added to the set of spanning protease. The resulting set of 16 representative structures are shown in Table 6.2, also shown are those structures that maximally load on a principal component .

These results align closely with expectations that major structural changes should affect binding energy. From the proteases studied, the main delineations are the presence of a heavily mutated loop region, the absence of a ligand in the crystal structure, and HIV-1 versus HIV-2. Drug resistance mutations also seem to play an important role.

The first two principal components can be visualized in a 2-dimensional space; cf. Figure 6.3a. The set of representative structures roughly bounds the periphery, while the consensus structure is centrally located. A multidimensional scaling plot of the same data using Sammon's nonlinear mapping (NLM)[117, 130] in Figure 6.3b demonstrates this behavior even more clearly.

In contrast to PCA, NLM reduces dimensionality via explicit local gradient minimization of a "stress" (error) function:

$$E = \frac{\sum_{\mu=1}^{n-1}\sum_{\nu=\mu+1}^{n} \frac{[d^*(\mu,\nu)-d(\mu,\nu)]^2}{d^*(\mu,\nu)}}{\sum_{\mu=1}^{n-1}\sum_{\nu=\mu+1}^{n} d^*(\mu,\nu)} \tag{6.1}$$

reflecting cumulative error $d^*(\mu,\nu)-d(\mu,\nu)$ in measuring the distance between n pairs of points

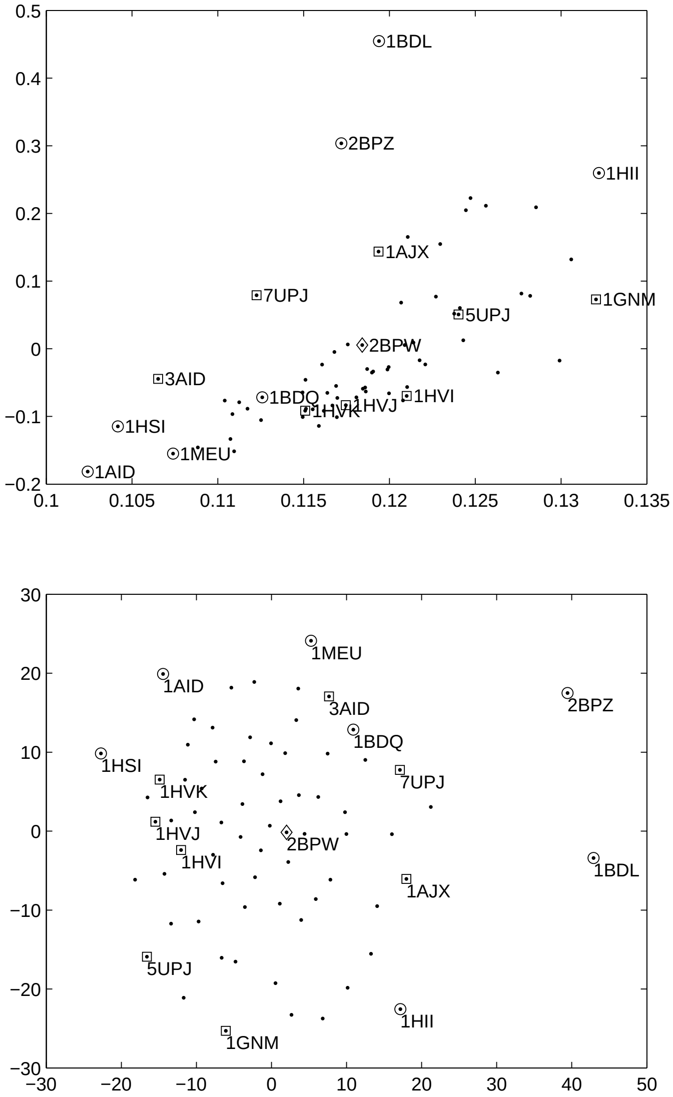


Figure 6.3: Representative protease structures plotted using (a) first two principal components and (b) multidimensional scaling with Sammon mapping. Maximally loaded structures are labeled using circles, other highly loaded structures are labeled using squares. The consensus protease structure, 2BPW, is represented with a diamond.

μ,ν in the original d^* data space vs. the reduced-dimensional space d. Given the stress function, minimization can be accomplished by a number of algorithms. Local search methods consider gradient change in stress within local neighborhoods, as potential placements in the reduced dimensional space are considered. In general, as with all local minimization procedures, there is no guarantee that the reduced dimensional solution is unique or globally optimum. In the current application, however, the Sammon mapping helps to confirm the basic pattern of the PCA solution, and provides additional indications that the particular mutants identified do indeed span the larger set.

Binding energy/sequence relationship

In ideal cases, relationships between protein sequence and function are obvious. For example, when dealing with protein crystal structures, factors other than sequence can have major effects. To determine the degree to which sequence and binding energy are coupled in this data set, a comparison between a sequence similarity matrix and a binding energy correlation matrix was performed. The sequence similarity matrix corresponds to the fraction of identical positions between sequence pairs. For binding energies, the matrix containing pairwise Pearson linear correlation coefficients is calculated. While the correlation between the two matrices is low, $r = 0.232$, the Mantel test demonstrates a statistically significant relationship between them. Using 100,000 random permutations, the empirically derived p-value is 0.015.

6.1.3 Discussion

The huge computational resources provided by the FAAH project have provided a wealth of docking information. In addition to the primary purpose of identifying novel inhibitors, this data can be used to calibrate and focus future experiments. Several novel analyses were carried out using the large body of docking results. Considering protease structures in the context of *in silico* dockings against diverse libraries of ligands provides a perspective on how similarities and dissimilarities among them that could not be anticipated, for example on the basis of sequence identity alone.

In a comparison of binding energies between compounds specifically designed to act as protease inhibitors and approximately random compounds drawn from the NCI Diversity Set, a threshold of -7.0 kcal/mol works well to discriminate between putative specific and non-specific binding with HIV protease. Applying this threshold to datasets may be useful in filtering out noise in weakly binding compounds. While this cutoff is specific to AutoDock and the protease system, the general approach is broadly applicable to other users of AutoDock (a widely used tool) and other docking systems.

The consensus protease structure, along with the other representatives, constitute a limited set which captures the breadth of the entire set of protease structures. Rather than directly capturing specific structural elements, these structures are characterized by their affinity with a diverse set of ligands. The PCA-based approach for choosing representatives is able to capture protease structures that lie on the periphery of the data set (Figure 6.3). Further applications of this technique may be useful in broader structural comparison and classification. In the more immediate future, the set of representatives will allow FAAH to continue screening larger libraries while maintaining breadth in its range of targets.

The enormous search capacity provided by the FAAH computing platform allows *in silico* experimentation on an unprecedented scale. It is not a coincidence, however, that the primary techniques we describe in this paper are all designed to *restrict* our experiments to especially informative cases; selection of the "centroid" wild-type structure and "spanning" viral structures provide two examples. Despite strong growth in computing power we can anticipate for the foreseeable future, high-throughput experimental methods and growing libraries of potential ligands generate a range of potential experiments that dwarf even these resources. Techniques supporting the judicious selection of informative structures and ligands will need to grow apace.

6.2 Empirical docking entropy

6.2.1 Introduction

The AutoDock3 and AutoDock4 empirical free energy force fields have been calibrated against a set of several hundred ligand-protein complexes of known structure and binding constants.[39,40] In our experience, this force field has been effective for the prediction of binding constants with tight-binding complexes, but we have noticed two significant problems.

First, we often find an incorrect conformation with slightly more favorable energy than the experimentally-observed conformation. However, these incorrect conformations are found with very low frequency when multiple docking experiments are performed: incorrect, low energy conformations will be found in ~1% of docking experiments, and the correct conformation will be found in 25-100% of the experiments. Thus, in these cases, a simple procedure that chooses the conformation of best energy from a set of multiple docking experiments will yield an incorrect conformation.

Second, the current force field poorly predicts the free energy of binding of weakly-interacting molecules. An example from APS reductase (adenosine 5'-phosphosulfate reductase), the subject of this report, highlights the problem. Experimentally, 5'-AMP binds tightly but 3'-AMP, which has a similar number of atoms and functional groups, binds weakly. However, in AutoDock both are predicted to bind tightly with similar binding constants. However, by looking at the frequency that a given conformation is found in reiterated docking experiments, a difference may be seen, as shown in Figure 6.4. When these compounds are docked multiple times, AutoDock finds a consistent conformation for 5'-AMP in many docking experiments, whereas 3'-AMP adopts many different conformations and the low energy conformations of 3'-AMP are only found in a small fraction of docking experiments.

We have observed this many times in other systems: if a given molecule shows a consistent conformation in many docking simulations, we have far more confidence in the result. Our current hypothesis is that the frequency of finding a given conformation is providing information on the energy landscape of binding, and that a high frequency is a measure of favorable entropy in the binding process. Recent work has shown that the energetic contribution of this configurational entropy will be high. A study by Chang, Chen, and Gilson[131] has estimated that the configurational entropy of binding of amprenavir to HIV-1 protease is 26.4 kcal/mol, of which 1.8 kcal/mol is due to the loss of conformational entropy when the molecule moves from freely flexible in solution to its constrained position in the active site, and the bulk of the penalty is due to loss of vibrational entropy in the restrictive binding site.

Ideally, we would like to quantify the binding energy of the entire range of conformations available to the ligand and protein, and use this explicitly to evaluate the conformational entropy. However, these types of calculations, such as the Mining Minima calculation employed by Chang, Chen and Gilson, are too computationally expensive for typical docking studies. Instead, several laboratories are exploring methods for using information from the docking simulation or from inexpensive approximations of the range of conformations to evaluate this entropic component.[42–44, 132] Many of these methods perform multiple docking experiments, cluster the resulting conformations by similarity, and then use a measure of the cluster size to estimate the conformational entropy. The assumption is that the docking protocol is providing information on the characteristics of the local energy landscape, and that large clusters of conformations are indicative of favorable entropic characteristics of this landscape.

In this report, we evaluate several methods for their ability to predict correctly the ex-

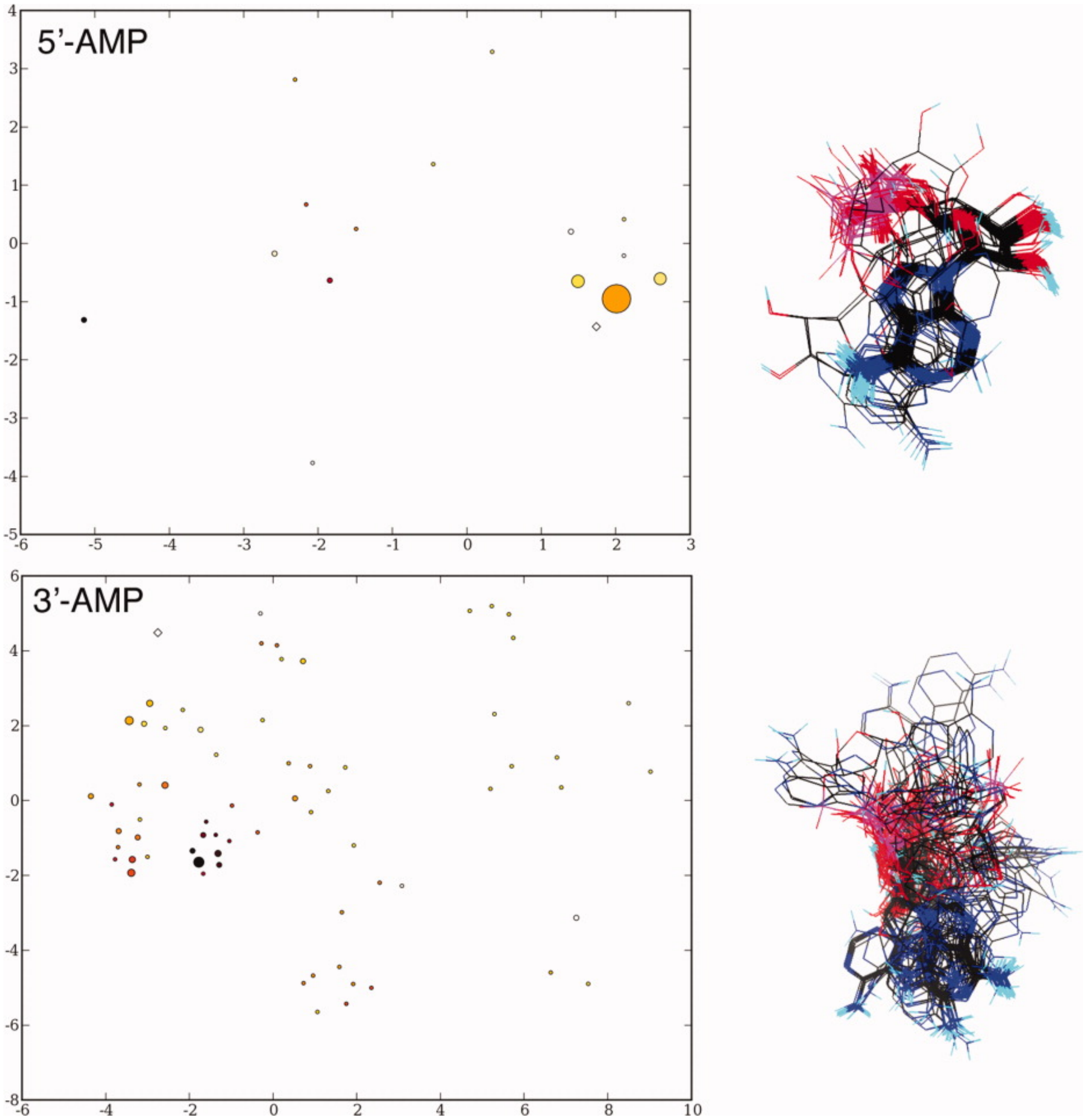


Figure 6.4: Clusters analysis of docking for 5’-AMP and 3’-AMP. The graphs on the left use Sammon mapping to preserve the approximate separation in conformational space between clusters. Each circle represents a cluster of conformations within 2 Å RMSD of each other, and the size of the circle is proportional to the number of conformations in the cluster. The expected bound conformation is shown with a diamond. The images on the right show all of the docked conformations. 5’-AMP binds tightly, and many of the docked conformations cluster into one large group at the expected conformation. 3’-AMP, however, binds weakly and shows a wide scattering of small clusters.

pected bound conformations of nucleotide analogues in APS reductase (Caroll, K.S. Manuscript in preparation).[133, 134] All of these methods seek to characterize the local energy landscape and use this information to estimate an entropic contribution to the binding free energy. The methods perform a sparse sampling of the landscape by reiterated docking or random sampling, making the implicit assumption that the sampled points will represent the features of the entire local landscape. We have found that APS reductase is an excellent test for these methods because experimental binding constants are available for a series of compounds of similar size and chemical composition, but with a wide range of binding constants. This provides a more critical test set than the typical databases used in most studies, which typically include a diverse collection of ligand-protein complexes, but all are specific, tight-binding complexes.

