12-16-2014

HIV Drug Resistant Prediction and Featured Mutants Selection using Machine Learning Approaches

Xiaxia Yu

HIV DRUG RESISTANT PREDICTION AND FEATURED MUTANTS SELECTION USING MACHINE LEARNING APPROACHES

by

XIAXIA YU

Under the Direction of Robert W. Harrison, Ph.D.

ABSTRACT

HIV/AIDS is widely spread and ranks as the sixth biggest killer all over the world. Moreover, due to the rapid replication rate and the lack of proofreading mechanism of HIV virus, drug resistance is commonly found and is one of the reasons causing the failure of the treatment. Even though the drug resistance tests are provided to the patients and help choose more efficient drugs, such experiments may take up to two weeks to finish and are expensive. Because of the fast development of the computer, drug resistance prediction using machine learning is feasible.

In order to accurately predict the HIV drug resistance, two main tasks need to be solved: how to encode the protein structure, extracting the more useful information and feeding it into the machine learning tools; and which kinds of machine learning tools to choose. In our research, we first proposed a new protein encoding algorithm, which could convert various sizes of proteins into a fixed size vector.

This algorithm enables feeding the protein structure information to most state of the art machine learning algorithms. In the next step, we also proposed a new classification algorithm based on sparse representation. Following that, mean shift and quantile regression were included to help extract the feature information from the data. Our results show that encoding protein structure using our newly proposed method is very efficient, and has consistently higher accuracy regardless of type of machine learning tools. Furthermore, our new classification algorithm based on sparse representation is the first application of sparse representation performed on biological data, and the result is comparable to other state of the art classification algorithms, for example ANN, SVM and multiple regression. Following that, the mean shift and quantile regression provided us with the potentially most important drug resistant mutants, and such results might help biologists/chemists to determine which mutants are the most representative candidates for further research.

INDEX WORDS: HIV-1 Drug resistance prediction, Delaunay triangulation, Sparse representation, Machine learning, Classification algorithms, Mean shift

HIV DRUG RESISTANT PREDICTION AND FEATURED MUTANTS SELECTION USING MACHINE LEARNING APPROACHES

by

XIAXIA YU

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2014

MACHINE LEARNING APPROACHES FOR GENOTYPE-PHENOTYPE PREDICTION AND FEATURED DRUG RESISTANT MUTANTS RETRIEVAL

by

XIAXIA YU

Committee Chair: Robert W. Harrison

Committee: Irene T. Weber

Yi Pan

Yanqing Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2014

DEDICATION

To my parents Jinhong Yu and Xiaoling Peng,

To my husband Yi Gao,

To my children Sophie and Allen

ACKNOWLEDGEMENTS

I would like to thank everyone who support and help me during my Ph. D study. Thank you!

First, I would like to thank my advisers, Dr. Robert W. Harrison and Dr. Irene T. Weber. Thank you, because during these years of study, I've not only learned the knowledge and the techniques of the area, but also how to become a good researcher and active learner. Thank you, for all your supports during my down time, guiding me out of it. Without that, I couldn't survive in my doctorate study. It is my honor to be the student of you.

I also would like to thank my thesis committee, Dr. Yi Pan and Dr. Yanqing Zhang. Thank you so much for your helpful suggestions and guidance, which improve my dissertation research a lot. Thank you, Dr. Yanqing Zhang, for your warm encouragement and comforting words after my qualify exam and dissertation proposal. You cheered me up!

Thank you, Yuanfang and Johnny! Thank you for teaching me how to collect data, process data in such a detail. It's really a great experience to collect data with you. I've learned a lot from you!

Thank you to everyone in our great group: Tracy, Ying, Guoxing, Brian, Ting, Bin, Hongmei, Xiaodan, Amit, Na'el and Erin. Thank you for your help during these years. It’s always a pleasure to spend time with you: setting up experiments, attending group meetings, and discussing the problems, and sharing the happiness with you.

Thanks to all my friends, for your friendship and support.

To my parents, my dear husband, for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Decision making is everywhere in our daily lives. We use our past experience to make the decision. However, sometimes it is hard to make an optimal decision, sometimes we are uncertain about the correctness of our decisions, and sometimes it is not that straightforward to find the relations or useful information from our past experience, so on and so forth. Due to all those reasons, machine learning is one of the options to help us make decisions. Similar to our decision making procedure, machine learning methods also use the past experience to automatically make decisions. In this process, the past experience is called training data, while the new situation is called the testing data. The given results could be considered as the decisions. Machine learning approaches could be used almost everywhere, for example, web search engine[1], stock market analysis[2], biomedical diagnosis[3], and so on. This dissertation focuses on using machine learning tools to solve biomedical problems.

In bioinformatics area, understanding the relationship between the protein sequence, structure and function renders a key component during the past decades[4, 5]. Traditionally, chemical/biological experiments, nuclear magnetic resonance (NMR) or X-ray crystallography, for instances, could be used to retrieve the relationship between them. However, even the minimal experiments are expensive and time consuming. Moreover, even with all the experiments and time, it is highly possible that no useful information could be obtained from the structure, or no good structures could be obtained. Nowadays, due to the rapid development and the wide spread of the computers, *in silico* experiments are introduced to solve this problem. With the advance of the computational power, we are enabled to have access to more and more knowledge of proteins.

Since the first case of AIDS was found in United States in early 1980s, AIDS has become one of the most severe diseases all over the world. It is known that AIDS is caused by HIV. However, due to the characteristics of the retrovirus, drug resistance is commonly seen during the anti-AIDS treatment and

often causes the failure of the treatment. Therefore, computational method is necessary to shorten the patients' waiting time, saving both time and money.

In this study we are focusing on using machine learning methods to predict the mutants' drug resistance to certain drugs, and furthermore proposed new algorithms to identifying the most representative drug resistant mutants among the whole drug resistance data. With these goals in mind, our research focuses on three aims:

- Is there an efficient way to encode both protein sequence and structure information?
- Is there an accurate method to predict whether a given mutant is drug resistant from sequence data? Does including structural data in the classification improve the accuracy?
- Can machine learning be used to identify critical or important mutations and aid in the design of biological/chemical experiments?

In order to achieve these three goals, our research starts from proposing a new encoding method to represent protein structure by using Delaunay Triangulation. In this method, the alpha carbon position is used to represent the whole amino acid residue, and the average distance of the same amino acid pairs were recorded to generate the adjacency matrix, and therefore based on these adjacency matrices, the fixed size vector could be obtained to represent each protein structure. Following that, we further tested such encoding method on more data, and then performed this on the prediction of the drug resistance property of certain mutants of the HIV-1 protease. By utilizing the recent advances in the sparse signal representation and compressive sensing, we proposed a sparse dictionary technique for the purpose of the drug resistance prediction. The cross-validation shows high consistency by using the publicly available data set. Furthermore, mean shift algorithm is included to extract the most important feature from the categories. Such results might be able to guide the experimental design for the biological/chemical study of the HIV-1 drug resistance.

To meet the needs of predicting the potential mutants by using computational methods, our research includes the following subjects:

Chapter 2: HIV/AIDS Background and its drug resistance: In this part, the general background of the HIV-1 protease, reverse transcriptase and their inhibitors used during the HIV/AIDS treatment is introduced. Moreover, the cause of the drug resistance, together with the importance of why this needs to be studied is also present in this part.

Chapter 3: Justification of the research topics: we will demonstrate a brief literature review on computational drug resistance, protein representation, as well as the sparse representation, a new technique we used as a classifier in our study.

Chapter 4 and chapter 5 on solving our first aim: finding new protein representation methods. In *Chapter 4: Encoding protein structure with functions using Delaunay* Triangulation: we proposed a new encoding method to represent the protein structure. Following that, in *Chapter 5: An application of new protein encoding methods using Delaunay Triangulation:* one application of our new proposed protein encoding methods using Delaunay Triangulation is demonstrated. In this application, we tested on both HIV protease and HIV reverse transcriptase and included multiple linear regression as the classification tool.

Chapter 6 and chapter 7 on solving our second aim: developing a new classification algorithm to distinguishing between the drug resistant and the non-drug resistant mutants. In *Chapter 6: Sparse representation for prediction of HIV-1 protease drug resistance*, we focus on retrieving the protein characteristics, including the property of drug resistance, and the folding information, from protein's sequence information. Specifically, we study the problem of HIV-1 protease drug resistant mutant prediction: We proposed a new classification algorithm based on sparse representation to predict the drug resistant property of the given HIV-1 protease mutant. In *Chapter 7: Prediction of HIV drug resistance from geno-*

type with encoded three-dimensional protein structure, more results were demonstrated here to solve the classification problem using our newly developed algorithm.

In *Chapter 8: Identifying essential features for the representative mutants from drug resistance data.* In this part of the research, we focus on finding out the most representative potential mutants which are resistant to certain drugs. The finding of such mutants might be a guide for biologists/chemists to select the most likely mutants for more research.

In *Chapter 9: Future work and summaries*. In this chapter, we present some possible future directions to improve/continue this work. After that, we summarize all the work presented in this dissertation, and make a conclusion based on this work.

2 HIV/AIDS Background and its drug resistance

2.1 The Current Status of AIDS

It has been almost three decades since the first case of AIDS was found in the United States in the early 80s, last century. At the end of year 2012, about 35.3 million people are living with HIV, and among them about 2.7 million people are newly infected[6]. Moreover, by the end of year 2011, nearly 30 million people died because of AIDS[7]. Currently, there is no effective vaccine or cure for AIDS; however, because of the Highly Active Antiretroviral Therapy (HAART), which was proposed in mid-1990s and the idea is to use three or four different drugs with different targets during the treatment to obtain a successful therapy, the infected growth rate is stablized (as shown in Figure 1)[7] and the death rate decreased to 47% in 1997 only one decade after the first AIDS case was found (as shown in Figure 2)[8].

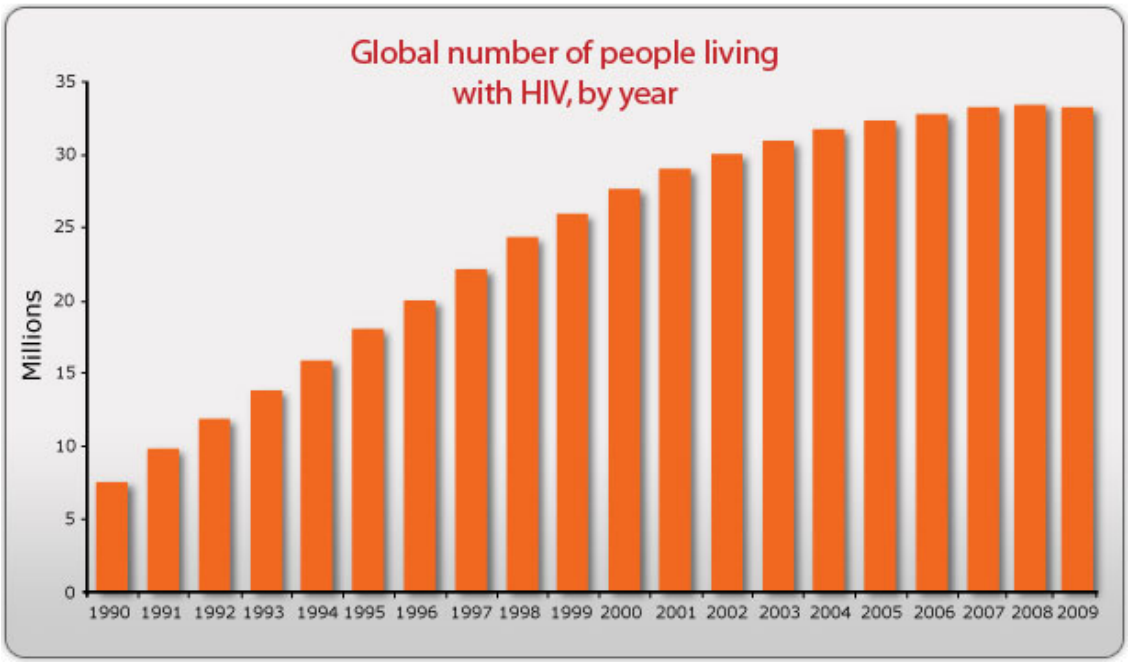


Figure 2.1.1.1 Global number of people living with HIV, by year[7]

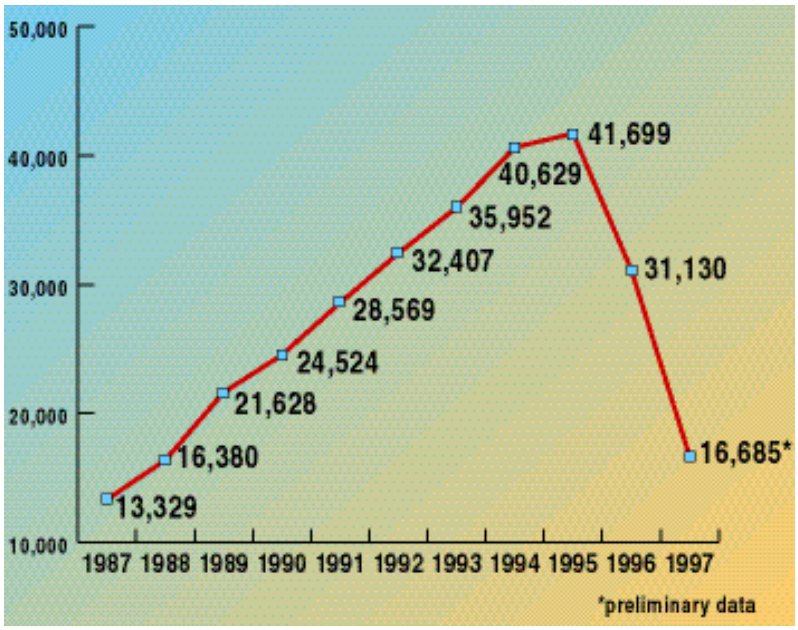


Figure 2.1.1.2 AIDS Deaths Since 1987[8]

Table 2.1.1.1 Approved antiretroviral drugs in the USA and Europe[9]

Generic name	Brand name	Manufacturer	Approval Date
Zidovudine	Retrovir	GlaxoSmithKline	03/19/1987
Didanosine	Videx (tablet)	Bristol-Myers Squibb	10/09/1991
	Videx EC (capsule)	Bristol-Myers Squibb	10/31/2000
Zalcitabine	Hivid	Hoffmann-La Roche	06/19/1992
Stavudine	Zerit	Bristol-Myers Squibb	06/24/1994
Lamivudine	Epivir	GlaxoSmithKline	11/17/1995
Saquinavir	Invirase (hard gel capsule)	Hoffmann-La Roche	12/06/1995
	Fortovase (soft gel capsule)	Hoffmann-La Roche	11/07/1997
Ritonavir	Norvir	Abbott Laboratories	03/01/1996
Indinavir	Crixivan	Merck	03/13/1996
Nevirapine	Viramune	Boehringer Ingelheim	06/24/1996
Nelfinavir	Viracept	Agouron Pharmaceuticals	03/14/1997
Delavirdine	Rescriptor	Pfizer	04/04/1997
Efavirenz	Sustiva (USA)	Bristol-Myers Squibb	09/17/1998
	Stocrin (Europe)	Merck	09/17/1998
Abacavir	Ziagen	GlaxoSmithKline	12/17/1998

Amprenavir	Agenerase	GlaxoSmithKline	04/15/1999
Lopinavir+ritonavir	Kaletra	Abbott Laboratories	09/15/2000
	Aluvia (developing world)	Abbott Laboratories	09/15/2000
Tenofovir	disoproxil fumarate (TDF)	Viread Gilead Sciences	10/26/2001
Enfuvirtide	Fuzeon	Hoffmann-La Roche & Trimeris	03/13/2003
Atazanavir	Reyataz	Bristol-Myers Squibb	06/20/2003
Emtricitabine	Emtriva	Gilead Sciences	07/02/2003
Fosamprenavir	Lexiva (USA)	GlaxoSmithKline	10/20/2003
	Telzir (Europe)	GlaxoSmithKline	10/20/2003
Tipranavir	Aptivus	Boehringer Ingelheim	06/22/2005
Darunavir	Prezista	Tibotec Inc.	06/23/2006
Maraviroc	Celsentri (Europe)	Pfizer	09/18/2007
	Selzentry (USA)	Pfizer	09/18/2007
Raltegravir	Isentress	Merck & Co. Inc.	10/12/2007
Etravirine	Intelence	Tibotec Therapeutics	11/18/2008
Fixed dose drug combinations			
Lamivudine and zidovudine	Combivir	GlaxoSmithKline	09/27/1997
Abacavir, zidovudine and lamivudine	Trizivir	GlaxoSmithKline	11/14/2000
Abacavir and lamivudine	Epzicom (USA)	GlaxoSmithKline	08/02/2004
	Kivexa (Europe)	GlaxoSmithKline	08/02/2004
TDF and emtricitabine	Truvada	Gilead Sciences	08/02/2004
Efavirenz, emtricitabine and TDF	Atripla	Bristol-Myers Squibb&Gilead Sciences	07/12/2006

Up till now, researchers and scientists have worked hard, and developed a total of twenty-five Food and Drug Administration (FDA) approved antiretroviral drugs for the treatment of HIV/AIDS. All these drugs are categorized into six different classes: seven nucleoside reverse transcriptase inhibitors (NRTIs); one nucleotide reverse transcriptase inhibitors (NtRTIs); four non-nucleoside reverse transcriptase inhibitors (NNRTIs); ten protease inhibitors (PIs); two cell entry inhibitors; and two integrase inhibitors (INIs)[9]; and target on different steps in HIV-1 life cycle: viral entry, reverse transcription, integration, and viral maturation[10]. All the drugs are listed in Table 1.