6.2.2 Methods

Motivation

In the most general case, we seek to evaluate the entropic contribution of binding (ΔS) through use of a conformational integral:[43]

$$\Delta G \quad = \quad T\,ln\big(\frac{\sigma_l\sigma_p}{\sigma_{pl}}\frac{c_0 N_a}{8\pi^2(2\pi)^{ntor}}V_B\big) \tag{6.2}$$

where the conformational integral V_B is:

$$V_B \quad = \quad \int_{\Gamma} exp\big[-(U_{pl}(r,\Omega)-E_{pl})/RT\big]\,dr\,d\Omega$$

In these equations, the σ terms account for any symmetry in the molecules, with values of 1 for asymmetric molecules, c_0 = 1 mol/L N_a is Avogadro's constant, $ntor$ is the number of torsional degrees of freedom in the ligand, $U_{pl}(r,\Omega)$ is the energy of each complex conformation, Γ is the region of integration (typically a small space that includes conformations with similar binding modes), and E_{pl} is the ground energy of the complex in solution. The vectors r and Ω define the 3 translational and the $3+ntor$ rotational motions of each complex.

We test several simple approximations to this integral, based on conformations obtained in reiterated AutoDock docking experiments and by directly sampling the local energy landscape. Our goal is to provide an efficient empirical method for estimating this entropic contribution. We seek to improve the estimation of binding constants by rescoring trial docked conformations, combining this estimated conformational entropy, which is derived from reiterated docking experiments, with predicted enthalpic and desolvation contributions used during the docking simulation of each conformation.

In all of these methods, we begin with a set of conformations obtained from docking simulation or from random sampling, and we assume that these sparse samples may be used to characterize the entire local energy landscape. It is important to keep in mind that the evolutionary search method used in AutoDock, which combines a genetic algorithm with a local search,[39] is not designed to be a uniform (Monte Carlo) sampling process, but instead to be successful at finding extreme (minimum) values of the energy function. Thus, it is not directly giving the information needed to estimate the conformational integral, but may be used to infer properties of the energy landscape and conformational entropies. Note also that: (1) the method is heuristic and stochastic, and thus does not guarantee convergence, so the search must be repeated multiple, statistically independent times, and (2) it generates a history of the search process as a by-product. Both of these properties provide opportunities and limitations for use in estimation of entropic contributions, and help to motivate our random sampling experiments, below.

Cluster Size Method

We have tested two methods of using the cluster size as an estimate of the configurational integral. In these methods, we hypothesize that the probability of finding a conformation in a given cluster is capturing information on the local energy landscape. As mentioned earlier, this hypothesis relies on the properties of Lamarkian genetic algorithm used in AutoDock for searching of conformations, which is a stochastic and heuristic method designed to find extreme minimum values of the complex energy landscape. Our hypothesis is that the docking method is more successful for wide energetic wells, and thus the success of finding a given conformation is proportional to the vibrational entropy.

The first is a probability based on a simple conformation-centered RMSD, which we will refer to as the "RMSD" method. For each conformation i, RMSD values $d_{i,j}$ are calculated over all conformations j not equal to i, and the fraction less than a given threshold d_{max} is evaluated. In this work, we used a threshold of d_{max} = 2Å RMSD.

$$P_i^{RMSD} = \frac{N_{d_{j \neq i} \leq d_{max}}}{N}$$

where the numerator is the number of conformations with RMSD less than the threshold and N is the total number of conformations. The second is a probability based on a distance-weighted RMSD, which we will refer to as the "wRMSD" method:

$$P_i^{wRMSD} = \frac{\sum_{j \neq i} exp(-d_{i,j}^2/2\sigma^2)}{N}$$

where the constant $\sigma = 2$Å. If we assume that the favorable region of conformational space is proportional to these probabilities, then the conformational entropy may be estimated as:

$$\Delta G_i = -W^{RMSD} RT\, ln(P_i^{RMSD})$$

where W^{RMSD} is an empirically-determined weight.

Random Sampling Method

We also estimated a value of the conformational integral based on a random sampling of the local energy landscape around each docked conformation. As noted by one reviewer, this method has much in common with the MINTA[135] and Mining Minima[136] methods. 100,000 conformations were generated with small random displacements from the docked conformation. Translational displacements were chosen from a random distribution with bounds -0.5 to 0.5Å, rotational displacements were generated by picking a random axis and rotating by a random angular displacement with bounds -0.5 to 0.5 rad, and torsional displacements were generated with a random angular displacement with bounds of -0.5 to 0.5 rad.

The conformational integral was calculated as:

$$\tilde{V_B}(r_i, \Omega_i) = \frac{\sum_j exp\big((\Delta E(r_j, \Omega_j) - \Delta E(r_i, \Omega_i))/RT\big)}{N}$$

where the ΔE values are predicted energies from AutoDock and the summation is performed over the N=100,000 samples j around the conformation of minimum energy i. The vibrational contribution to the free energy is then calculated as in eq. 6.2.

Binding Constants for Ligands with APS Reductase Binding constants are available for 22 ligands bound to APS reductase (Table 6.3). Values of K_i were determined under single turnover conditions from the dependence of the observed rate constant (k_{obs}) at a given inhibitor concentration under conditions of subsaturating APS, such that K_i is equal to the K_d.[133,137] Kinetic

Table 6.3: Results of docking. ΔG_{obs}, the experimental free energy of binding; N, the number of docked conformations in the cluster of best energy; ΔG_{AD4}, the predicted free energy of binding from AutoDock; RMSD, the root mean square difference in coordinates between docked conformation and analogous atoms in the crystallographic structure; ntor, the number of torsional degrees of freedom in the molecule; and N, ΔG_{AD4}, and RMSD are provided for the cluster of best energy and the largest cluster.

		Best energy			Largest cluster			
	ΔG_{obs}	N	ΔG_{AD4}	RMSD	N	ΔG_{AD4}	RMSD	ntor
5'AMP	-8.07	2	-8.73	4.04	61	-7.96	0.81	6
7deazaAMP	-7.51	1	-8.36	3.46	77	-8.14	0.81	6
5'ADP	-7.29	3	-10.07	3.03	41	-9.98	0.78	8
3'deoxyAMP	-7.21	3	-8.29	3.13	81	-8.29	0.81	5
5'PMP	-6.30	1	-8.37	3.33	60	-7.73	0.90	6
NmethylAMP	-5.97	45	-8.24	0.82	same			6
8aminoAMP	-4.95	50	-8.29	1.63	same			6
2aminoAMP	-4.76	2	-9.28	3.99	20	-8.19	0.96	6
3'phosphoAMP	-4.76	4	-9.07	3.21	same			7
2methoxyAMP	-4.57	2	-8.51	3.56	17	-8.13	1.31	6
bmethAPS	-4.22	57	-9.34	0.81	same			8
2'deoxyAMP	-4.13	1	-8.75	4.10	31	-7.24	1.10	5
adenosine	-3.93	5	-5.58	3.64	27	-4.44	0.69	5
dimethylAMP	-3.90	13	-8.15	3.03	same			6
5'IMP	-3.44	2	-8.60	3.12	23	-7.26	1.48	7
3'deoxyadenosine	-3.17	1	-5.48	4.74	99	-5.27	0.61	4
5'phosphoribose	-2.73	2	-6.93	3.62	7	-6.06	1.93	5
3'AMP	-2.27	8	-9.25	3.87	same			6
2'deoxyadenosine	-2.00	4	-5.90	4.78	13	-4.90	3.17	4
ribose	-1.77	2	-3.65	9.99	56	-4.29	1.73	4
adenine	-1.76	21	-4.13	2.81	23	-3.81	1.49	0
5'IDP	-1.54	1	-9.9	3.90	13	-9.00	0.90	9

data were nonlinear-least squares fit to a model of competitive inhibition. Each K_d reflects the average of at least two independent experiments, and the standard deviation was less than 10% of the value of the mean. The synthesis, characterization and biochemical analysis of the analogues used in this computational study will be reported elsewhere (Caroll, K.S. Manuscript in preparation).

Docking with AutoDock4 Docked conformations and predicted free energies of association were obtained for 22 nucleotide analogues using AutoDock4 (`http://autodock.scripps.edu`). Coordinates for APS reductase were obtained from C. David Stout prior to release–they are identical with subunit B in entry 2goy at the Protein Data Bank.[134] Coordinates for the enzyme were processed in AutoDockTools by adding all hydrogens, assigning charges with the Gasteiger method,[40, 138] and merging non-polar hydrogen atoms. Coordinates for the nucleotides were constructed in InsightII starting with the conformation of the APS nucleotide bound at subunit B in the crystallographic structure. Charges were assigned in ADT and non-polar hydrogen atoms merged. Docking experiments were then performed in AutoDock4 using the default docking parameters, with 2,500,000 energy evaluations for each docking experiment and finding 100 separate docked conformations for each nucleotide.

A test of the role of sugar conformation in the nucleotide was performed using 5'-ADP conformations from entries 1e19, 1m7g, 1o0h and 1rdq from the Protein Data Bank (`http://www.pdb.org`), which were judged to have different sugar conformations based on the distance between C5' and N9, and the torsion angle through atoms C5'-C4'-C1'-N9. These ADP coordinates were prepared and docked similarly to the other nucleotides.

Since crystallographic results are only available for the ligand APS, RMSD values were calculated based on the distance between the nucleotide atoms and the modeled nucleotide, which was created to overlap the analogous atoms in the crystallographic conformation of APS. Thus, the RMSD values in this paper refer to the similarity of the binding modes to the observed mode of APS.

Calibration of Empirical Terms Linear regressions and statistical analysis were performed using the free software R, forcing the regression to include the origin in all cases.

6.2.3 Results

Docking of Nucleotides to APS Reductase

For each of the 22 nucleotides, we performed 100 docking experiments, and clustered the resulting conformations using a 2Å threshold. The results, shown in Table 6.3, are typical of results of AutoDock docking experiments. In 3/22 compounds, the conformation with best energy was in the proper position, but in the remaining 19, they were greater than 2Å RMSD different than the crystallographic position. If, however, we look at the best conformation in the largest cluster, 18/22 conformations are within 2Å of the expected location.

These types of results, which are commonly obtained for AutoDock experiments, are the motivation for the current work. Tight binding ligands, such as 5'-AMP (Figure 6.4a), show excellent clustering and weakly binding ligands, such as 3'-AMP (Figure 6.4b), show poor clustering, although both show similar predicted binding energies. The docking protocol, as revealed in the clustering, is capturing some aspect of the binding energetics that is missing from the current empirical free energy force field.

Table 6.4: Results of regression

			Coeff (t value)	
	Std. error	Multiple R^2	ΔG_{AD4} term	Cluster term
ΔG_{AD4}	1.81	0.86	0.577(11.6)	n/a
ΔG_{AD4}+RMSD	1.74	0.88	0.658(9.6)	1.148(1.6)
ΔG_{AD4}+wRMSD	1.71	0.88	0.658(10.2)	1.098(1.9)
ΔG_{AD4}+Vb	1.69	0.89	1.381(3.4)	1.030(2.0)

Conformational Entropies From Cluster Size

Table 6.4 includes results from regression analysis. Observed binding energies were fit with models that included the predicted AutoDock4 energy and one of the two clustering models: the 2Å threshold model RMSD or the distance weighted model wRMSD. In both cases, modest improvement was seen. The standard error of the predicted binding energy was reduced slightly, and the multiple R-squared increased.

Table 6.5 shows the effectiveness of the cluster size models in rescoring. The first column shows the poor predictive ability of the basic AutoDock4 method: when looking at only the conformation of best energy, only 3/22 identify the proper conformation (these results are also shown in Table 6.3). The second and third columns show the results when the cluster size measure is included. Both methods show excellent predictive ability, ranking the expected conformation as the best in 15/22 or 16/22 cases.

The significance of this result may be estimated by comparison with a statistical method based on Bernoulli trials. We calculated the fraction of dockings with RMSD less than 2Å for each compound, which ranges from 0.00 for 3'-phospho-5'-AMP to 0.99 for 3'-deoxyadenosine. Using these fractions, we can estimate the expected number of correct conformations we would obtain by randomly choosing a conformation for each compound. This analysis estimates that random choice would give a correct answer in 12.35 cases, with a standard deviation of 1.72, out of the 22 compounds.

Conformational Entropies from the Local Conformational Integral

Ideally, we could like to be able to start with a single docked conformation and, by analyzing the local energy landscape, evaluate this entropic contribution to the binding strength. As a first step towards this goal, we have randomly sampled the conformational space around each docked conformation and calculated a conformational integral based on the energy landscape. This is partially effective for improving the prediction of free energies and in reranking. The regression showed a small improvement in the standard error, and the method was able to

Table 6.5: Results of rescoring. RMSD values are given for the docked conformation of best energy as determined by each method, with values > 2.00Å in bold. The final line gives the number of conformations in each column with RMSD < 2.00Å.

	AD4	wRMSD	RMSD	Fit V_b	Lowest RMSD
5'AMP	**4.04**	0.83	0.81	0.95	0.78
7deazaAMP	**3.46**	0.80	0.81	0.85	0.77
5'ADP	**3.03**	0.78	0.78	**2.72**	0.69
3'deoxyAMP	**3.13**	0.90	0.90	0.81	0.77
5'PMP	**3.33**	0.89	0.89	0.93	0.81
NmethylAMP	0.82	0.84	0.82	**2.42**	0.78
8aminoAMP	1.63	1.63	1.63	0.80	1.52
2aminoAMP	**3.99**	0.96	0.97	**3.87**	0.81
3'phosphoAMP	**3.21**	**3.21**	**3.20**	**3.08**	**2.76**
2methoxyAMP	**3.56**	1.31	1.31	**3.89**	0.81
bmethAPS	0.81	0.81	0.81	**2.61**	0.63
2'deoxyAMP	**4.10**	1.10	1.39	**2.72**	0.94
adenosine	**3.64**	0.68	0.69	**4.81**	0.59
dimethylAMP	**3.03**	**2.69**	**3.22**	**2.71**	1.35
5'IMP	**3.12**	**2.83**	1.48	0.88	0.99
3'deoxyadenosine	**4.74**	0.59	0.61	**5.16**	0.56
5'phosphoribose	**3.62**	**3.78**	**3.78**	0.93	1.75
3'AMP	**3.87**	**3.82**	**3.87**	**3.96**	**3.02**
2'deoxyadenosine	**4.78**	**4.78**	**4.78**	1.79	1.12
ribose	**10.62**	1.72	1.73	1.55	1.55
adenine	**2.81**	**2.81**	**2.81**	1.58	1.48
5'IDP	**3.90**	0.98	0.98	2.56	0.81
	3	15	16	10	20/22

rank 10/22 compounds.