2.1.1 HIV-1 life cycle

AIDS is caused by human immunodeficiency virus type 1 (HIV-1), which is one of the retroviruses. The life cycle of the HIV consists of two phases: the early phase and the late phase.

The early phase includes three steps before the replication of the viral genome. The first step is binding: the virus recognizes the CD4 protein, which usually acts as an immune recognizer, and then binds to the host cell. Following that, the virus enters the host cell, and then HIV reverse transcriptase helps the genome RNA convert to DNA. After that, the genome DNA is transported into the nucleus and HIV integrase helps integrate it into the host DNA.

In the late phase, HIV genomic materials and messenger RNA (mRNA) are created by the host cell RNA polymerase. Using these mRNAs, HIV polyproteins are translated. Then during budding, an outer envelope coats the new virus particles, and the new coated virus moves outside of the host cell. In the last step, maturation, HIV protease cleaves the HIV polyproteins into small pieces, and synthesizes the matured HIV virions, which are able to infect other healthy cells. All the steps are shown in Figure 3.

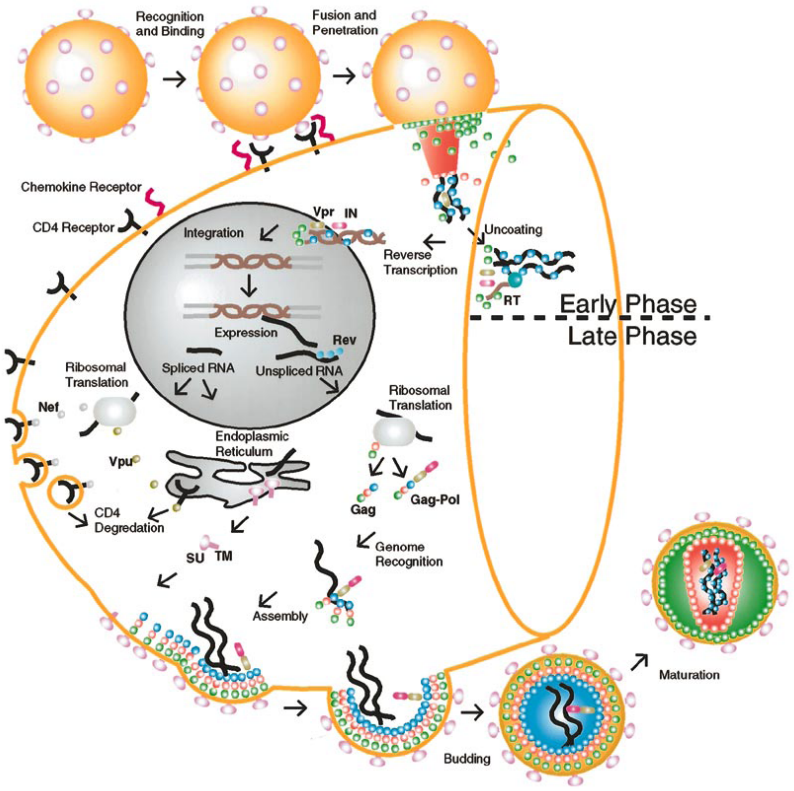


Figure 2.1.1.1 life cycle[11]

2.1.2 HIV-1 protease and its inhibitors

Among all the HIV-1 proteins, the structure of HIV-1 protease was first determined in 1989 [12, 13]. It's a homodimer with two identical subunits, and each one has 99 amino acids. The structure of the HIV-1 protease could be considered as three parts: dimer interface, active site and flap region (as shown in Figure 4). The dimer interface connects two subunits, and helps to stabilize the structure of the HIV-1 protease. The active site cavity is the place where the inhibitors bind to the HIV-1 protease. The substrate binding is connected via hydrogen bonds and van der Waals interactions. The flap region is flexible and could change the conformation easily and is very important for the enzymatic activity of the HIV-1 protease. Such character could enhance the binding between the protease and the inhibitor (or substrate) at the active site[14]: without the inhibitor binding to the active site, the flaps are slightly open. Once the inhibitor binds to the protease, the flaps could fold down to improve the protease-inhibitor binding.

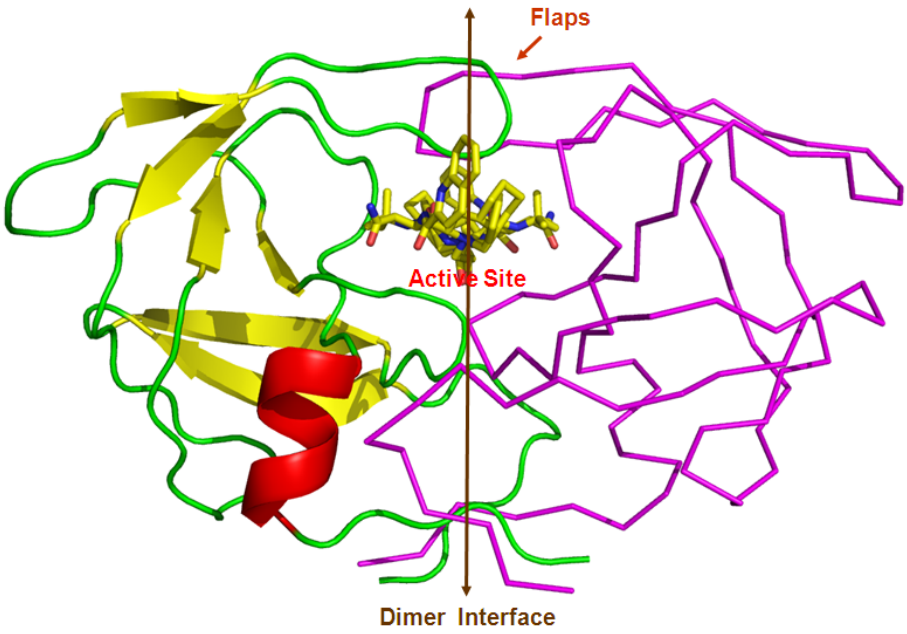


Figure 2.1.2.1 Structure of HIV protease dimer with saquinavir inside the active site. The α-helix is in red; the β-sheet is in yellow arrow in the left subunit. The right subunit is in magenta.

HIV-1 protease inhibitors (PIs) were developed to bind to the active site, and prevent the maturation of the virions. In this case, the newly synthesized viruses are unable to infect other cells. Since HIV-1 protease is crucial for the maturation of the HIV-1 polyproteins by catalyzing the hydrolysis of certain peptide bonds in them[15], the inhibitors of HIV protease have proved to be effective anti-viral drugs[16].

The first PI was developed in year 1995, and after applying this treatment to the patients, the HIV death rate has decreased sharply[17] and the lifetime of the AIDS patients has been increased[8]. Up till now, a total of ten PIs have been approved by the US Food and Drug Administration (FDA). They are saquinavir, ritonavir, indinavir, nelfinavir, (fos)amprenavir, lopinavir, atazanavir, tipranavir and darunavir, listed chronologically by the FDA approval date. These PIs bind in the active site of HIV protease, and prevent the cleavage of the virus polyproteins. Therefore, the viruses cannot form mature particles to infect other host cells[18, 19].

2.1.3 HIV-1 reverse transcriptase and its inhibitors

The HIV-1 reverse transcriptase helps synthesizing the DNA based on the information given by mRNA using either RNA-dependent DNA polymerase or DNA-dependent DNA polymerase. The structure of HIV-1 reverse transcriptase was determined in 1995 at resolution 2.35 Å[20], 2.7 Å[21], and 3.2 Å[22]. HIV-1 reverse transcriptase is a dimer with two different monomers: one is p66 with the length of 560 residues; and the other one is p51 with the length of 440 residues, and the structure is shown in Figure 5[22, 23].

The sequence of p51 is identical to the first 440 residues in p66; however, they differ variously in structural conformation. The structure of p66 is often considered as illustration of the right hand[22] and includes the fingers, a palm, a thumb and a RNAseH, which is the residues 441-660[23]. The polymerase active site is inside the palm region, and contains three aspartic acids, similar to that inside the HIV-1 protease active site. These three aspartic acids help to binding the polymerase to the active site[24]. The structure of p51 has no enzymatic function, but helps to stabilize the structure of p66[25].

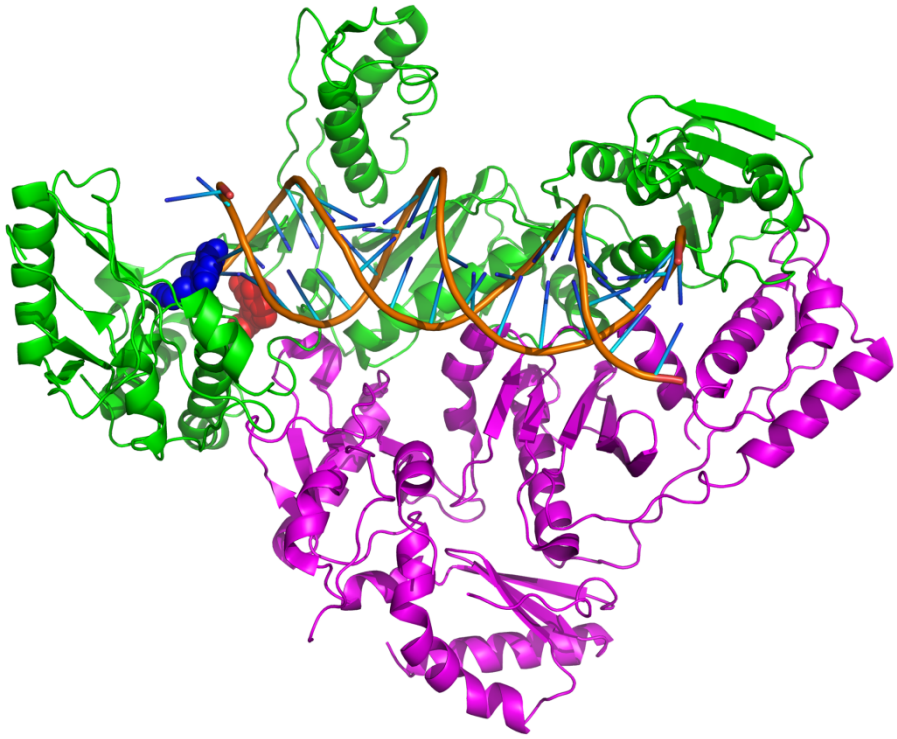


Figure 2.1.3.1 Domain structure of the HIV-1 reverse transcriptase. The structure of HIV-1 RT dimer in complex with DNA and bound NNRTI and NRTI from [26, 27]. The p66 subunit is shown in green and the p51 subunit is shown in purple. NRTI is shown in blue, and NNRTI is shown in red. Double stranded DNA is shown in orange.

There are two classes of HIV-1 reverse transcriptase inhibitors: Nucleotide analog reverse transcriptase inhibitors (NtARTIs or NtRTIs) and Non-nucleoside reverse transcriptase inhibitors (NNRTIs). NtRTIs are structural analogs, and it mimics the function and mechanisms of the natural substrate of the enzymes. By competing with the natural substrate, it could incorporate with the newly synthesized DNA. Because of this, NtRTIs are not HIV-1 reverse transcriptase specific, and could be used to other antiviruses, for instance, HIV-2, SIV, murine leukemia virus, visna virus, etc[28, 29]. NNRTIs class is more specific compared to NtRTIs class, and target to HIV-1 reverse transcriptase. The inhibitors in this class could bind to the HIV-1 reverse transcriptase in the palm domain of the p66 sub-unit. Such interaction could

reduce the enzymatic activities of the HIV-1 reverse transcriptase[30]. The different characteristics of these two categories are summarized in Table 2[31].

Table 2.1.3.1 Characteristics of HIV-1 RT inhibitors: NtRTIs and NNRTIs[31].

Characteristics	NtRTIs	NNRTIs
Chemical structure	Analogs of the natural substrates, i.e. nucleosides	Chemically diverse, non-nucleoside
Active form	Metabolic conversion to 5'-triphosphates by host-cell enzymes	No metabolic conversion
Mechanism of action	Incorporate into growing DNA chain, terminate chain synthesis	Induce conformational changes in RT, reducing catalytic activities
Type of inhibition	Competitive with the natural substrates (dNTPs)	Non-competitive/uncompetitive
Binding site on the RT	Catalytic site	Allosteric (non-substrate) hydrophobic pocket
Spectrum	Broad spectrum antiretrovirals	HIV-1 specific RT inhibitors
Selectivity	Low to moderate	Very high

.

2.2 Drug Resistance

2.2.1 HIV-1 protease and reverse transcriptase drug resistance

Due to the lack of proofreading[32, 33] and high mutation rate[34, 35], mutations are commonly seen in HIV-1 genome[36]. Drug resistance occurs during the treatment of the AIDS, which may cause the failure of the treatment. Surveys using conventional bulk sequencing in North America and Europe show that for the untreated patients, the primary drug resistance rate is 8-20%[10]. Most of the mutations may decrease the susceptibility to certain drugs; however, in some rare cases, certain mutations may increase the drug efficiency, for instance N88S could increase the susceptibility of FPV[36]. HIV-1 PI-resistant mutations were found in the active site, dimer interface, flap region as well as the surface of protease. Currently, twenty five or more residues out of ninety nine have been found in PI-resistance (as shown in Figure 6)[16].

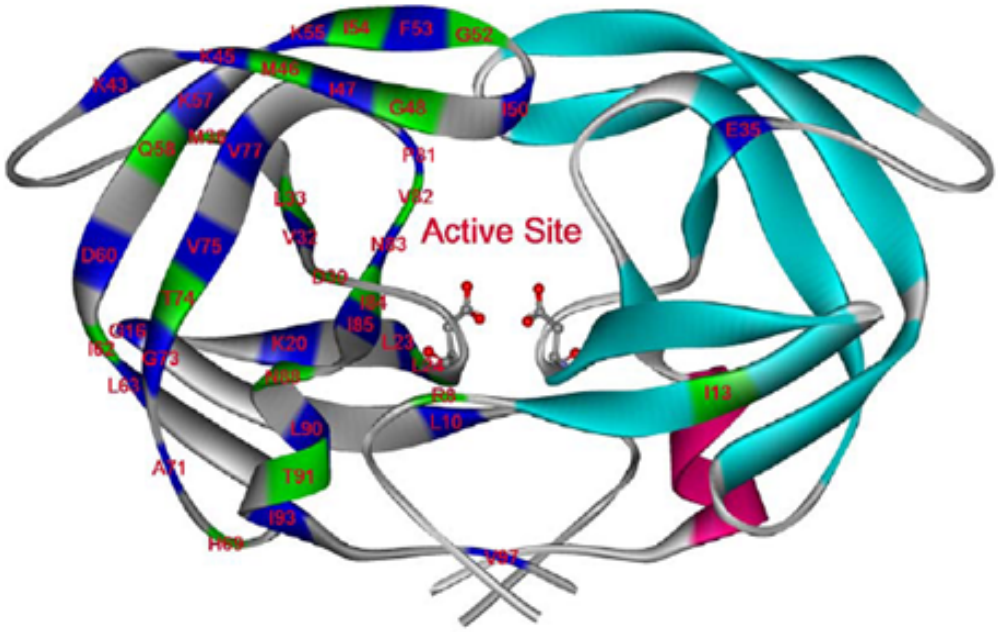


Figure 2.2.1.1 Crystal structure of HIV protease with sites of drug mutation. The active site aspartic acid residues (Asp25) of each monomer are shown in a stick representation. Positions of drug-resistant mutations are indicated in blue and green.[14]

Currently, there are three proposed mechanisms for the drug resistance of HIV-1 protease inhibitors: one is that, because of the mutations, the structural conformation of the HIV-1 protease changes, and therefore directly affects the interactions between the inhibitors or substrate and the HIV-1 protease at the active site[37]. The second one is that those mutations indirectly change the ability of the protease to bind inhibitor[38, 39]. The third one is that the mutations at the dimer interface may decrease the stability of the protease [38-40], and thus weaken the enzymatic function of the HIV-1 protease.

The drug resistance is also found for HIV-1 reverse transcriptase inhibitors in both NtRTIs and NNRTIs. Almost all the NtRTIs mutations were found to alter a direct interaction to the active site of the enzyme[41]. Over 40 amino acid mutations are found in the NNRTIs related mutants (as shown in Figure 7)[42, 43], and more detailed explanation could be obtained in the review[10]. The mutants are found in the palm region[43, 44], p51 sub-unit[45], connection domain of p66 sub-unit[46], between the thumb region[47], as well as the RNAseH domains[48]. One mechanism of the drug resistance is that most the

mutations could decrease the RNAseH cleavage activities; while in some rare cases, such enzymatic activities may increase due to the mutations[49].