Comparing two of the compounds from this study, we can see how these conformational integrals capture the underlying landscape. 5'-AMP and 3'-AMP have the same number and type of atoms and the same number of torsional degrees of freedom, but widely different experimental binding constants. In docking, 5'-AMP gives a tight cluster of 61/100 docked conformations in the expected location, whereas weaker-binding 3'-AMP shows a scatter of different, small cluster conformations.

Looking at the energy landscape around the docked conformation, as shown in Figures 6.5 and 6.6, we find that 5'-AMP has a broader energy well than 3'-AMP. Thus, small motions of 3'-AMP will run up against large steric contacts, whereas small motions of 5'-AMP do not encounter bad contacts.

Unfortunately, these types of correlations were difficult to extract for other compounds, where the structural similarity was not as great. Looking at the entire set, the greatest trend was a strong correlation between the value of the conformational integral and the number of torsional degrees of freedom in the molecule. This is not a surprise, since this merely reflects the magnitude of the entropy involved in freezing these torsional degrees of freedom into a confined space of the active site. The more subtle effect of the local shape of that active site, as seen in the 5'-AMP vs. 3'-AMP landscapes, is overshadowed by this larger effect.

6.2.4 Discussion

The ultimate goal of this work is to find a computationally tractable method to evaluate the conformational entropy of binding, and thus improve our predicted binding energies. This is essential for the future success of docking in computer-aided drug design, where the common presence of false positives and false negatives during virtual screening is a major problem in current studies.

The results presented here suggest that the cluster size is an effective and cheap method for evaluating these entropies, and may be used to improve both the ranking of different complexes, and for the identification of proper binding modes within a single complex. These cluster size methods, however, are not satisfying from a conceptual level, since they are relying on some unknown combination of the overall energy landscape and the details of the docking protocol. Ideally, we would like to develop a method that analyses the energy landscape, both locally and globally, and uses that information to identify the major binding modes and affinities.

Our attempt to characterize the local energy landscape through random sampling has

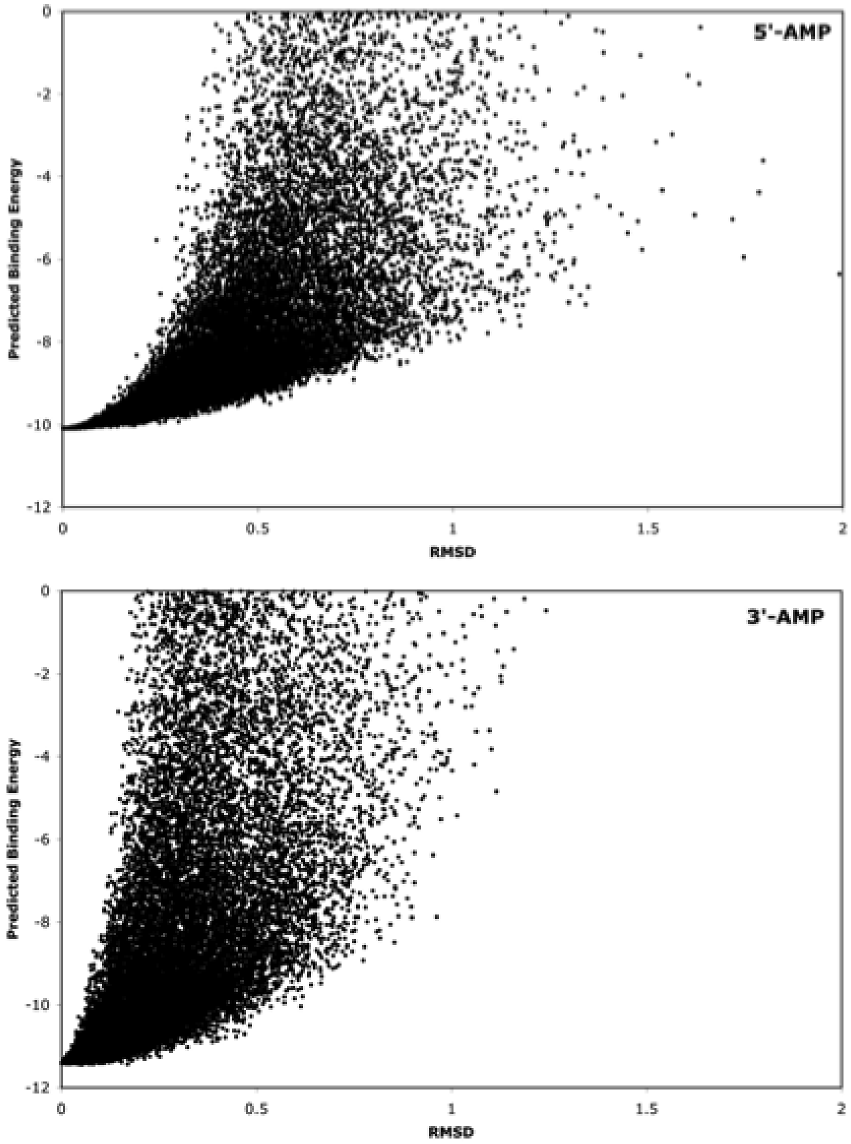


Figure 6.5: Analysis of the local energy landscape. Each point represents a small random change in conformation away from the most favorable bound conformation. RMSD values are calculated between the perturbed conformation and the starting conformation. 5'-AMP shows a wide basin, with very few unfavorable conformations until they are a distance of about 0.5 Å RMSD from the bound conformation. 3'-AMP shows a narrower basin, with many unfavorable conformations as distances less than 0.25 Å RMSD.

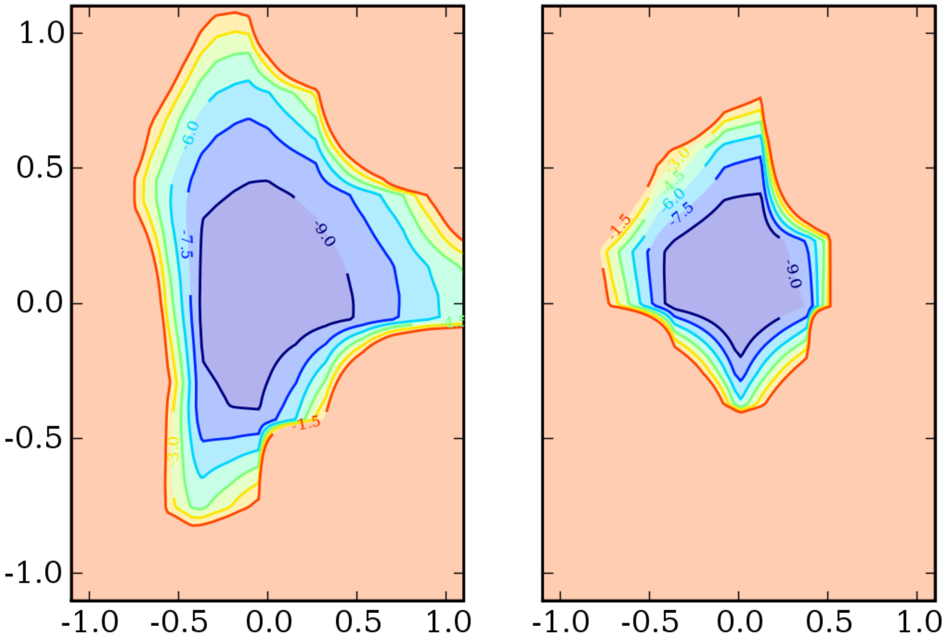


Figure 6.6: Analysis of the local translational energy landscape for 5'-AMP (left) and 3'-AMP(right). Conformations were sampled in the range of -1Å to +1Å in the x and y directions around the most favorable bound conformation. The energy of the sampled conformations is shown here, with the outer contour at -1.5 kcal/mol and additional contours at -1.5 kcal/mol increments.

provided some provocative, but not definitive, results. The results presented in Figure 6.5 show that there are significant differences in the local energy landscape for two forms of AMP, differences that correlate strongly with the large difference in binding constants between these two compounds. However, this principle did not generalize over the entire set. Our current hypothesis is that the docking analysis, and thus the clustering, is capturing information over a larger area of conformational space that we sampled in this work, and that sampling of this larger space will be necessary to develop an effective method for directly evaluating the conformational entropy contribution to binding. However, use of the cluster size in multiple docking experiments is a fast and easy way to estimate this contribution, and is a viable method for improving current docking results.

6.3 HIV protease inhibitor screening

Decades of AIDS research have resulted in the development of a number of anti-HIV drugs effective against several viral proteins. HIV protease has historically been one of the main pharmaceutical targets, and there are currently 10 FDA-approved protease inhibitors. Structure-based drug design has played an important role in the development of protease inhibitors,[5] and this knowledge has been applied in high-throughput screening (HTS) and virtual screening projects involving HIV protease.[35,139]

However, HTS experiments are often very costly, and reagents alone can cost tens of thousands of dollars.[140] Further, hit rates from HTS are generally low. For example, the HTS screen described by by Doman et al. reported a hit rate of 0.021% for inhibitors of protein tyrosine phosphatase-1B.[34] Furthermore, apparent hits in HTS experiments are often found to be promiscuous inhibitors.[141] Virtual screening, though normally less expensive than HTS, suffers from high false positive rates.[37,119] The individual weaknesses of each approach have driven some groups to successfully combine high-throughput and virtual screening for drug discovery.[34,118,142]

To improve screening for anti-HIV compounds, we established a three-step strategy. Identified anti-HIV compounds from the National Cancer Institute's (NCI) cell-based AIDS Antiviral Screen[¶] of small molecules were virtually screened for protease binding, thereby allowing selected compounds to be further evaluated using protease inhibitor assays. The NCI cell-based screen measured the ability of over 40,000 compounds to protect human cells against HIV infection via a chromogenic assay.[143] HIV protease is only one of the many possible targets for

[¶] http://dtp.nci.nih.gov/docs/aids/aids_screen.html the roughly 1,500 compounds found to protect cells from viral-induced death. In a cell-based screen the compounds could be affecting viral proteins or host factors involved in HIV infection. However, the identified compounds are also likely to contain a higher proportion of HIV protease inhibitors than found in a library of random compounds. As an additional benefit, any verified protease inhibitors would ostensibly have low cytotoxicity and be absorbed into human cells, allowing faster compound optimization.

By applying virtual screening to the subset of the anti-viral compounds obtained from cell-based screening putative protease inhibitors can be identified and then directly tested for protease inhibition. For virtual screening, we used the protein-ligand docking program AutoDock[39] to estimate binding energies with respect to particular locations on protease. Current FDA-approved HIV protease inhibitors all target the enzyme's active site. Although novel active site inhibitors will continue to provide useful leads, other sites can be considered as well.[10] The current virtual screening study therefore considers the active site and also an alternative protease binding site, termed the "exo-site."

Both the active site and exo-site, as shown in Figure 6.7, act as potential drug interaction sites for all of the compounds obtained from the NCI. Molecular dynamics simulations have shown that restricting protein flexibility in the exo-site region can affect enzyme function, suggesting this site may serve as a drug target.[11] In evaluating candidate compounds for testing, both sites are considered and criteria beyond predicted binding energy are used. Choosing compounds based on binding energy alone can be problematic, as grid-based energy calculations are often biased by ligand size.[144] Therefore, all tested compounds satisfy criteria based on: (1) predicted binding energy, (2) conformational clustering, and (3) differences between active site and exo-site binding.

Use of these criteria eliminates more than 90% of the initial pool of docked compounds. By narrowing the field, the final compounds can be ordered and tested against HIV protease in biochemical assays, which represents the final phase of our process. Given the typical hit rates from HTS of a random compound library, even a single inhibitor would represent a modest success, but our final results indicate multiple hits for each binding site. This finding demonstrates that a virtual screen for inhibitors of specific viral activity, such as protease, from compounds obtained from a cell-based screening for non-specific anti-viral activity, can narrow the pool of inhibitor candidates to the point where only a handful of compounds must be tested against a specific anti-viral target. For HIV researchers investigating novel drug targets, this strategy should be directly applicable to the discovery of lead compounds.

6.3.1 Methods

Anti-HIV inhibitor library

The NCI AIDS Antiviral Screen contains information for 43,850 compounds. Of these, 617 were determined to be highly active against HIV and 1,195 were determined to be moderately active. 2D chemical structures for 1,585 total active compounds were successfully retrieved from the NCI. Using Marvin 4.1.6 (ChemAxon, `http://www.chemaxon.com`), hydrogen atoms and 3D coordinates were assigned. After eliminating 109 molecules with exotic atom types[‖], 1,476 compounds remained for computational molecular docking. Gasteiger charges were assigned to all atoms and rotatable bonds were assigned using AutoDockTools.

Docking protocol

The 2BPW protein structure from the Protein Data Bank[85] was used for all dockings in this study, as it was previously found to be representative of wild-type HIV proteases.[139] All water molecules were removed from the structure. Polar hydrogens were added and Kollman charges were assigned to all atoms. Affinity grids encompassing the two docking sites were calculated with 0.375 Angstrom spacing using AutoGrid 4.00. Grids coordinates and sizes are set as shown in Figure 6.7.

AutoDock 4.00[39] was used to evaluate ligand binding energies over the conformational search space using the Lamarckian genetic algorithm. Default docking parameters were used with the following exceptions: sw_rho, 0.5; sw_lb_rho, 0.005; ga_num_evals, 1500000; ga_run, 100; rmstol, 1.0.

Determination of inhibitory concentration values

All reactions were run in 100 μl total volume in 96-well microtiter plates, with buffer containing 50 mM MES (pH = 5.5), 200 mM NaCl, 1 mM DTT, 0.0002% Triton X-100, and 5% glycerol. Wild-type protease concentration was 25 nM with initial substrate concentration at 30 μM, which approximates the Km under these conditions.

HIV protease activity was measured with a fluorogenic hexapeptide substrate (Abz-Thr-Ile-Nle-p-nitro-Phe-Gln-Arg-NH2) using an FLX-800 Microplate Fluorescence Reader (Bio-Tek Instruments, Inc., Winooski, VT). Changes in fluorescence were measured over 15 minutes at 37°C, with 340/30 nm excitation and 420/50 nm emission filters. Initial reaction rates were de-

[‖] Only the following atoms types were allowed: H, C, N, O, F, P, S, Cl, Zn, Br, and I.