NNRTI RAM	Prevalence in Samples for RCRT (%)	Prevalence in NNRTI-Resistant Samples (%)	Number of Nucleotide Substitutions (Codon Change)	FC[b] in SDMs							Reference(s) in which the association with NNRTI resistance was demonstrated
				EFV	NVP	ETR	TMC278[d]	RDEA806[e]	IDX899[f]	UK453061[g]	
V90I	4.68	6.84	1 (GTT → ATT)	1.7	3.8	1.5	NA	NA	0.9	NA	Vingerhoets et al. (2005, 2007)
A98G	3.46	7.76	1 (GCA → GGA)	2.2	8.1	2.5	NA	NA	NA	NA	Byrnes et al. (1993), Bacheler et al. (2000)
L100I	2.96	6.92	1 (TTA → ATA)	13.1	NA	1.8	0.8	NA	0.4	2.9	Johnson et al. (2007), Mellors et al. (1993)
K101E	3.89	9.01	1 (AAA → GAA)	2.9	4.3	1.7	2.5	NA	NA	8.7	Byrnes et al. (1993), Bacheler et al. (2000)
K101P	0.87	2.04	2 (AAA → CCA)	97.4	>733.4	6.2	46.5	NA	NA	NA	Rhee et al. (2004)
K101Q	3.02	5.95	1 (AAA → CAA)	3.8	NA	3.4	1.4	NA	NA	NA	Bacheler et al. (2000), Kleim et al. (1999), Ceccherini-Silberstein et al. (2007)
K103H	0.08	0.18	2 (AAA → CAC)	15.6	17.8	1.6	NA	NA	NA	NA	Harrigan et al. (2005)
K103N	24.34	56.96	1 (AAA → AAC)	26.7	56.2	0.7	0.8	0.5	1.1	NA	Johnson et al. (2007), Nunberg et al. (1991)
K103S	1.11	2.6	2 (AAA → AGC)	4.9	36.2	0.9	1.2	NA	NA	NA	Rhee et al. (2004), Kleim et al. (1999)
K103T	0.07	0.17	1 (AAA → ACA)	1.4	>37.0	1.3	NA	NA	NA	NA	Rhee et al. (2004)
V106A	1.01	2.37	1 (GTA → GCA)	1.8	86.1	0.5	0.7	NA	NA	3.1	Johnson et al. (2007), Larder (1992)
V106I	2.62	3.47	1 (GTA → ATA)	NA	NA	NA	NA	NA	NA	NA	Vingerhoets et al. (2007), Kleim et al. (1997), Taylor et al. (2000)
V106M	0.49	1.15	2 (GTA → ATG)	2.3	6.9	0.8	1	NA	NA	NA	Johnson et al. (2007), Loemba et al. (2002)
V108I	4.81	10.68	1 (GTA → ATA)	1.2	2.7	0.5	NA	NA	NA	4.9	Johnson et al. (2007), Byrnes et al. (1993)
E138G	0.42	0.8	1 (GAG → GGG)	2.3	NA	3.8	NA	NA	NA	NA	Pelemans et al. (2001)
E138K	0.34	0.5	1 (GAG → AAG)	1.8	1.7	2.4	2.2	NA	0.8	5.8	Balzarini et al. (1994)
E138Q	0.42	0.91	1 (GAG → CAG)	7.1	NA	5.1	NA	NA	NA	NA	Pelemans et al. (2001), McCreedy et al. (1999)
V179D	1.75	3.25	1 (GTT → GAT)	6.2	5.7	2.6	1.6	NA	NA	NA	Byrnes et al. (1993), Palmer et al. (2003)
V179E	0.75	1.47	2 (GTT → GAG)	4	2.6	1.1	1.1	NA	NA	NA	Byrnes et al. (1993), Vingerhoets et al. (2006)
V179F	0.15	0.34	1 (GTT → TTT)	<0.4	1.6	0.1	0.1	NA	NA	NA	Vingerhoets et al. (2005, 2006)
V179G	0.08	0.15	1 (GTT → GGT)	0.6	NA	0.6	NA	NA	NA	NA	Miller et al. (2003)
V179I	8.55	11.82	1 (GTT → ATT)	0.9	1.3	0.8	NA	NA	NA	NA	Turner et al. (2004)
Y181C	10.66	24.95	1 (TAT → TGT)	2.2	207.6	3.9	2.6	3.2	2.7	2.2	Johnson et al. (2007), Nunberg et al. (1991)
Y181I	0.45	1.05	2 (TAT → ATT)	1.6	>55.5	12.5	14.3	NA	11.5	0.8	Johnson et al. (2007), Shih et al. (1991)
Y181V	0.27	0.63	2 (TAT → GTT)	2.8	2155.9	17.4	12.6	NA	NA	NA	Vingerhoets et al. (2006), Shih et al. (1991)
Y188C	0.18	0.43	1 (TAT → TGT)	2.1	36.5	0.2	NA	NA	NA	0.3	Johnson et al. (2007), Richman (1993)
Y188H	0.44	1.02	1 (TAT → CAT)	7.6	5.5	0.3	NA	NA	NA	NA	Johnson et al. (2007), Sardana et al. (1992)
Y188L	2.72	6.37	2 (TAT → CTT)	31.9	173.4	0.9	2.8	6.3	NA	NA	Johnson et al. (2007), Shih et al. (1991)
V189I	1.73	2.79	1 (GTA → ATA)	1.2	2.9	0.8	NA	NA	NA	NA	Vingerhoets et al. (2007), Kleim et al. (1996)
G190A	8.16	19.09	1 (GGA → GCA)	6.8	105	0.8	1	NA	0.4	NA	Johnson et al. (2007), Bacolla et al. (1993)
G190C	0.07	0.17	2 (GGA → TGC)	NA	NA	NA	NA	NA	NA	NA	Huang et al. (2003)
G190E	0.29	0.67	1 (GGA → GAA)	NA	NA	NA	NA	NA	NA	NA	Bacheler et al. (2000), Kleim et al. (1993)
G190Q	0.2	0.47	2 (GGA → CAA)	NA	NA	NA	NA	NA	NA	NA	Huang et al. (2003), Kleim et al. (1994)
G190S	1.66	3.89	2 (GGA → AGC)	94.4	177.1	0.2	0.2	NA	NA	NA	Johnson et al. (2007), Kleim et al. (1994)
H221Y	3.75	8.34	1 (CAT → TAT)	2.4	4.4	2.5	NA	NA	NA	NA	Ceccherini-Silberstein et al. (2007), Gonzales et al. (2003), Saracino et al. (2006), Perno et al. (2006)
P225H	2.34	5.38	1 (CCT → CAT)	2.2	2.8	1	NA	NA	0.3	NA	Johnson et al. (2007), Bacheler et al. (2000), Kleim et al. (1999)
F227C[c]	0.01	0.02	1 (TTC → TGC)	27.2	24.3	3.6	NA	NA	NA	56.4	Andries et al. (2004)
F227L	0.97	2.24	1 (TTC → CTC)	0.7	2.9	0.4	NA	NA	NA	6.3	Balzarini et al. (1994), Parkin et al. (2000)
M230I	0.08	0.19	1 (ATG → ATA)	4.7	13.1	2.4	NA	NA	NA	NA	Kleim et al. (1999)
M230L	0.5	1.16	1 (ATG → CTG)	5.7	13.9	3.4	2.5	NA	1.5	NA	Rhee et al. (2004), Huang et al. (2003)
P236L	0.09	0.21	1 (CCT → CTT)	2.4	4.6	1.3	NA	NA	NA	0.3	Johnson et al. (2007), Dueweke et al. (1993a,b)
K238N	0.35	0.76	1 (AAA → AAC)	2.7	NA	1.7	NA	NA	NA	NA	The Stanford University HIV Drug Resistance Database (2007)
K238T	1.79	4.11	1 (AAA → ACA)	3.4	NA	2.4	NA	NA	NA	NA	Rhee et al. (2004), Demeter et al. (1998)
Y318F	0.85	1.97	1 (TAT → TTT)	0.6	1.5	1.4	NA	NA	NA	NA	Pelemans et al. (1998), Harrigan et al. (2002)

Abbreviations: NNRTI: non-nucleoside reverse transcriptase inhibitor; RAM: resistance-associated mutation; RCRT: routine clinical resistance testing; FC: fold change in 50% effective concentration; SDM: site-directed mutant; EFV: efavirenz; NVP: nevirapine; ETR: etravirine; NA: not available.

[a] Adapted from Tambuyzer et al. (2009).
[b] FC values were determined for EFV, NVP, ETR and TMC278 as described in Vingerhoets et al. (2005). Results were reported as median values from 2 or more measurements.
[c] This SDM also contained the L234I mutation in the RT.
[d] FC values for TMC278 were extracted from Rimsky et al. (2009).
[e] FC values for RDEA806 were calculated from data presented in Girardet et al. (2007).
[f] FC values for IDX899 were calculated from data presented in Jakubik et al. (2008).
[g] FC values for UK453061 were extracted from Corbau et al. (2007).

Figure 2.2.1.2 Common NNRTI resistance associated mutations, and their impact on the susceptibility of HIV-1 to NNRTIs[10, 42, 43].

3 JUSTIFICATION OF THE RESEARCH TOPICS

3.1 Literature review on computational HIV/AIDS drug resistance prediction

As mentioned in Chapter 2, due to the high replication rate and no proofreading mechanism, resistant strains are commonly seen during the HIV/AIDS treatment. Because the importance of each mutant is not equal to the drug resistance, and the mutation pattern is difficult to retrieve [50]. Moreover, even though the measurement of the genotype isolates obtained from the patients could determine the relative resistance to certain drugs using the genotypic and phenotypic assays[51], such expensive experiments may take up to two weeks to complete. Furthermore, due to the huge existing data nowadays, it is more convenient to introduce computer methods to predict the relative resistance to certain drugs, whose results could consider as the reference during the AIDS treatment, to shorten the assay time and provide a more rapid, cheap and proper treatment to the patients. Under this circumstance, predicting the phenotypes from the genotypes is a crucial research topic and many different kinds of methods have been used to solve this problem.