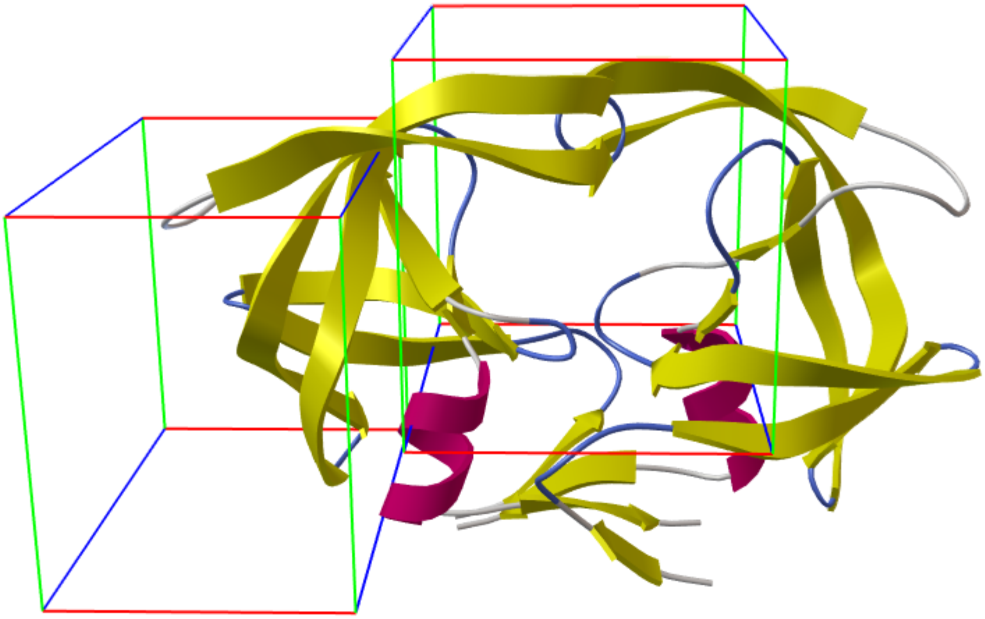


Figure 6.7: Location of docking grids on HIV protease structure. The right box encompasses the enzyme's active site, while the left box corresponds to the exo-site (alternate binding site).

termined by linear regression of the initial 2 minutes of the reaction using KC4 (Bio-Tek). IC_{50} values were determined by non-linear regression using the initial reaction rates versus inhibitor concentration with Prism 4.0c for Macintosh (Graphpad Software, San Diego, CA).

Promiscuous inhibition assays

The α-chymotrypsin, β-galactosidase, and horseradish peroxidase enzymes were used to test for promiscuous inhibition. All reactions were run in 100 μl total volume in 96-well microtiter plates. For the α-chymotrypsin and -galactosidase assays, reactions were performed as previously reported by McGovern et al.,[145] with the exception of enzyme concentration. Instead, the enzymes were diluted based on activity units: 0.5 units/ml for α-chymotrypsin and 10 units/ml for β-galactosidase. The chromogenic α-chymotrypsin substrate, succinyl-ala-ala-pro-phe-p-nitro-anilide, was used at 200 μM, while ONPG, the β-galactosidase substrate, was used at 1 mM.

For horseradish peroxidase, reactions were run in a buffer containing 50 mM citrate and 100 mM sodium phosphate at pH 5.0. A chromogenic substrate, o-phenylenediamine dihydrochloride, was used at an initial concentration of 1.125 mM. Each reaction contained 5 nM horseradish peroxidase and 0.03% hydrogen peroxide. A uQuant Microplate Spectrophotometer (Bio-Tek) was used to measure horseradish peroxidase activity, as a function of OD_{450}. Changes in color were measured for 5 minutes at room temperature.

6.3.2 Results

Docking

Molecular docking via AutoDock 4 was applied to narrow the pool to less than 100 candidates. A straightforward approach could simply rank the compounds according to predicted binding energy. However, we elaborate this basic ordering in three ways:

1. Previous virtual screening work involving HIV protease has shown that an energy threshold of -7.0 kcal/mol is sufficient to separate known protease inhibitors from many randomly selected compounds.[139]

2. Analysis of the entropic properties of protein-ligand docking have shown links between conformational clustering and binding energy.[43, 146] When particular solutions are found repeatedly, the underlying energy landscape may be guiding the search, indicating greater entropic favorability, and thus binding energy. Protein-ligand docking programs typically

Table 6.6: Virtual screen selection criteria and summary of experimental results.

	Active site	Exo-site
Met ΔG threshold	996	707
Met clustering threshold	196	135
Met ΔG difference threshold	75	48
Compounds tested	27	9
Number of hits	3	2

generate an ensemble of results, which can be clustered to determine the degree of convergence in a solution. In this case, each docking incorporates conformations from 100 independent runs with an RMSD tolerance of 1 Å. In the experiments reported here, a candidate compound must have a cluster containing at least 10% of the runs.

3. Finally, it is common to see compounds displaying non-specific activity across a range of compounds, even in virtual screens.[36] To address this problem, the use of two distinct docking sites can be exploited. Since a specific inhibitor is unlikely to bind strongly to two disparate sites, compounds with similar predicted binding affinity at both sites can be eliminated from consideration. Due to differences in the properties of the active and exo-sites, predicted binding energy tends to be more favorable in the active site, leading to different thresholds for putative active site and exo-site binders. All putative active site binders considered for biochemical testing had $\Delta G_{active} - \Delta G_{exo}$ < -2.5 kcal/mol, while all putative exo-site binders had $\Delta G_{active} - \Delta G_{exo}$ > 1.5 kcal/mol.

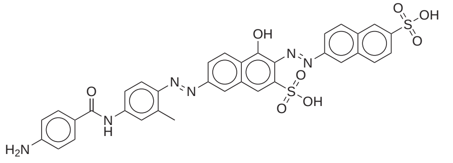
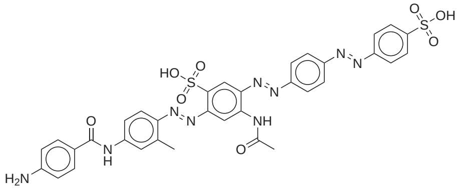
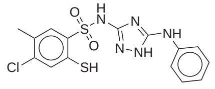
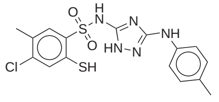
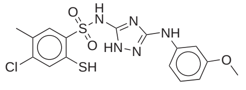
Together, these thresholds eliminated the majority of the 1,476 compounds from consideration (see Table 6.6).

75 putative active site binders remained, as well as 34 putative exo-site binders. Since the exo-site has not been validated as a drug target *in vitro*, the choice was made to focus primarily on compounds affecting the active site. After further eliminating several macrocyclic and high molecular weight compounds, 71 putative active site binders and 9 putative exo-site binders were ordered from the NCI. However, only 36 of the 80 compounds ordered were available and directly evaluated for their ability to disrupt protease activity in protease-substrate activity assays.

Inhibition of HIV protease

All 36 compounds received from the NCI were evaluated against HIV protease at concentrations of 1, 5, and 25 μM to estimate protease inhibition as discussed in the Methods. For

Table 6.7: Verified hits from HIV protease inhibitor screen. NSC 45621 and 79594 are predicted to bind in the exo-site, while the others are predicted to bind in the active site. 95% confidence intervals resulting from non-linear regression of triplicate experiments are shown in parentheses.

NCI ID	Chemical structure	IC_{50}
NSC 45621		733 nM (684 - 785 nM)
NSC 79594		695 nM (633 - 763 nM)
NSC 661073		1.47 μM (1.32 - 1.65 μM)
NSC 666714		2.14 μM (1.87 - 2.45 μM)
NSC 666717		962 nM (0.821- 1.13 μM)

the purposes of this study, compounds with an IC_{50} less than 25 μM are considered potential protease inhibitors. Though FDA-approved protease inhibitors are effective at low nanomolar concentrations, a low micromolar goal is more realistic for unoptimized compounds. Five compounds showed significant inhibition in this range, and their structures are shown in Table 6.7.

Further tests with a wider concentration range were used to determine IC_{50} values, as shown in Figure 6.8. The three putative active site inhibitors are nearly identical in structure, differing only by functional groups in a phenyl ring. These differences appear to have minor effects on protease inhibition, as the IC_{50} values for these three compounds vary significantly, though all are in the low micromolar range. Structures of the two putative exo-site inhibitors are quite similar to each other as well.

Current HTS experiments often discover promiscuous inhibitors, which inhibit through concentration-based aggregation, rather than more desirable specific inhibitors.[141] To assess promiscuous inhibition, the five compounds found to be effective against HIV protease were tested with increased levels of detergent (0.01% Triton X-100) and with protease concentration increased to 50 nM. In both cases, no significant changes in inhibition were detected (data not

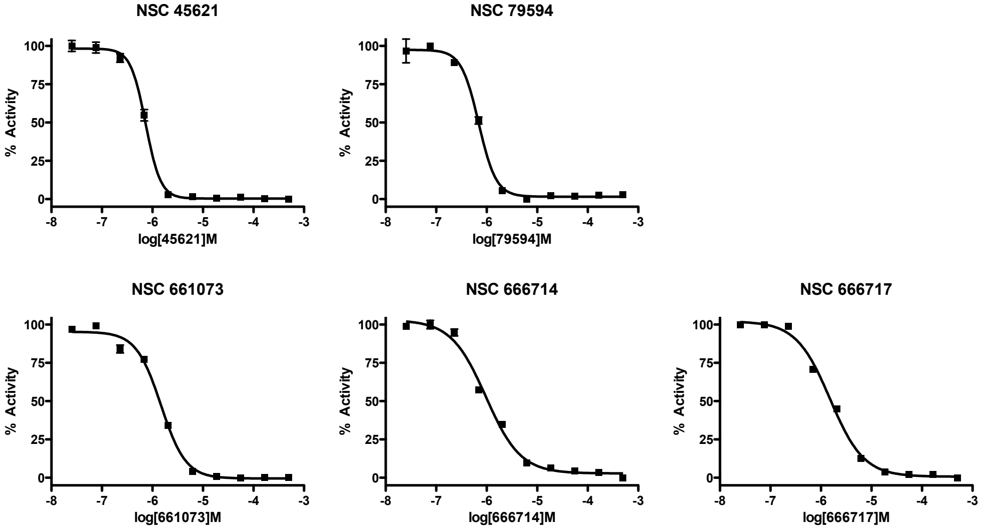


Figure 6.8: Dose-response curves for compounds exhibiting low micromolar or sub-micromolar inhibition of HIV protease. Curves for the putative exo-site binders are shown at the top, the bottom three curves correspond to putative active site binders.

shown). Also, as shown in Figure 6.8, the dose-response curves were not especially steep. All of these observations are consistent with non-promiscuous inhibition.

Additionally, these five compounds were tested against horseradish peroxidase, α-chymotrypsin, and β-galactosidase in chromogenic assays. With compound concentrations of up to 500 μM, no inhibition of horseradish peroxidase or β-galactosidase was detected for any of the five protease inhibitors. Some inhibition of α-chymotrypsin was noted for all 5 compounds, with apparent IC_{50} values between 50 and 500 μM.

Potential allosteric inhibition

The predicted binding mode of NSC 45621 is shown in Figure 6.9. In this conformation, the compound fits into a long groove along the side of the protease, and is likely to interfere with the protein's mobility. The predicted binding mode of NSC 79594 (not shown) is similar.

If the exo-site acts as an allosteric inhibition site, then compounds binding in this area should cause non-competitive inhibition. As shown in Figure 6.10, both putative exo-site binders display decreasing V_{max} as inhibitor concentration increases, consistent non-competitive inhibition. In addition, the dose-response curves of both NSC 45621 and NSC 79594 show Hill coefficients between 2 and 3. Since HIV protease is symmetric, each dimer contains two exo-sites, allowing two separate binding events. With some level of cooperativity in binding, a Hill coefficient greater than 2 would be expected. However, it must also be noted that non-

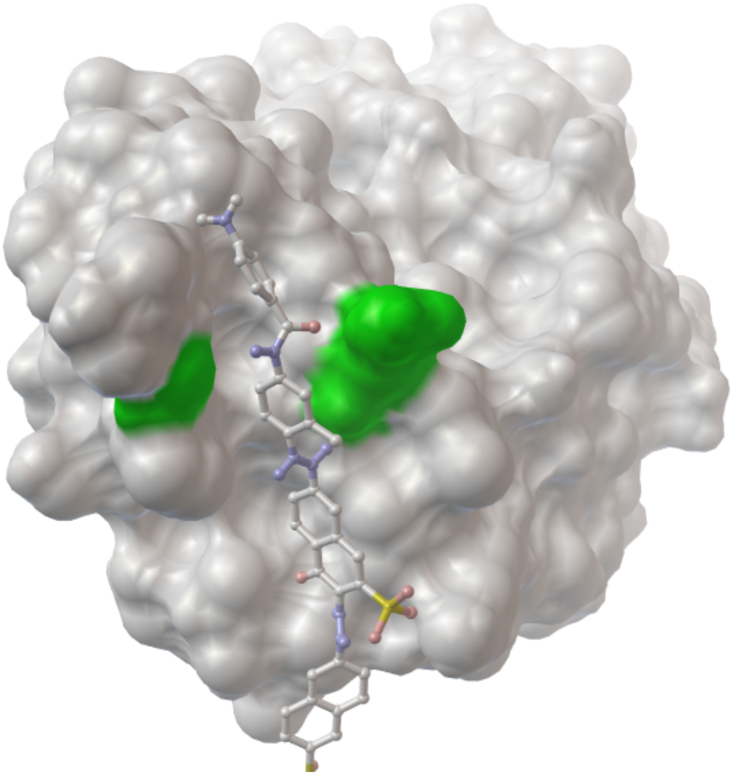


Figure 6.9: Predicted binding mode for NSC 45621. The colored regions of the protease correspond to Gly40 and Gln61, which were constrained in a previous study.[11]

competitive inhibition and high Hill coefficients are associated with promiscuous inhibition.[141]

6.3.3 Discussion

From an existing cell-based screen of anti-HIV compounds, computational molecular docking was used to discover novel inhibitors of HIV protease. By combining high-throughput cell-based and virtual screening, a small number of potential protease inhibitors were identified from the original compound library, allowing direct testing against HIV protease *in vitro*. Of 36 compounds tested, 5 were found to inhibit protease with IC_{50} values all less than 2.5 μM, and as low as 700 nM. While typical HTS experiments yield verified hit rates of less than 0.1%, the hit rate for this approach was 13.9%. While direct comparisons are problematic, this represents an increase of several logs over the HTS screen conducted by Doman et al., which had a hit rate of 0.021%.[34]

Our virtual screening strategy allows specific regions of the protease enzyme to be interrogated for possible drug binding. While FDA-approved protease inhibitors all target the enzyme's active site, the entire chemical library was docked against a putative alternate binding site as well as the active site. This facilitated the use of a novel analysis technique which compared predicted binding energies in each of the sites in order to filter out compounds predicted to bind indiscriminately. The five verified inhibitors are divided between each target, which provides a rich opportunity for further study. Although the mechanism of inhibition has not yet been experimentally verified for all compounds, the docking results suggest the possibility that two of the compounds may be allosteric inhibitors.