3.1.1 Genotypic-resistance interpretation systems

The genotypic interpretation algorithms, which could be considered as the knowledge based methods, are also used to predict the drug resistance. These kinds of algorithms either use a set of rules or a score of 'penalty' for each drug, which is provided by groups of HIV experts. The input of the algorithms is the list of mutations, while the output is the drug resistance categories, for example susceptible, resistance, intermediate, or others[52]. The output categories of each algorithm differ from each other. Following two paragraphs present the details of each algorithm in both rule-based and score-based systems.

The first kind of algorithms is also considered as the rule-based systems that infer drug-resistance levels from sequence information such as Agence Nationale de Recherches sur le SIDA (ANRS)[53], Rega Institute version 5.5 (Rega-5.5)[54] and Visible Genetics version 6 (Toronto, Ontario, Canada) (VGI-6)[52]. All these three algorithms report three levels of the resistance: susceptible, resistant, and intermediate. The rules used in these algorithms are sets of the Boolean expressions. These are "designed to provide reasonable interpretations for the large number of remaining possible mutation combinations"[52].

The Stanford University HIV Drug Resistance Database (Stanford HIVdb)[55] and mutation rate based score[56] are examples of the second kind of the algorithms. The HIVdb algorithm reports a total of five levels of the resistance: susceptible, potential low-level resistance, low-level resistance, intermediate resistance, and high-level resistance. The penalty score used in the algorithm is defined as follows: for each mutation, a drug penalty score is assigned by the algorithm. Then to determine the drug resistance category, the total scores are added and the sum is used to infer the final result.

Also, a combined rule-based and penalty-based method named AntiRetroScan (ANS) is proposed and applied to both HIV-1 protease and reverse transcriptase inhibitors[57]. This system is developed at the University of Sienna and is maintained on the Italian Antiretroviral Resistance Cohort Analysis Website. More frequently used genotypic-resistance interpretation systems are discussed and reviewed in[58].

The use of such genotypic resistance interpretation system helps the physicians during the treatment, and had better outcomes comparing to those who didn't use it[59]. However, these methods provide little insight on the genetic and molecular basis of drug resistance and often give inconsistent results when analyzing the same input mutation data[52, 58]. Furthermore, because different algorithms use different rules, the outcome of drug resistance levels and the approach to deal with the data shortage, produces the inter-algorithm discordance[52, 58]. Moreover, it is a tedious work for experts to

provide the mutation information for each mutant. Meanwhile, due to the HIV-1's high mutation rate, it is difficult to interpret in such a time and energy consuming approach. Therefore, bioinformatics-assisted anti-HIV therapy is needed and developed in a rapid speed.

3.1.2 Bioinformatics-assisted anti-HIV therapy

As mentioned in the last section, bioinformatics-assisted anti-HIV therapy is needed, and such algorithms have several advantages compared to the traditional systems: First, the results given by this kind of algorithms are more global and quantitative. Comparing to the results given by the experts, such results are less subjective. Second, the constructed computational models could be used for different data sets, and therefore limit the potential bias. Third, it is difficult for humans to deal with large numbers of variables, but computers are good at it. Computational approaches are good at revealing the patterns between the mutations.[60]

In most common case, the input of the bioinformatics methods is the viral genotype; while the output of the algorithms is the resistance values of the virus to certain drug/inhibitor. The general procedure is as follows: first, the algorithm study the training data set of both the input and their correspond output; then by using statistical, classification, or other algorithms, a computational model is learned and constructed for these data; finally, at the last step, a new viral genotype is given to the model, and the predicted resistance value is generated by the model. From this predicted resistance value, the given genotype could be predicted as drug resistant or not, or somewhere in between, to certain drug. The different algorithms used to construct the computational model could further classify as the statistical learning methods, classification methods together with the molecular structure based methods. The details of these methods are discussed in the following sections.

3.1.2.1 *Statistical methods*

In the past decades, many statistical learning methods have been introduced in predicting the phenotypes from the genotypes [52, 58, 61-63]. These methods could be treated as the regression problems, and the resistance values are directly predicted. The cross-validation is included to assess the performance of the algorithms. A reliable algorithm should have the squared correlation coefficients and mean-squared errors between the measured and the predicted resistance value between 0.7 to 0.8, and 0.2 to 0.3, respectively[60].

In [50], Bayesian variable partition model is used to detect resistant mutation combinations and find the interaction patterns of drug resistance to certain inhibitors. Following that, molecular dynamics (MD) is introduced to explain how these mutations interact with each other on molecular basis.

In [64], linear regression model is used to predict the in vitro susceptibility phenotype and virology response during the treatment. The most significant mutations and interactions are given, and a high concordance with in vitro measurement is presented.

In [65], cluster analysis, recursive partitioning, and linear discriminant analysis are applied on Adult AIDS Clinical Trials Group (ACTG) protocol 333. The results from the three methods show in consistence that residues 10, 63, 71, and 90 have in vitro resistance to IDV and SQV. Similarly, in [66, 67], existing cluster analysis, discriminant analysis, and recursive partitioning techniques are used to construct the model and test on IDV.

Also, non-parametric methods are proposed to solve these high dimensionality data[68, 69].

3.1.2.2 *Computational classification techniques*

Despite statistical learning methods, classification algorithms could also be used to solve this problem. By using this kind of methods, a resistance-factor cutoffs[63] is needed to categorize whether each mutant is drug resistant or not. A reliable algorithm should have the errors rates below 10% [60].

In [62, 70, 71], the Geno2pheno system uses decision trees and support vector machines to predict the phenotypic drug resistance values. The output of it is the normalized predicted resistance value, together with the observed fold-changes among the untreated patients.

In [72], artificial neural networks (ANN) was used to train, validate, and test on 1322 clinical samples, and two neural network models were established. The result shows that the predictor has the correlation coefficient with R^2=0.88. In the same year, in [73], ANN was also used to test on SQV and IDV, with the accuracy of 60%-70%.

3.1.2.3 *Molecular structure based methods*

Fundamentally, the HIV drug resistance is caused by the change of the structure and the enzymes' drug target sites. The molecular structure can also be used to predict the drug resistance value to the mutations to certain drug/inhibitor. This approach includes the molecular docking methods, the homology-based modeling methods[74], as well as the molecular dynamics simulations[75].

The computational structure-based methods used in molecular modeling are often used for structure optimization and scoring ligand-protein docking structure. Such procedures are similar to the drug resistance prediction, and could be used to solve this problem [74, 76, 77].

Combined sequence-structure approaches are also included to solve the problem: a Delaunay tessellation derived four-body statistical potential mutagenesis method together with support vector machine (SVM) and random forest classification methods is applied to predict the drug resistance for HIV-1 reverse transcriptase inhibitor, Nevirapine (NVP) [78] and more inhibitors later [79]. More detailed information about the structure-based phenotyping is discussed in[76, 80].

3.2 Literature review on sparse representation

In recent years, the compressive sensing/sparse representation[81], paper[82] provides a nice framework for the purpose of combining capacity and efficiency and solving the dilemma between

learning capacity and efficiency. In the sparse representation theory, it is observed that the natural (one or more dimensional) signals are often sparse when represented in certain non-adaptive basis or tight frames. As a result, compressive sensing has been employed to show that a very large class of signals can be accurately (or in some cases exactly) reconstructed from far fewer samples than suggested by conventional sampling theory. Classical signal processing techniques lead to sufficient sampling by employing the band-limitedness of signals. In the compressive sensing approach, one defines sufficient sampling conditions based on the compressibility of a signal relative to a given dictionary designed for the problem at hand. From an opposite perspective, given a set of signals of interest, an over-complete dictionary can be constructed so that the signals can be represented sparsely[83]. In particular, the idea of sparse representation has now drawn much attention in image restoration[84], denoising[85], deblurring[86], signal processing[87], face detection[88], texture modeling[89-92], etc. In them, the redundancy in the over-complete dictionary gives rise to the sparse representation which enables both the efficiency in processing and the capacity of handling highly complex large data sets.

3.3 Mean shift

Mean shift clustering was first introduced in 1975 by Fukunaga and Hostetler[93] with the purpose of seeking the mode of a density function in the given sample set. Fukunaga and Hostetler[93] also suggested that mean shift clustering is an instance of gradient ascent by using decreasing distance functions, which often referred as kernel, from a given point to a point in the sample set. This algorithm was not widely used until 1995 when Cheng[94] developed a more generalized formulation of the algorithm. By clarifying the relationship between mean shift and the optimization, the algorithm could potentially be applied on clustering and global optimization problems. Applications of the mean shift algorithm range from image/video segmentation, image representation/retrieval, discontinuity-preserving smoothing[95, 96], higher level tasks like appearance-based clustering[97, 98], tracking including blob tracking[99] and face tracking[100], shape detection and recognition[101], so on and so forth. After-

wards, applications extend to other fields like biology. These applications include analysis of structural variation in genome[102], DNA microarray analysis[103], time-warped gene expression analysis[104], with many other implementations.