Since these compounds inhibited HIV protease in the high nanomolar/low micromolar range, it was possible that promiscuous inhibition played a role. Assays with horseradish peroxidase and β-galactosidase showed no inhibition from any of these compounds, even at concentrations far greater than the IC_{50} values for HIV protease. At high concentrations, the five compounds did inhibit α-chymotrypsin. However, as a serine protease, there may be enough structural similarity between α-chymotrypsin and HIV protease to account for the weak inhibitory effects of the compounds found. Taken together with the horseradish peroxidase and β-galactosidase experiments, though, promiscuous inhibition seems unlikely since inhibition was not seen in all cases, even at concentrations of 500 μM. Moreover, these compounds are meant for use as scaffolds or probes for further rounds of the drug development cycle, and so may be seen as a stepping stone toward higher affinity inhibitors with greater specificity.

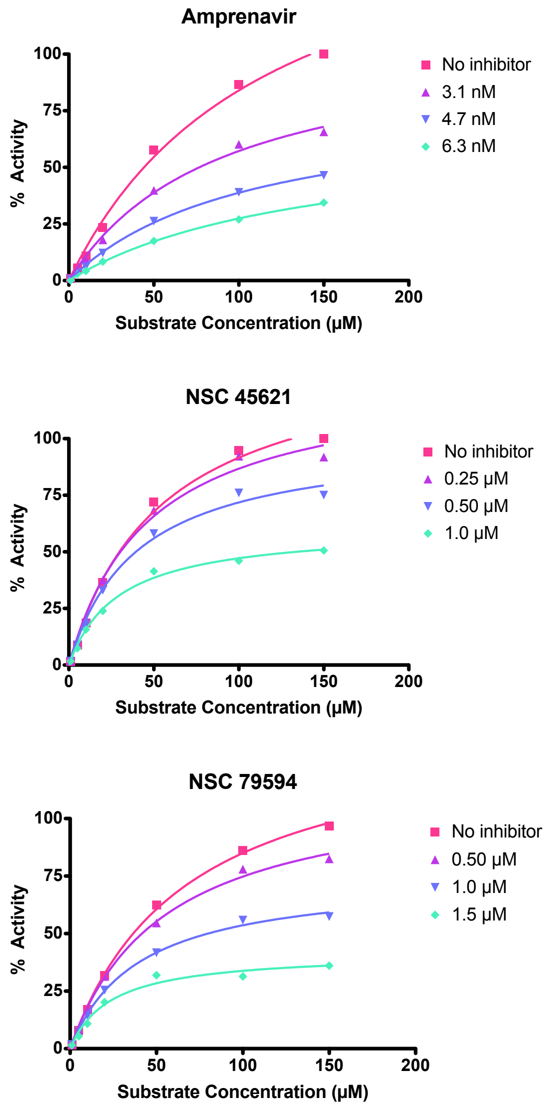


Figure 6.10: Saturation curves for putative exo-site binders and amprenavir (control), using Michaelis-Menten curve fits with increasing inhbitor concentrations.

Optimization of these compounds should benefit due to the source of the chemical library. Because these compounds have already been screened in human cell-based assay, there is reason to believe that they are absorbed into cells and may have low cytotoxicity. With these properties, the lead optimization of identified compounds should be streamlined. Since a modest virtual screen followed by small-scale biochemical verification of activity has led to the discovery of five low micromolar and sub-micromolar protease inhibitors, future drug discovery efforts focused on additional HIV targets could benefit from this strategy. Given similar high-throughput cell-based experimental findings for bacterial and other viral pathogens, our approach may be applicable for identification of starting compounds that target various pathways in these pathogens.

Acknowledgements

Chapter 6.1 is a reprint in full of materials that appeared in *Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries*. Max W. Chang, William Lindstrom, Arthur J. Olson, and Richard K. Belew. *Journal of Chemical Information and Modeling* 2007, **47**(3):1258–1262. I was the primary investigator and author of this paper.

Chapter 6.2 contains material that appeared in *Empirical entropic contributions in computational docking: Evaluation in APS reductase complexes*. Max W. Chang, Richad K. Belew, Kate S. Carroll, Arthur J. Olson, David S. Goodsell. *Journal of Computational Chemistry* 2008, **29**(11):1753–1761. I was the primary researcher of this work, David Goodsell was the primary author of the paper.

Chapter 6.3 is an extended version of a preprint of *Virtual screening for HIV protease inhibitors from a library of antiviral compounds obtained via a cell-based screen*. Max W. Chang, Michael J. Giffin, William M. Lindstrom, Jr., Arthur J. Olson, Richard K. Belew, and Bruce E. Torbett. Submitted to *Journal of Medicinal Chemistry*. I was the primary investigator and author of this work.

Chapter 7

HIVLink

7.1 Introduction

Since the mid-1990s, there has been tremendous growth in the amount of biological data available in online databases. GenBank, the NIH sequence database has grown from roughly 1 million sequences in 1995 to over 80 million in 2008.[147] During this time, there has also been major growth in scientific publishing related to the life sciences. The number of articles indexed at PubMed, the premier biomedical literature database, exceeds 17 million, encompassing more than 5,000 journals. This enormous corpus spans many fields and has benefited from the use of traditional text search methods. However, specific groups of users with common interests may benefit from specialized techniques.

The HIVLink program was designed to serve an audience interested in HIV drug resistance. Although there are nearly 200,000 HIV-related articles in PubMed, researchers may be interested in a particular area, such as clinical practice or drug development. By first restricting the corpus to a specific subset of literature, it becomes possible to incorporate features that are especially useful for this focused body of work. For example, important synonyms can be taken into account while performing a search. Additionally, following retrieval of a large set of articles, useful aggregate statistics can be displayed, such as the frequency distribution of relevant drugs and mutations.

HIVLink also exploits several techniques of interest to a general audience: the inclusion of citation indexing and spreading activation search,[148–152] which is a unique approach to finding related articles. The spreading activation search treats a set of documents as a network, with connections formed by textual similarity and citations. This type of search allows a more expansive notion of similarity in literature, allowing the discovery of related work that would not be found by a traditional search engine. A spreading activation search is also easily used to search by example. In other words, a specified set of known relevant articles can be used to perform a search, where the results are based on citation and overall text similarity to the given articles. HIVLink's graphical user interface also provides a quick display of temporal and citation information.

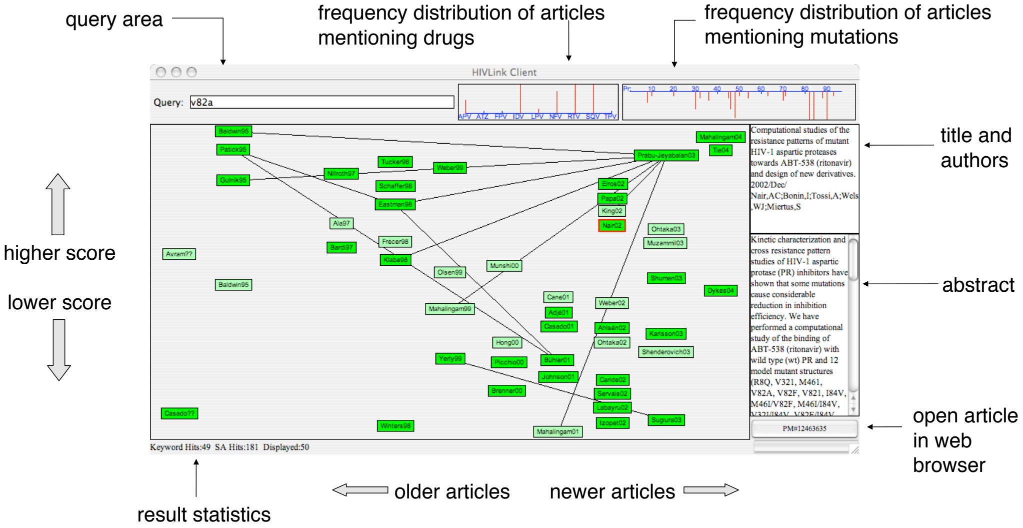
7.2 Interface

The HIVLink application window is shown in Figure 7.1. The main window displays the result of a query, where each article is represented by a rectangular node. Each node is labeled based on the first author's surname and the year of publication (see Figure 7.2a). Darker nodes indicate that the query was matched exactly, while the lighter nodes represent articles that were determined to be relevant by the spreading activation search (see section 7.5 for details). The articles are arranged vertically based on their score, more relevant articles are found toward the top. Horizontal arrangement is based on the year in which the article was published, with newer articles toward the right. Black edges between nodes indicate a citation relation, i.e. the newer article cites the older one.

A number of subpanels display information related to the selected article or the entire set of documents retrieved. The currently selected article's title, authors, and abstract are shown on the right. Pushing a button with the article's PubMed ID will open the selected article's PubMed record. Two bar graphs at the top of the window display aggregate information from all articles shown in the main window. The rightmost graph (Figure 7.2b) shows the prevalence of mutations at specific positions mentioned in the abstracts of the retrieved articles. Clicking on part of the graph will select articles that refer to the appropriate mutations, which can be added to a clipboard or used for further searches. The other graph (Figure 7.2c) shows the number of articles which mention specific protease inhibitors. Similar to the graph of mutation frequencies, clicking on a drug will select the corresponding articles.

7.3 Corpus

The NCBI's PubMed service provides titles, abstracts, and keywords for the articles that it indexes. PubMed contains roughly 185,000 articles related to HIV, as of May, 2008. The current version of HIVLink indexes only PubMed records related to HIV protease, which total



Query: v82a

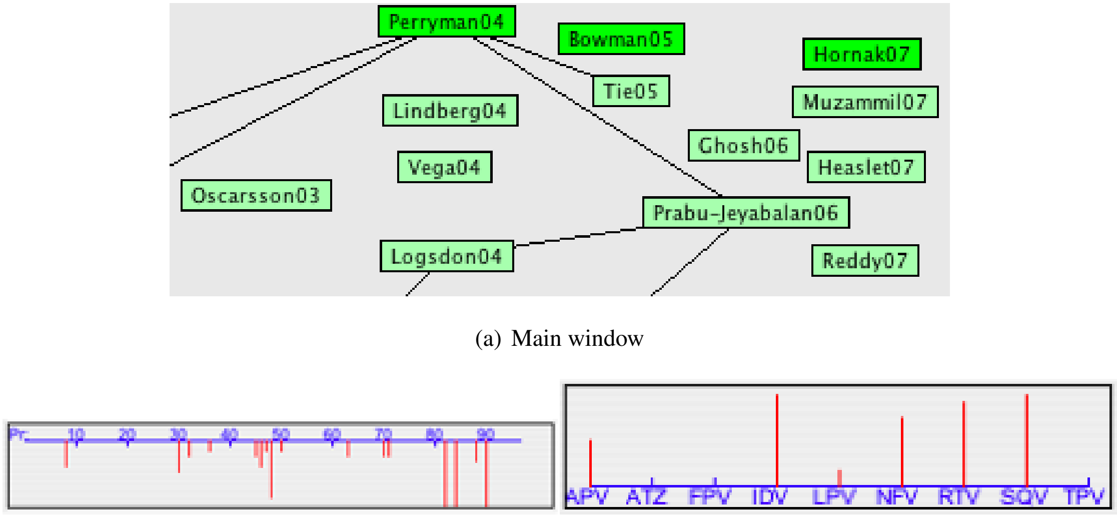
Computational studies of the resistance patterns of mutant HIV-1 aspartic proteases towards ABT-538 (ritonavir) and design of new derivatives.
2002/Dec/
Nair,AC;Bonin,I;Tossi,A;Wels,WJ;Miertus,S

Kinetic characterization and cross resistance pattern studies of HIV-1 aspartic protase (PR) inhibitors have shown that some mutations cause considerable reduction in inhibition efficiency. We have performed a computational study of the binding of ABT-538 (ritonavir) with wild type (wt) PR and 12 model mutant structures (R8Q, V321, M461, V82A, V82F, V821, I84V, M46I/V82F, M46I/I84V, V32I/I84V, V82F/I84V

PM#12463635

Keyword Hits:49 SA Hits:181 Displayed:50

Figure 7.1: The HIVLink application window.



(a) Main window

(b) Mutation frequency

(c) Drug frequency

Figure 7.2: Details of HIVLink's interface.

Table 7.1: HIV drug names and synonyms.

Brand names	Generic name	Abbreviation	Experimental codes
Agenerase	amprenavir	APV	141W94, VX-478
Crixivan	indinavir	IDV	MK-639
Fortovase, Invirase	saquinavir	SQV	Ro-31-8959
Kaletra, Aluvia	lopinavir	LPV	ABT-378
Norvir	ritonavir	RTV	ABT-538
Reyataz	atazanavir	ATZ	BMS-232632
Viracept	nelfinavir	NFV	AG-1343

13,719, using the following query:

> (protease AND (HIV-1 OR HIV OR (human[Text Word] AND immunodeficiency[Text Word]))) OR HIV protease OR HIV protease inhibitors

Across a wide range of search, citations are a highly desirable means of finding related articles,[153] Google's PageRank technique shows how much a keyword search engine can benefit from this type of information.[154] However, citation information is not currently used by PubMed's related article facility. For a limited set of PubMed articles, full text is available from PubMedCentral, which includes citation information. Of the PubMed articles related to HIV protease, approximately 1,000 are present in PubMedCentral and have associated citation information.

7.4 Synonymy

Synonymous relationships among words can make broad searches difficult. HIV drugs can be referred to by brand name, generic name, an abbreviation, or even an experimental code (see Table 7.1). In searching for "amprenavir," for instance, a researcher is also presumably interested in articles where "VX-478" or "APV" is mentioned. In making this assumption, a focused corpus is important, as the context of "APV" is clear when discussing HIV protease and its inhibitors, but not across all biomedical literature. For example, "APV" may also refer to avian polyomavirus.

Similarly, mutations in HIV proteins are often referenced using a shorthand notation. A mutation in HIV protease at position 84 from the wild-type isoleucine to valine is signified as "I84V." However, the wild-type residue is sometimes omitted, and the mutation shown as "84V." As these references are interchangeable, a search engine should treat them as equivalent. These examples are relatively simple, but show that some kinds of synonymy can be addressed with simple tables. The spreading activation search technique, described in the following section, may also be of use, but the general problem remains deep and connected to full natural language processing.

7.5 Spreading activation search

Traditional search engines find documents whose contents match a user's query, based on the presence of keywords. However, a keyword-based may be inadequate for some users' purposes. Other problems, such as changing vocabulary, may also make searches more difficult. For example, in the early 1990s, some articles referred to HIV "proteinase" rather than "protease," which has become the dominant term. Since the context of these articles is similar, the overall level of textual similarity between documents can be used to improve retrieval.

HIVLink uses search engine results from a user's query to "activate" matching articles, taking synonyms into account. These articles propagate activity to other articles with similar text or a citation relationship. This propagation continues for several iterations, in a process known as "spreading activation" (see Figure 7.3). At the end of this process, the results are ordered by activity rather than the keyword-based search. This allows the integration of different sources of evidence that may match the user's interests. Although an article may not match a user's request explicitly, a closely related article can often be relevant.

In some situations, a user may have a set of articles, and wish to find additional related articles. A traditional search would likely involve a combination of keyword-based search and a review of cited works. With HIVLink, a search can be "seeded" using a list of PubMed IDs. The use of spreading activation search involves both text similarity and available citation information, using both as a factor in finding related articles. Furthermore, relationships between multiple articles are taken into account. For instance, when two "seed" articles are linked to another work, activity propagates along both edges, and indicates greater relevance.

7.6 Implementation details

HIVLink is a cross-platform application written in Java. It has been tested successfully on Windows, Linux, Solaris, and OS X platforms. The program is split into server and client applications, which communicate through Java's RMI facility. The GUI is Swing-based, and there is a command line interface as well. Traditional keyword searching is handled by Lucene, a high-performance text search engine.[155]

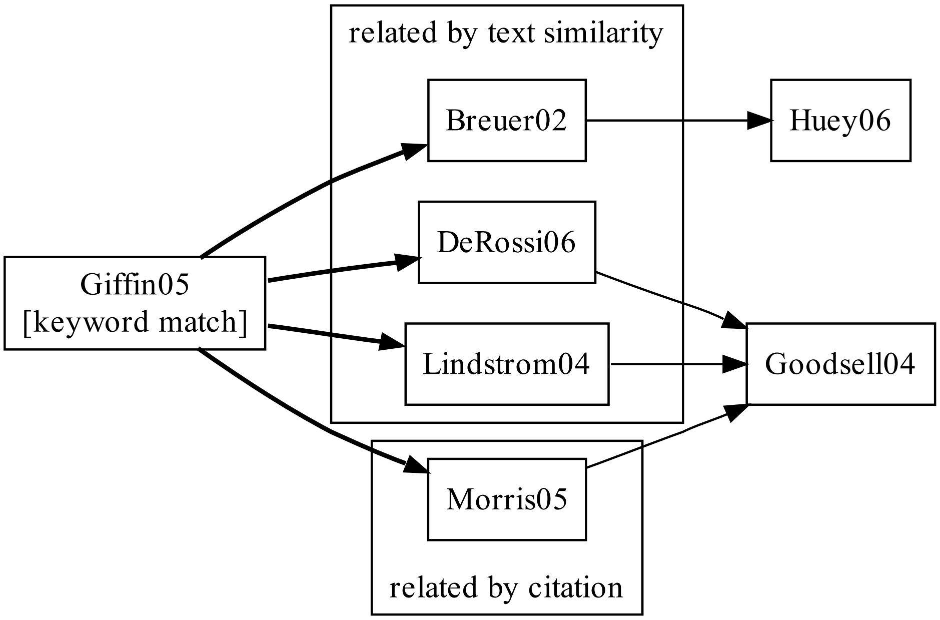


Figure 7.3: An example of associative retrieval by spreading activation search. The initial article with matched keyword(s) will "spread" the search to include articles related by overall text similarity or citation. In a second round, the search will spread further to include articles that are not directly related to the initial match.

7.7 Conclusion

HIVLink is a prototype system designed to explore a new methodology for retrieval system design. It developed out of a close collaboration with biologists actively engaged in the investigation of structural features that emerge with drug resistance. The techniques described here can be extended to support investigation of other major drug classes that are typically part of AIDS therapy. Effort will be devoted to incorporating additional forms of evidence, such as authorship, to augment the spreading activation search. Future versions will benefit from relevance feedback studies derived from observations of user behavior.

Acknowledgements

This chapter contains material that appeared in *Purposeful retrieval: Applying domain insight for topically-focused groups of biologists*. Richard K. Belew and Max W. Chang. In *Search and Discovery in Bioinformatics: SIGIR 2004 Workshop*. Edited by Javed Mostafa and Padmini Srinivasan. 2004. I was the secondary researcher and author of this work.

Chapter 8

Conclusions

This dissertation has focused on two areas related to HIV drug resistance: anticipating evolution of the virus in response to drug therapy and the development of new inhibitors. In modeling the evolution of the virus, a general model of fitness was developed, which was factored into replication capacity and drug resistance components for individual proteins. The protease and reverse transcriptase enzymes remain the main HIV drug targets, and were the main focus of this work. Special emphasis was placed on protease, which has been a major target for structure-based drug design.

In predicting the replication capacity component of viral fitness, several sequence-based machine learning methods were shown to be imprecise with currently available data. As an alternative, a phylogenetic approach examining sequence diversity for HIV protease homologs was able to predict impairments in enzyme function that correlated with experimentally observed catalytic efficiency.

To complement the study of replication capacity, the contribution of individual mutations to drug resistance was measured by structural modeling of protease-ligand interactions. A combination of the structure-based models and estimates of replication capacity were able to predict over half of the major resistance mutations for clinically-approved HIV protease inhibitors. This approach was extended to identify resistance mutations for a novel protease inhibitor, AB2. Protease mutants containing the 47V, 53L, and 84V mutations were constructed using site-directed mutagenesis, then tested for resistance *in vitro*. Each of the mutations conferred resistance against AB2 individually and showed further increases when combined, up to 16-fold increase in IC_{50} for the triple mutant. The same mutations were predicted to have a lesser effect of amprenavir resistance, which was also experimentally verified *in vitro*.

Moving beyond the prediction of specific resistance mutations, the evolution of HIV in response to drug therapy was studied by simulating a viral population. The core of this simulation incorporated a viral fitness function based on predictions of replication capacity and drug resistance. In addition, the epistatic interactions between mutations were modeled using the results of a covariation analysis on viral isolates. The fitness function showed significant accuracy in ranking the relative fitness of various drug resistant protease mutants. Incorporating this function into the simulation framework allowed the evolution of a viral population in the presence of inhibitors. The prevalence of mutations in a simulated population correlated with the clinically observed frequencies for protease inhibitor treatments, showing that the simulation was capturing substantial aspects of viral fitness. Next, the simulation was used to predict the effects of protease inhibitor combination therapy, finding that combinations of clinically-approved inhibitors were no more effective than single inhibitor treatments.

These existing protease inhibitors all share a common target, the enzyme's active site, and exhibit substantial cross-resistance. Use of the simulation allows experiments involving hypothetical inhibitors, such as an allosteric protease inhibitor. Resistance mutations were assumed to arise easily against this putative inhibitor when used individually, limiting its effectiveness as a single inhibitor treatment. However, when combined with clinically-approved protease inhibitors, treatment was highly effective, in contrast to combinations of inhibitors that targeted the active site. Even relatively weak inhibitors that target novel binding sites should serve as excellent complements to existing inhibitors.

In seeking out new inhibitors, models of protein-ligand binding can play an important role. Early studies focused on finding representative protease structures and estimating entropic contributions based on conformational clustering. These findings were applied in a subsequent virtual screen for inhibitors of HIV protease. A key feature of this screen was the use of prescreened compounds with anti-HIV activity, rather than a random library. *In vitro* testing of the best candidates from the virtual screen found five inhibitors with IC_{50} near 1 micromolar. Importantly, two of these inhibitors were predicted to bind outside of the protease active site and displayed properties consistent with allosteric inhibition. Additional tests of this binding mode will be performed in future studies, but the use of a virtual screen was key in targeting this novel site and minimizing the number of compounds tested *in vitro*.

8.1 Future directions

8.1.1 Simulation of HIV infection and drug resistance evolution

The simulation of HIV evolution showed better correlation with protease-based data sets than reverse transcriptase. Further research will address this deficiency, focusing on the prediction of replication capacity. This area would benefit from additional information for data mining, which would be useful in improving regression results. A more general approach involving structural modeling would be helpful, but remains difficult for several reasons. Even for protease, which is well-characterized structurally, the large degree of flexibility in its substrates makes docking studies intractable. Beyond calculating the affinity between protease and the substrates, modeling k_{cat} remains a challenge.

Extending the simulation to incorporate multiple compartments would be a useful step in modeling infection. Currently, infection is represented by a single population that can be eradicated with sufficiently high levels of inhibitor. This does not occur *in vivo*, partially due to viral reservoirs in different parts of the body. Each of these locations exhibits different properties, such as drug accessibility and structural organization, which could greatly impact the outcome of different drug treatments.

Another factor not yet explored was the role of time in the simulation. Future work will attempt to connect the rate of evolution *in silico* to the results observed during serial passage experiments. Genetic selection episodes from clinical records also provide temporal information, as well as viral load and CD4 counts. These values could be used as the first steps in extending this simulation to model *in vitro* drug resistance evolution and interaction with the immune system.

8.1.2 Docking and entropy

The entropy-related studies in Chapter 6 represent the first steps in finding a fast method for determining the contribution of vibrational entropy. Using cluster size as an approximation of an entropic contribution was found to improve binding energy predictions and the ranking of docked conformations. The clustering of independent docking runs depends on the docking search process that, in turn, depends on the underlying energy landscape. Initial efforts to perform local sampling of the energy landscape were not able to improve docking results as much as cluster-based methods, but this will continue to be an active avenue of exploration. In the short term, exploring larger conformational spaces may be sufficient to improve the local sampling method and will be useful in further characterizing energy landscapes in low-energy regions.

Beyond utility in improving binding energy estimates, better understanding of protein-ligand energy landscapes has the potential to improve the underlying search mechanics in AutoDock. The current search algorithm incorporates a Solis-Wets local search, which attempts to descend energy gradients using randomly generated steps. Key parameters in this process control the average size of these steps and, indirectly, the size of space explored. In the current version of AutoDock, these parameters are based on global search performance on a small set of test systems.[40] However, enumeration of the energy landscape through an extension of the local sampling procedure provides the opportunity to more directly set the parameters that guide local search. The step size, for instance, should be guided by the average size of local minima. Improvements in the local search process would lead to greater efficiency for the overall docking process.

Entropic considerations may be incorporated more directly into the overall docking process. Currently, AutoDock runs are performed sequentially and independently, with clustering calculated after all searches are complete. With little additional effort, the runs could be executed in parallel, rather than sequentially, and with clustering performed after each generation. A significant level of convergence in clustering would signal early completion, sparing computational effort. Further improvements involving entropy are possible in the search algorithm. While the genetic algorithm component of the AutoDock search procedure is driven through recombination and selection of the most favorable conformations, much information is lost. AutoDock may evaluate the binding energy of millions of conformations, while only reporting on a handful. More detailed tracking of these conformations over the duration of the search process could reveal more global aspects of the energy landscape and further improve estimates of binding energy.

8.1.3 Finding novel HIV inhibitors

The virtual screening process described in Chapter 6.3 was successful in discovering new inhibitors for HIV protease. A key feature of this work was the use of a pre-screened chemical library effective against HIV. As the work in FightAIDS@Home continues to focus on HIV protease, this set of compounds may prove useful in targeting specific mutants. More generally, virtual screens with this library should be useful against other HIV proteins which have known structures. For example, crystal structures are available for portions of the integrase and Nef proteins, making them viable targets for virtual screening. Although inhibitors of various potency are available for both of these proteins, further study may yield novel scaffolds or alternative binding sites.

The putative allosteric protease inhibitors shown in Chapter 6.3 remain subjects of active interest. Attempts are underway to validate the binding modes of these inhibitors through x-ray crystallography and molecular dynamics simulations. Apart from experiments involving these particular inhibitors, more direct studies of the exo-site region are also important in order to verify that impeding structural flexibility of the protein at this site is effective in disrupting protease function. This could be tested by engineering a disulfide bond in the exo-site region through site-directed mutagenesis. Alternatively, a tethered-ligand strategy could be useful in directly targeting this site.[156]

Throughout this process, it will be important to keep in mind the development of drug resistance. In targeting regions of HIV protease outside of the active site, resistance is likely to arise easily because mutations will have lesser effects on viral replication capacity. Allosteric inhibitors of HIV protease will be most useful in augmenting existing protease inhibitor-based treatments.

Bibliography

1. Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, Hahn BH: **Origins and evolution of AIDS viruses: estimating the time-scale**. *Biochem Soc Trans* 2000, **28**(2):275–282.

2. MacNeil A, Sarr AD, Sankalé JL, Meloni ST, Mboup S, Kanki P: **Direct evidence of lower viral replication rates in vivo in human immunodeficiency virus type 2 (HIV-2) infection than in HIV-1 infection**. *J Virol* 2007, **81**(10):5325–5330.

3. Frankel AD, Young JA: **HIV-1: fifteen proteins and an RNA**. *Annu Rev Biochem* 1998, **67**:1–25.

4. Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, Korber B (Eds): *HIV Sequence Compendium 2005*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory 2005.

5. Wlodawer A, Vondrasek J: **Inhibitors of HIV-1 protease: a major success of structure-assisted drug design.** *Annu Rev Biophys Biomol Struct* 1998, **27**:249–284.

6. National Institute of Allergy and Infectious Diseases: **How HIV Causes AIDS** 2004, [http://www.niaid.nih.gov/factsheets/howhiv.htm].

7. Chou KC: **Prediction of human immunodeficiency virus protease cleavage sites in proteins**. *Anal Biochem* 1996, **233**:1–14.