AIM 1: Developing a new encoding algorithm to retrieve the protein structure information

4 Encoding Protein Structure with Functions on Graphs[105]

4.1 Abstract

The application of machine learning and datamining to the analysis and prediction of protein structure is a research area with potentially high impact in both computer science and biology. Proteins structures are inherently complicated objects with a mixture of crisp and fuzzy properties. Therefore developing effective representations for them is a research problem in itself, while quantifying and predicting properties and structure is of immediate importance in structural biology. This paper focuses on developing a compact, effective, efficient and accurate representation of protein structure that is compatible with widely used machine learning tools like the SVM. Graphs based on Delaunay triangulation are used to represent the structure, and then functions are constructed from these graphs to develop constant-size representations of protein structure that are tightly bound to the amino acid sequence. The representations preserve sufficient information to be valuable for model vs. experimental structure classification and regression analysis of model quality.

4.2 Introduction

The accurate and predictive association of protein sequence, protein structure and protein function is one of the “holy grails” of structural bioinformatics. Developing effective and efficient encoding of protein structure is a necessary step towards achieving this aim. In this paper, we develop a novel class of encoding algorithms, based on Delaunay and related triangulations, for protein and other complicated three-dimensional objects, which are highly effective and efficient. Typically an atom or subset of atoms or centroid of atoms is chosen as a fiducial marker per amino acid residue. These fiducial markers, after encoding, are then used as input for machine learning or data mining analyses. If there

are N fiducial markers then there are $O(N^2)$ distances between them. For example, Figure 8 shows the distances for the protein with the pdb id 2b0v[106]. Therefore a naïve representation of the distances can result in a large and highly variable representation which may make machine learning more difficult than it needs to be.

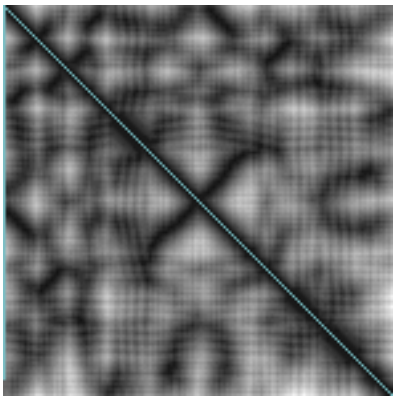


Figure 3.1.2.1 Distance plot of 2B0V. The distance between two alpha-carbon atoms is plotted in this gray-scale image. These are the raw data that will be compressed for machine learning. Darker areas are closer in space than lighter areas. Note that some residues that are quite far apart in sequence space are close together in 3-dimensions.

To be an effective measure for machine learning the representation must be constant-size thus minimizing the possibility of spurious feature detection (for example learning to discriminate between model structures of 80 residues and experimental structures of 81 residues), suppressing irrelevant features, and emphasizing important features of the un-encoded data.

A simple and widely used way to encode would be to select a sliding window of some width, typically about 20 residues, and then train on instances of the window from model and experimental structures [for example see [107, 108]]. Protein structures consist of stretches of highly regular structure such as alpha-helices and beta sheets, with more variable turn and loop conformations connecting them. The alignment of a sliding window with respect to these regular and variable features is likely to have an impact on training accuracy, and unfortunately it is not trivial to predict the locations of these features from sequence data alone. More importantly, using a banded representation discards infor-

mation about residues that are close in 3-dimensional space, but are distant in sequence. Figure 2 shows how the data are limited by this approximation.

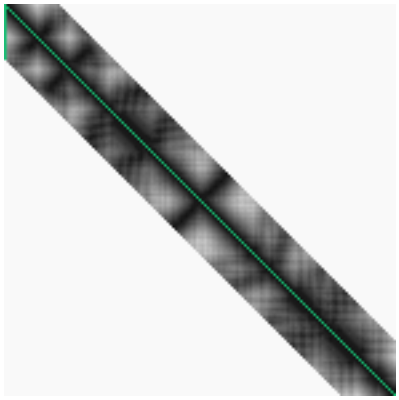


Figure 3.1.2.2 Selecting a 20 residue wide "banded" encoding for 2B0V results in this image. Note that all of the long-range information in Figure 1 has been deleted.

It is precisely this information that determines the fold of the protein. Finally, if we train the machine learning approach against individual small pieces of protein, then we need to define an algorithm for combining the scores across a variable number of smaller pieces of protein. Defining a general voting algorithm for variable numbers of data can be problematic. Therefore it is important to find encodings that can handle the whole protein fold, rather than pieces of it.

Our encoding begins by calculating a graph based on critical contacts within a protein. Following Richard's use of Voronoi tessellation[109], we used the dual of the tessellation, Delaunay triangulation, to define a unique graph for each protein structure. This is simpler than directly using volumes or surface areas derived from Voronoi tesselation[110], but is still a fully rigorous description of protein structure. Since Delaunay triangulation can be expensive to calculate, we also tested a "defective Delaunay" triangulation that is sparser than the Delaunay triangulation, faster to calculate, and which reproduces many of the same features. As a control we also assessed the performance of distance cutoff based triangulation to test the importance of local structural geometry in defining an accurate encoding. For this work we used the alpha-carbon atom as a fiducial, clearly other atoms or centroids could be used, but as

every residue possesses an alpha-carbon this is sufficient for evaluating the effects of differences in the functions and graphs. The graphs are symmetric and undirected. A weight consisting of the two kinds of amino acid and the distance between them is associated with each non-zero element of the adjacency matrix or arc of the graph. An example of an adjacency matrix is shown in figure 3, where the adjacent points (those that would be non-zero in the matrix) are shown.

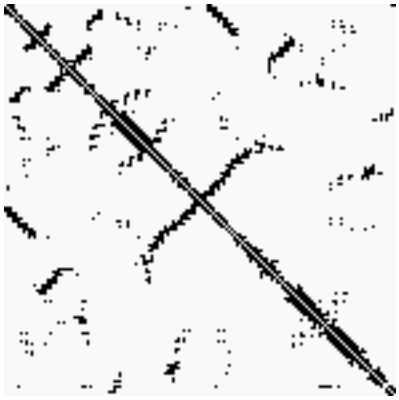


Figure 3.1.2.3 The adjacency matrix for Delaunay triangulation of 2B0V is shown here. The distances for points that are colored in this figure are selected from figure 1. Note that features that are distant in sequence space but adjacent in 3-dimensional space are selected. While similar to the features that are "close" in figure 1, the Delaunay triangulation selects a subset of the "close" distances as well as some longer-range distances.

These graphs are an intermediate representation, of size O(N) instead of $O(N^2)$, and therefore still have a size dependence, but are already more space-efficient than the naïve approach. Functions are then calculated from these graphs that contain both sequence and structure information. We evaluated five functions, 1) the average distance per kind of arc (210 features corresponding to each unique pair of amino acids), 2) total distance per kind of arc (210 features), 3) number of instances of any given kind of arc (210 features), 4) frequency of each kind of arc (210 features), and 5) the Cartesian product of average distance and number of instances (420 features).

One of the difficulties in assessing machine learning encodings is differentiating between the effect of the tuning and selection of the machine learning tool and the effect of the representation of the data on the accuracy. In order to remove this variability and to ensure that the differences in encoding

were reflected rather than our ability to choose parameters for a software tool, a single SVM engine, svm_light[111] was used with a linear kernel and default parameters. In the future we can tune parameters and choose other machine learning tools and to improve the accuracy of the classification and regression. The focus of this paper is the comparison between different representations and therefore we did not vary the machine learning approach as that would invalidate the comparison. We also believe that it is important to demonstrate that the encoding is sufficiently linear to work with simpler machine learning tools. Svm_light was able to classify and regress the data with other kernels like polynomial and radial basis kernels, but the linear kernel worked well and therefore was used.

4.3 Methods

4.3.1 Datasets:

For classification a set of 1447 protein structures with internal sequence identities of less than 25% was downloaded from the Pisces culling server[112]. Benchmark data sets were generated by shifting the sequence by one residue and by reversing the sequence. Small sequence shifts are typical in low-identity homology models[113] and protein structures have been determined with the sequence completely reversed (in error) so these benchmarks are representative of realistic errors.

For regression analysis the MOULDER benchmark suite defined by[114] was downloaded. This dataset consists of 20 individual proteins with 300 miss-aligned homology models each. Both RMS error and the fraction of residues within 3.5Å of correct positions are associated with each data point in the MOULDER dataset. We found that regression against the fraction of residues within 3.5Å of correct positions performed much better than regression against RMS errors, which reflects the fragility of RMS as a measure of model quality [115].

4.4 Delaunay Triangulation:

The Delaunay triangulation is defined by sets of points which lie on a sphere with no other points within that sphere[116, 117]. The naïve direct iteration algorithm was used to identify the triangulation, although inter-atomic distances larger than 10Å were excluded to speed it up. Since van derWaals contacts and hydrogen bonds, the closest non-bonded distances, are much shorter than 10Å and the structures are densely packed, the use of such a cutoff is justified on chemical and structural grounds.