8. de Oliveira T, Engelbrecht S, Janse van Rensburg E, Gordon M, Bishop K, zur Megede J, Barnett SW, Cassol S: **Variability at human immunodeficiency virus type 1 subtype C protease cleavage sites: an indication of viral fitness?** *J Virol* 2003, **77**(17):9422–9430.

9. Miller M, Schneider J, Sathyanarayana BK, Toth MV, Marshall GR, Clawson L, Selk L, Kent SB, Wlodawer A: **Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution**. *Science* 1989, **246**(4934):1149–1152.

10. Hornak V, Simmerling C: **Targeting structural flexibility in HIV-1 protease inhibitor binding.** *Drug Discov Today* 2007, **12**(3-4):132–138.

11. Perryman AL, Lin JH, McCammon JA: **Restrained molecular dynamics simulations of HIV-1 protease: the first step in validating a new target for drug design.** *Biopolymers* 2006, **82**(3):272–284.

12. Lengauer T, Sing T: **Bioinformatics-assisted anti-HIV therapy**. *Nat Rev Microbiol* 2006, **4**(10):790–797.

13. Petropoulos CJ, Parkin NT, Limoli KL, Lie YS, Wrin T, Huang W, Tian H, Smith D, Winslow GA, Capon DJ, Whitcomb JM: **A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1**. *Antimicrob Agents Chemother* 2000, **44**(4):920–928.

14. Nijhuis M, Schuurman R, de Jong D, Erickson J, Gustchina E, Albert J, Schipper P, Gulnik S, Boucher CA: **Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy**. *AIDS* 1999, **13**(17):2349–2359.

15. Mammano F, Trouplin V, Zennou V, Clavel F: **Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: virus fitness in the absence and in the presence of drug**. *J Virol* 2000, **74**(18):8524–8531.

16. Bühler B, Lin YC, Morris G, Olson AJ, Wong CH, Richman DD, Elder JH, Torbett BE: **Viral evolution in response to the broad-based retroviral protease inhibitor TL-3**. *J Virol* 2001, **75**(19):9502–9508.

17. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW: **Human immunodeficiency virus reverse transcriptase and protease sequence database**. *Nucleic Acids Res* 2003, **31**:298–303.

18. Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J: **Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype**. *Proc Natl Acad Sci U S A* 2002, **99**(12):8271–8276.

19. Wang K, Jenwitheesuk E, Samudrala R, Mittler JE: **Simple linear model provides highly accurate genotypic predictions of HIV-1 drug resistance**. *Antivir Ther* 2004, **9**(3):343–352.

20. Rhee SY, Taylor J, Wadhera G, Ben-Hur A, Brutlag DL, Shafer RW: **Genotypic predictors of human immunodeficiency virus type 1 drug resistance**. *Proc Natl Acad Sci U S A* 2006, **103**(46):17355–17360.

21. Jenwitheesuk E, Samudrala R: **Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations**. *BMC Struct Biol* 2003, **3**:2–2.

22. Shenderovich MD, Kagan RM, Heseltine PN, Ramnarayan K: **Structure-based phenotyping predicts HIV-1 protease inhibitor resistance**. *Protein Sci* 2003, **12**(8):1706–1718.

23. Jenwitheesuk E, Samudrala R: **Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach**. *Antivir Ther* 2005, **10**:157–166.

24. Cheng Y, Prusoff W: **Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction.** *Biochem Pharmacol* 1973, **22**(23):3099–108.

25. Copeland R: *Evaluation of enzyme inhibitors in drug discovery: a guide for medicinal chemists and pharmacologists*. Hoboken, NJ: Wiley-Interscience 2005.

26. Nowak MA, May RM, Anderson RM: **The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease**. *AIDS* 1990, **4**(11):1095–1103.

27. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD: **HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time**. *Science* 1996, **271**(5255):1582–1586.

28. Phillips AN: **Reduction of HIV concentration during acute infection: independence from a specific immune response**. *Science* 1996, **271**(5248):497–499.

29. Nowak M, Nowak M: *Virus Dynamics: Mathematical Principles of Immunology and Virology*. Oxford 2001.

30. Dixit NM, Perelson AS: **Complex patterns of viral load decay under antiretroviral therapy: influence of pharmacokinetics and intracellular delay**. *J Theor Biol* 2004, **226**:95–109.

31. Rosin CD, Belew RK, Morris GM, Olson AJ, Goodsell DS: **Coevolutionary analysis of resistance-evading peptidomimetic inhibitors of HIV-1 protease**. *Proc Natl Acad Sci U S A* 1999, **96**(4):1369–1374.

32. Bocharov G, Ford NJ, Edwards J, Breinig T, Wain-Hobson S, Meyerhans A: **A genetic-algorithm approach to simulating human immunodeficiency virus evolution reveals the strong impact of multiply infected cells and recombination**. *J Gen Virol* 2005, **86**(Pt 11):3109–3118.

33. Leonard JN, Schaffer DV: **Computational design of antiviral RNA interference strategies that resist human immunodeficiency virus escape**. *J Virol* 2005, **79**(3):1645–1654.

34. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK: **Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B**. *J Med Chem* 2002, **45**(11):2213–2221.

35. Cheng T, Goodsell D, Kan C: **Identification of Sanguinarine as a Novel HIV Protease Inhibitor from High-Throughput Screening of 2,000 Drugs and Natural Products with a Cell-Based Assay**. *Letters in Drug Design & Discovery* 2005, **2**(5):364–371.

36. Schneider G, Bohm HJ: **Virtual screening and fast automated docking methods**. *Drug Discov Today* 2002, **7**:64–70.

37. Shoichet BK: **Virtual screening of chemical libraries**. *Nature* 2004, **432**(7019):862–865.

38. Alvarez JC: **High-throughput docking as a source of novel drug leads**. *Curr Opin Chem Biol* 2004, **8**(4):365–370.

39. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ: **Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function**. *Journal of Computational Chemistry* 1999, **19**(14):1639–1662.

40. Huey R, Morris GM, Olson AJ, Goodsell DS: **A semiempirical free energy force field with charge-based desolvation**. *J Comput Chem* 2007, **28**(6):1145–1152.

41. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ: **The Amber biomolecular simulation programs**. *J Comput Chem* 2005, **26**(16):1668–1688.

42. Rosenfeld RJ, Goodsell DS, Musah RA, Morris GM, Goodin DB, Olson AJ: **Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling**. *J Comput Aided Mol Des* 2003, **17**(8):525–536.

43. Ruvinsky AM, Kozintsev AV: **New and fast statistical-thermodynamic method for computation of protein-ligand binding entropy substantially improves docking accuracy**. *J Comput Chem* 2005, **26**(11):1089–1095.

44. Ruvinsky AM, Kozintsev AV: **Novel statistical-thermodynamic methods to predict protein-ligand binding positions using probability distribution functions**. *Proteins* 2006, **62**:202–208.

45. Ruvinsky AM: **Role of binding entropy in the refinement of protein-ligand docking predictions: Analysis based on the use of 11 scoring functions**. *J Comput Chem* 2007.

46. Segal MR, Barbour JD, Grant RM: **Relating HIV-1 sequence variation to replication capacity via trees and forests**. *Stat Appl Genet Mol Biol* 2004, **3**.

47. Birkner MD, Sinisi SE, van der Laan MJ: **Multiple testing and data adaptive regression: an application to HIV-1 sequence data**. *Stat Appl Genet Mol Biol* 2005, **4**.

48. Stone EA, Sidow A: **Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity**. *Genome Res* 2005, **15**(7):978–986.

49. Doyon L, Croteau G, Thibeault D, Poulin F, Pilote L, Lamarre D: **Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors**. *J Virol* 1996, **70**(6):3763–3769.

50. Zhang YM, Imamichi H, Imamichi T, Lane HC, Falloon J, Vasudevachari MB, Salzman NP: **Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites**. *J Virol* 1997, **71**(9):6662–6670.

51. Maguire MF, Guinea R, Griffin P, Macmanus S, Elston RC, Wolfram J, Richards N, Hanlon MH, Porter DJ, Wrin T, Parkin N, Tisdale M, Furfine E, Petropoulos C, Snowden BW, Kleim JP: **Changes in human immunodeficiency virus type 1 Gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro**. *J Virol* 2002, **76**(15):7398–7406.

52. Pettit SC, Henderson GJ, Schiffer CA, Swanstrom R: **Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease**. *J Virol* 2002, **76**(20):10226–10233.

53. Prabu-Jeyabalan M, Nalivaika EA, King NM, Schiffer CA: **Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease**. *J Virol* 2004, **78**(22):12446–12454.

54. Nijhuis M, van Maarseveen NM, Lastere S, Schipper P, Coakley E, Glass B, Rovenska M, de Jong D, Chappey C, Goedegebuure IW, Heilek-Snyder G, Dulude D, Cammack N, Brakier-Gingras L, Konvalinka J, Parkin N, Kräusslich HG, Brun-Vezinet F, Boucher CA: **A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism**. *PLoS Med* 2007, **4**.

55. Burges C: **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery* 1998, **2**(2):121–167.

56. Witten IH, Frank E: *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2nd edition 2005.

57. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES: **Characterization of single-nucleotide polymorphisms in coding regions of human genes**. *Nat Genet* 1999, **22**(3):231–238.

58. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions**. *Genome Res* 2001, **11**(5):863–874.

59. Wang W, Kollman PA: **Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance**. *Proc Natl Acad Sci U S A* 2001, **98**(26):14937–14942.

60. Friedman N, Ninio M, Pe'er I, Pupko T: **A structural EM algorithm for phylogenetic inference**. *J Comput Biol* 2002, **9**(2):331–353.

61. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA: **Complete mutagenesis of the HIV-1 protease**. *Nature* 1989, **340**(6232):397–400.

62. Rhee SY, Fessel WJ, Zolopa AR, Hurley L, Liu T, Taylor J, Nguyen DP, Slome S, Klein D, Horberg M, Flamm J, Follansbee S, Schapiro JM, Shafer RW: **HIV-1 Protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance**. *J Infect Dis* 2005, **192**(3):456–465.

63. Rosé JR, Babé LM, Craik CS: **Defining the level of human immunodeficiency virus type 1 (HIV-1) protease activity required for HIV-1 particle maturation and infectivity**. *J Virol* 1995, **69**(5):2751–2758.

64. Moody MD, Pettit SC, Shao W, Everitt L, Loeb DD, Hutchison CA, Swanstrom R: **A side chain at position 48 of the human immunodeficiency virus type-1 protease flap provides an additional specificity determinant**. *Virology* 1995, **207**(2):475–485.

65. Pazhanisamy S, Stuver CM, Cullinan AB, Margolin N, Rao BG, Livingston DJ: **Kinetic characterization of human immunodeficiency virus type-1 protease-resistant variants**. *J Biol Chem* 1996, **271**(30):17979–17985.

66. Nillroth U, Vrang L, Markgren PO, Hultén J, Hallberg A, Danielson UH: **Human immunodeficiency virus type 1 proteinase resistance to symmetric cyclic urea inhibitor analogs**. *Antimicrob Agents Chemother* 1997, **41**(11):2383–2388.

67. Ridky TW, Kikonyogo A, Leis J, Gulnik S, Copeland T, Erickson J, Wlodawer A, Kurinov I, Harrison RW, Weber IT: **Drug-resistant HIV-1 proteases identify enzyme residues important for substrate selection and catalytic rate**. *Biochemistry* 1998, **37**(39):13835–13845.

68. Chou KC, Tomasselli AG, Reardon IM, Heinrikson RL: **Predicting Human Immunodeficiency Virus Protease Cleavage Sites in Proteins by a Discriminant Function Method**. *Proteins: Structure, Function, and Genetics* 1996, **24**:51–72.

69. Cai YD, Liu XJ, Xu XB, Chou KC: **Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein**. *Journal of Computational Chemistry* 2002, **23**:267–274.

70. Narayanan A, Wu X, Yang ZR: **Mining viral protease data to extract cleavage knowledge**. *Bioinformatics* 2002, **18**:S5–S13.

71. Yang ZR, Chou KC: **Mining Biological Data Using Self-Organizing Map**. *Journal of Chemical Information and Computer Sciences* 2003, **43**:1748–1753.

72. Yang ZR, Chou KC: **Bio-support vector machines for computational proteomics**. *Bioinformatics* 2004, **20**(5):735–741.

73. Rognvaldsson T, You L: **Why neural networks should not be used for HIV-1 protease cleavage site prediction**. *Bioinformatics* 2004, **20**(11):1702–1709.

74. Yang ZR, Dalby AR, Qiu J: **Mining HIV protease cleavage data using genetic programming with a sum-product function**. *Bioinformatics* 2004, **20**(18):3398–3405.

75. Beck ZQ, Hervio L, Dawson PE, Elder JH, Madison EL: **Identification of Efficiently Cleaved Substrates for HIV-1 Protease Using a Phage Display Library and Use in Inhibitor Development**. *Journal of Virology* 2000, **274**:391–401.

76. Beck ZQ, Lin YC, Elder JH: **Molecular Basis for the Relative Substrate Specificity of Human Immunodeficiency Virus Type 1 and Feline Immunodeficiency Virus Proteases**. *Journal of Virology* 2001, **75**(19):9458–9469.

77. Chang CC, Lin CJ: *LIBSVM: a library for support vector machines* 2001. [Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm].

78. Quinlan J: *C4. 5: Programs for Machine Learning*. Morgan Kaufmann 1993.

79. Shafer RW: **Rationale and uses of a public HIV drug-resistance database**. *J Infect Dis* 2006, **194 Suppl 1**:51–58.

80. Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR, Pillay D, Schapiro JM, Richman DD: **Update of the Drug Resistance Mutations in HIV-1: Spring 2008**. *Top HIV Med* 2008, **16**:62–68.

81. Beerenwinkel N, Lengauer T, Däumer M, Kaiser R, Walter H, Korn K, Hoffmann D, Selbig J: **Methods for optimizing antiviral combination therapies**. *Bioinformatics* 2003, **19 Suppl 1**:16–25.

82. Brik A, Alexandratos J, Lin YC, Elder JH, Olson AJ, Wlodawer A, Goodsell DS, Wong CH: **1,2,3-triazole as a peptide surrogate in the rapid synthesis of HIV-1 protease inhibitors.** *Chembiochem* 2005, **6**(7):1167–1169.

83. Barbour JD, Wrin T, Grant RM, Martin JN, Segal MR, Petropoulos CJ, Deeks SG: **Evolution of phenotypic drug susceptibility and viral replication capacity during long-term virologic failure of protease inhibitor therapy in human immunodeficiency virus-infected adults**. *J Virol* 2002, **76**(21):11104–11112.

84. Johnson VA, Brun-Vezinet F, Clotet B, Conway B, Kuritzkes DR, Pillay D, Schapiro JM, Telenti A, Richman DD: **Update of the drug resistance mutations in HIV-1: Fall 2005**. *Top HIV Med* 2005, **13**(4):125–131.

85. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**:235–242.

86. Canutescu AA, Shelenkov AA, Dunbrack RL: **A graph-theory algorithm for rapid protein side-chain prediction**. *Protein Sci* 2003, **12**(9):2001–2014.

87. Heaslet H, Lin YC, Tam K, Torbett BE, Elder JH, Stout CD: **Crystal structure of an FIV/HIV chimeric protease complexed with the broad-based inhibitor, TL-3**. *Retrovirology* 2007, **4**:1–1.

88. Mitsuya Y, Winters MA, Fessel WJ, Rhee SY, Hurley L, Horberg M, Schiffer CA, Zolopa AR, Shafer RW: **N88D facilitates the co-occurrence of D30N and L90M and the development of multidrug resistance in HIV type 1 protease following nelfinavir treatment failure**. *AIDS Res Hum Retroviruses* 2006, **22**(12):1300–1305.

89. Deforche K, Silander T, Camacho R, Grossman Z, Soares MA, Van Laethem K, Kantor R, Moreau Y, Vandamme AM: **Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance**. *Bioinformatics* 2006, **22**(24):2975–2979.

90. Rhee SY, Liu TF, Holmes SP, Shafer RW: **HIV-1 subtype B protease and reverse transcriptase amino acid covariation**. *PLoS Comput Biol* 2007, **3**(5).

91. Beerenwinkel N, Däumer M, Sing T, Rahnenfuhrer J, Lengauer T, Selbig J, Hoffmann D, Kaiser R: **Estimating HIV evolutionary pathways and the genetic barrier to drug resistance**. *J Infect Dis* 2005, **191**(11):1953–1960.

92. Deforche K, Camacho R, Grossman Z, Silander T, Soares MA, Moreau Y, Shafer RW, Van Laethem K, Carvalho AP, Wynhoven B, Cane P, Snoeck J, Clarke J, Sirivichayakul S, Ariyoshi K, Holguin A, Rudich H, Rodrigues R, Bouzas MB, Cahn P, Brigido LF, Soriano V, Sugiura W, Phanuphak P, Morris L, Weber J, Pillay D, Tanuri A, Harrigan PR, Shapiro JM, Katzenstein DA, Kantor R, Vandamme AM: **Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors**. *Infect Genet Evol* 2006.

93. Poon AF, Kosakovsky Pond SL, Richman DD, Frost SD: **Mapping protease inhibitor resistance to human immunodeficiency virus type 1 sequence polymorphisms within patients**. *J Virol* 2007, **81**(24):13598–13607.

94. Larder BA: **Interactions between drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase**. *J Gen Virol* 1994, **75 (Pt 5)**:951–957.

95. Hu Z, Giguel F, Hatano H, Reid P, Lu J, Kuritzkes DR: **Fitness comparison of thymidine analog resistance pathways in human immunodeficiency virus type 1**. *J Virol* 2006, **80**(14):7020–7027.

96. Tang J, Hartsuck JA: **A kinetic model for comparing proteolytic processing activity and inhibitor resistance potential of mutant HIV-1 proteases**. *FEBS Lett* 1995, **367**(2):112–116.

97. Eron JJ, Haubrich R, Lang W, Pagano G, Millard J, Wolfram J, Snowden W, Pedneault L, Tisdale M: **A phase II trial of dual protease inhibitor therapy: amprenavir in combination with indinavir, nelfinavir, or saquinavir**. *J Acquir Immune Defic Syndr* 2001, **26**(5):458–461.

98. Yonetani T, Theorell H: **Studies on Liver Alcohol Hydrogenase Complexes. 3. Multiple Inhibition Kinetics in the Presence of Two Competitive Inhibitors.** *Arch Biochem Biophys* 1964, **106**:243–251.

99. Burch CL, Chao L: **Epistasis and its relationship to canalization in the RNA virus phi 6**. *Genetics* 2004, **167**(2):559–567.

100. Sanjuan R, Moya A, Elena SF: **The contribution of epistasis to the architecture of fitness in an RNA virus.** *Proc Natl Acad Sci U S A* 2004, **101**(43):15376–15379.

101. Sugiura W, Matsuda Z, Yokomaku Y, Hertogs K, Larder B, Oishi T, Okano A, Shiino T, Tatsumi M, Matsuda M, Abumi H, Takata N, Shirahata S, Yamada K, Yoshikura H, Nagai Y: **Interference between D30N and L90M in selection and development of protease inhibitor-resistant human immunodeficiency virus type 1**. *Antimicrob Agents Chemother* 2002, **46**(3):708–715.

102. Mo H, Parkin N, Stewart K, Lu L, Dekhtyar T, Kempf D, Molla A: **I84A and I84C mutations in protease confer high-level resistance to protease inhibitors and impair replication capacity**. *Antivir Ther* 2003, **8**:S56.

103. Perrin V, Mammano F: **Parameters driving the selection of nelfinavir-resistant human immunodeficiency virus type 1 variants**. *J Virol* 2003, **77**(18):10172–10175.

104. Doyon L, Tremblay S, Bourgon L, Wardrop E, Cordingley MG: **Selection and characterization of HIV-1 showing reduced susceptibility to the non-peptidic protease inhibitor tipranavir**. *Antiviral Res* 2005, **68**:27–35.

105. Belew RK, Chang MW: **Modeling recombination's role in the evolution of HIV drug resistance**. In *Artificial Life X: The Tenth International Conference on the Synthesis and Simulation of Living Systems*. Edited by Rocha LM, Bedau M, Floreano D, Goldstone R, Vespignani A, Yaeger L, MIT Press 2006:98–104.

106. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM: **Dynamics of HIV-1 recombination in its natural target cells**. *Proc Natl Acad Sci U S A* 2004, **101**(12):4204–4209.

107. Mansky LM, Temin HM: **Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase**. *J Virol* 1995, **69**(8):5087–5094.

108. Gao F, Chen Y, Levy DN, Conway JA, Kepler TB, Hui H: **Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious**. *J Virol* 2004, **78**(5):2426–2433.

109. Suzuki Y, Yamaguchi-Kabata Y, Gojobori T: **Nucleotide substitution rates of HIV-1**. *AIDS Rev* 2000, **2**:39–47.

110. Baker JE: **Reducing bias and inefficiency in the selection algorithm**. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*, Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc. 1987:14–21.

111. Rusert P, Fischer M, Joos B, Leemann C, Kuster H, Flepp M, Bonhoeffer S, Günthard HF, Trkola A: **Quantification of infectious HIV-1 plasma viral load using a boosted in vitro infection protocol**. *Virology* 2004, **326**:113–129.

112. Althaus CL, Bonhoeffer S: **Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1**. *J Virol* 2005, **79**(21):13572–13578.

113. Suryavanshi GW, Dixit NM: **Emergence of recombinant forms of HIV: dynamics and scaling.** *PLoS Comput Biol* 2007, **3**(10):2003–2018.

114. Deforche K, Camacho R, Van Laethem K, Lemey P, Rambaut A, Moreau Y, Vandamme AM: **Estimation of an in vivo fitness landscape experienced by HIV-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment.** *Bioinformatics* 2008, **24**:34–41.

115. Domingo E, Holland JJ: **RNA virus mutations and fitness for survival**. *Annu Rev Microbiol* 1997, **51**:151–178.

116. of Health D, Services H: **Panel on Clinical Practices for Treatment of HIV Infection. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents.** `[http://aidsinfo.nih.gov/]`.

117. Sammon JW: **A Nonlinear Mapping for Data Structure Analysis**. *IEEE Transactions on Computers* 1969, **C-18**(5):401–409.

118. Bajorath J: **Integration of virtual and high-throughput screening**. *Nat Rev Drug Discov* 2002, **1**(11):882–894.

119. Lyne PD: **Structure-based virtual screening: an overview.** *Drug Discov Today* 2002, **7**(20):1047–1055.

120. Shoichet BK, McGovern SL, Wei B, Irwin JJ: **Lead discovery using molecular docking**. *Curr Opin Chem Biol* 2002, **6**(4):439–446.

121. Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS: **Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock**. *Proteins* 2002, **46**:34–40.

122. Lee T, Laco GS, Torbett BE, Fox HS, Lerner DL, Elder JH, Wong CH: **Analysis of the S3 and S3' subsite specificities of feline immunodeficiency virus (FIV) protease: development of a broad-based protease inhibitor efficacious against FIV, SIV, and HIV in vitro and ex vivo**. *Proc Natl Acad Sci U S A* 1998, **95**(3):939–944.

123. Kutilek VD, Sheeter DA, Elder JH, Torbett BE: **Is resistance futile?** *Curr Drug Targets Infect Disord* 2003, **3**(4):295–309.

124. Knegtel RM, Kuntz ID, Oshiro CM: **Molecular docking to ensembles of protein structures**. *J Mol Biol* 1997, **266**(2):424–440.

125. Carlson HA, McCammon JA: **Accommodating Protein Flexibility in Computational Drug Design**. *Mol Pharmacol* 2000, **57**(2):213–218.

126. Fernandes MX, Kairys V, Gilson MK: **Comparing ligand interactions with multiple receptors via serial docking**. *J Chem Inf Comput Sci* 2004, **44**(6):1961–1970.

127. Hayashi Y, Sakaguchi K, Kobayashi M, Kobayashi M, Kikuchi Y, Ichiishi E: **Molecular evaluation using in silico protein interaction profiles**. *Bioinformatics* 2003, **19**(12):1514–1523.

128. Vinkers HM, de Jonge MR, Daeyaert ED, Heeres J, Koymans LM, van Lenthe JH, Lewi PJ, Timmerman H, Janssen PA: **Inhibition and Substrate Recognition–a Computational Approach Applied to HIV Protease**. *J Comput Aided Mol Des* 2003, **17**(9):567–581.

129. Wlodawer A: **Rational Approach to Aids Drug Design through Structural Biology**. *Annu Rev Med* 2002, **53**:595–614.

130. Lerner B, Guterman H, Aladjem M, Dinstein I, Romem Y: **On pattern classification with Sammon's nonlinear mapping – An experimental study**. *Pattern Recognition* 1998, **31**(4):371–381.

131. Chang CE, Chen W, Gilson MK: **Ligand configurational entropy and protein binding**. *Proc Natl Acad Sci U S A* 2007, **104**(5):1534–1539.

132. Bottegoni G, Cavalli A, Recanatini M: **A comparative study on the application of hierarchical-agglomerative clustering approaches to organize outputs of reiterated docking runs**. *J Chem Inf Model* 2006, **46**(2):852–862.

133. Carroll KS, Gao H, Chen H, Stout CD, Leary JA, Bertozzi CR: **A conserved mechanism for sulfonucleotide reduction**. *PLoS Biol* 2005, **3**(8).

134. Chartron J, Carroll KS, Shiau C, Gao H, Leary JA, Bertozzi CR, Stout CD: **Substrate recognition, protein dynamics, and iron-sulfur cluster in Pseudomonas aeruginosa adenosine 5'-phosphosulfate reductase**. *J Mol Biol* 2006, **364**(2):152–169.

135. Hayes J, Stein M, Weiser J: **Accurate Calculations of Ligand Binding Free Energies: Chiral Separation with Enantioselective Receptors**. *JOURNAL OF PHYSICAL CHEMISTRY A* 2004, **108**(16):3572–3580.

136. Head M, Given J, Gilson M: **Mining minima": direct computation of conformational free energy**. *J. Phys. Chem. A* 1997, **101**(8):1609–1618.

137. Fersht A: *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. Freeman, New York 1999.

138. Gasteiger J, Marsili M: **Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges**. *Tetrahedron* 1980, **36**(22):3219–3228.

139. Chang MW, Lindstrom W, Olson AJ, Belew RK: **Analysis of HIV wild-type and mutant structures via in silico docking against diverse ligand libraries**. *J Chem Inf Model* 2007, **47**(3):1258–1262.

140. Handen J (Ed): *Industrialization of Drug Discovery: From Target Selection Through Lead Optimization*. Boca Raton, FL: CRC Press 2005.

141. Shoichet BK: **Screening in a spirit haunted world**. *Drug Discov Today* 2006, **11**(13-14):607–615.

142. Jenkins JL, Kao RY, Shapiro R: **Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin**. *Proteins* 2003, **50**:81–93.

143. Weislow OS, Kiser R, Fine DL, Bader J, Shoemaker RH, Boyd MR: **New soluble-formazan assay for HIV-1 cytopathic effects: application to high-flux screening of synthetic and natural products for AIDS-antiviral activity**. *J Natl Cancer Inst* 1989, **81**(8):577–586.

144. Pan Y, Huang N, Cho S, MacKerell ADJ: **Consideration of molecular weight during compound selection in virtual target-based database screening.** *J Chem Inf Comput Sci* 2003, **43**:267–272.

145. McGovern SL, Caselli E, Grigorieff N, Shoichet BK: **A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening**. *J Med Chem* 2002, **45**(8):1712–1722.

146. Chang MW, Belew RK, Carroll KS, Olson AJ, Goodsell DS: **Empirical entropic contributions in computational docking: Evaluation in APS reductase complexes**. *J Comput Chem* 2008, **29**(11):1753–1761.

147. National Center for Biotechnology Information: **Growth of GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html].

148. Preece S: **A spreading activation network model for information retrieval**. *PhD thesis*, University of Illinois 1982.

149. Cohen P, Kjeldsen R: **Information retrieval by constrained spreading activation in semantic networks**. *Information Processing and Management: an International Journal* 1987, **23**(4):255–268.

150. Salton G, Buckley C: **Improving retrieval performance by relevance feedback**. *Journal of the American Society for Information Science* 1990, **41**(4):288–297.

151. Belew R: **Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents**. *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval* 1989, :11–20.

152. Doszkocs T, Reggia J, Lin X: **Connectionist models and information retrieval**. *Annual Review of Information Science and Technology (ARIST)* 1990, **25**:209–260.

153. Belew R: *Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge University Press 2000.

154. Brin S, Page L: **The anatomy of a large-scale hypertextual Web search engine**. *Computer Networks and ISDN Systems* 1998, **30**(1-7):107–117.

155. [`http://lucene.apache.org/`].

156. Erlanson DA, Braisted AC, Raphael DR, Randal M, Stroud RM, Gordon EM, Wells JA: **Site-directed ligand discovery**. *Proc Natl Acad Sci U S A* 2000, **97**(17):9367–9372.