4.5 Defective Delaunay Triangulation:

An approximate Delaunay triangulation can be performed by finding the closest atom to a given atom and then using the plane of the perpendicular bisector to eliminate atoms that are further away. This is then applied recursively until all the atoms are either excluded or identified as contacts. This algorithm produces an asymmetric graph, so the graph was forced to be symmetric by requiring that all atoms identified as belonging to the contact set of atom A, had atom A as a member of their contact set. This produces a sparse subset of the Delaunay graph. It also produces a convex hull around the central atom, although not necessarily the smallest convex hull.

4.6 Distance Only Triangulation:

In order to demonstrate the importance of using a triangulation algorithm, rather than a simple distance cutoff, a limited number of calculations were performed using a distance-based triangulation. The distances between all pairs of amino acids in the decoy and experimental structures were calculated and if the distance was less than 6Å the pair was added to the graph. 6Å was chosen based on the distribution of distances in the Delaunay contacts, as most of the Delaunay contacts were shorter than this value.

4.7 Machine Learning:

The tool svm_light was downloaded from http://svmlight.joachims.org/, compiled and used. N-fold cross validation tests were performed in addition to the leave one out tests implemented in svm_light. Care was taken to insure that all positive and negative instances of a given protein were removed from either a training or testing dataset when generating a set for cross-validation. This avoided the potential problem of having negative instances associated with a positive test item or positive instances associated with a negative test item and thus generating systematically optimistic (and incorrect) assessments of the training accuracy.

Five related functions were calculated from the triangulations. Since the adjacency matrices associated with each of the triangulations are symmetric, there are 210 unique pairs of amino acids. The sum of the distances for each kind of pair and the numbers of each kind of pair were directly summed from the adjacency matrices. Normalizing the sum of distances by the numbers of each kind produced the average distance, and normalizing the numbers of each kind by the total number of arcs produced a frequency measure. Finally, appending the numbers of each kind to the average distance produced a Cartesian product that was useful for probing the importance of normalization.

4.8 Results

4.8.1 Classification

The classification accuracy, assessed with 5-fold cross-validation, on the shifted and reversed sequence benchmarks is shown in table 1. The variance between samples ranged between 0.5 and 1.5% indicating the magnitude of difference that is significant.

Table 4.8.1.1 Classification results in percent. Results are from 5-fold cross-validation. The abbreviation in the model type dd stands for defective Delaunay, and the abbrevation cl stands for close where the graph was selected purely on distance criteria. Shift refers to using the decoys where the sequence has been shifted by one residue, and reverse refers to the decoys where the sequence has been reversed. Accuracy is (TP+TN)/(all data). Precision is (TP)/(FP+TP). Recall is (TN)/(FN+TN).

Model	Decoy	Accuracy	Precision	Specificity
frequency	shift	fail	fail	fail
frequency	reverse	fail	fail	fail
average	shift	75.6	75.2	76.2
average dd	shift	73.7	75.5	71.6
average cl	shift	65.7	67.5	60.6
average	reverse	73.2	71.6	76.9
number	shift	89.7	96.5	82.4
number dd	shift	89.1	94.5	83.0
number cl	shift	70.0	74.4	61.4
number	reverse	91.1	96.1	85.8
Total length	shift	90.0	96.6	83.0
Total length dd	shift	87	95.9	77.3
Total length cl	shift	73.7	75.5	70.1
Total length	reverse	91.9	96.7	86.7
Cartesian	shift	90.4	92.6	88.0
Cartesian dd	shift	88.3	92.0	83.9
Cartesian cl	shift	70.7	73	65.8
Cartesian	reverse	91.8	94.15	89.2

Since svm_light can easily perform leave one out estimates they were performed as well and the leave one out estimates are identical within the estimated variation to the 5-fold cross-validation estimates. The best results are seen with un-normalized data. Normalizing numbers of types of arcs to frequencies produced data sets where no SVM model could be found, and normalizing the total lengths along kinds of arcs to average lengths reduced the accuracy by about 15%. The Delaunay graph and the defective Delaunay graph produced essentially equivalent results. Using a graph constructed solely based on distances produced results that were worse than either the Delaunay or defective Delaunay

graph. Simply knowing the identities of the residues associated with the adjacency matrix (the number case below) was sufficient to accurately classify the data. Adding distance information to the number information improved results slightly.

4.8.2 Regression

Each of the 20 individual protein structures used in the Moulder benchmark[114] was removed and the system trained on the remaining structures and then evaluated on the removed structure resulting in a 20-fold cross validation. The fraction of residues with errors less than 3.5Å was used as a regression target. The results are shown in table 2. The difference between using the average distance and non-normalized distances is more pronounced with regression than with classification, and the average distance trained very poorly. The high variance in the correlations reflects that 20 structures are not enough to span the space of protein folds. However, the best correlation factors (77-79% for R^2) demonstrate that system can be quite accurate when the training data are sufficient. The defective Delaunay triangulation performs slightly worse than the Delaunay triangulation, which we believe is due to the greater degree of information that is dropped with the sparser graph.

Table 4.8.2.1 Accuracy of regression analysis on the Sali dataset. The abbreviation dd refers to the defective Delaunay graph. The R^2 correlation coefficient is shown in percent.

Model	Average R^2	Standard Deviation	Best R^2
average	12.6	13.1	43.9
number	53	18.7	79.3
Total distance	52.1	16.5	77.4
Cartesian	51.2	18.7	79.3
number dd	40.8	21.6	71
Total distance dd	38.7	20.4	65.4
Cartesian dd	43.3	21.7	70.7

4.9 Discussion

Encoding protein folds with a function applied to a triangulation derived graph results in an effective, compact, constant-sized code that is suitable for machine learning and data mining. The fact

that a simple linear SVM could be effectively trained for both classification and regression shows that the encoding is highly linear and effectively represents the features in the structure. It should be pointed out that this work only encoded structural features of the proteins, and no additional information such as hydrogen bonding, solvent exposure, measures of structural quality, sequence homology or profile information or knowledge based potential functions was used to assist the machine learning. Undoubtedly, with careful selection and training other features could be added to this model and improve its performance.

One conclusion of this work is highly suggestive. Normalizing the data to protein size, either by finding average distances or (worse) by converting from numbers of arcs to frequencies resulted in degradation of both classification accuracy and regression. This strongly suggests that the optimal encoding of protein structure should include a measure of protein size. Indeed, simply appending the numbers of arcs to the average distances (the Cartesian product above) restored the performance, although this could simply reflect the sufficiency of the numbers as a type of data. This suggests, as well, that deriving a highly accurate knowledge based potential to distinguish between native and non-native protein models without including terms that reflect protein size is likely to be very difficult, if not impossible.

4.9.1 Necessity of the Triangulation

Since the calculation of a triangulation is an extra, and potentially expensive, step in the encoding, It may be asked if the distances could simply be summed for each kind of residue pair in the model and this used as a measure for training. This may work for distance information, but it will be suboptimal because it includes more information that is needed. Indeed, when tested, the performance of a simple distance based triangulation was significantly worse at classification than either the Delaunay or defective Delaunay triangulations. This strongly suggests that exclusion of interactions from the triangulation based on the local molecular geometry is important for defining effective and accurate encodings.

The easiest way to see the importance of the intermediate triangulation step is to examine the accuracy of the measure based on the numbers of each kind of pair of residues in an arc of the graph. Simply knowing the numbers of each kind of residue pairs associated with an arc of the graph or non-zero element of the adjacency matrix of the triangulation is sufficient to give accurate results for both classification and regression. Adding distance information improved the results, but only by a small amount. Without the triangulation step the total numbers of inter-residue pairs is a function of the primary sequence and not the three-dimensional structure. Failing to use some form of triangulation results in data that cannot be used for classification or regression against structural metrics because the identical data could be derived in the absence of structure information. In essence, the triangulation binds the sequence information to the encoding so that the association between amino acid sequence and structure is established in the data.

4.9.2 Are the Triangulations Pseudo-Kernels?

Kernels in SVM's are distance measures or inner products in Hilbert space that are tuned to measure important distances in the data[118]. An encoding or representation of the data that is also a distance measure and therefore suitable for use with a linear SVM kernel can be thought of as a “pseudo-kernel” and may indeed be a candidate for inclusion as a “user-defined” kernel in an SVM package. Since the linear kernel worked well with our encoding, it is worth examining the metric properties of the encodings.

The adjacency matrices associated with the triangulations cannot solely by themselves be metrics, since the measure of distances between matrices depends on the definition of an appropriate norm. If the functions of the adjacency matrices that we define for use with the SVM are norms then they will obey triangle or transitive ordering rules (i.e. a>b and b>c implies a>c) and thus the triangulations will define, via functions applied to the graphs, linear “pseudo-kernels”. Since the functions defined in this paper output a vector containing solely positive elements by collapsing the adjacency matrix

based on the labels associated with each non-zero point in the matrix, and the norms of vectors are well-defined in terms of the properties of individual elements of the vectors, we need only examine the properties of the elements of the vectors to establish metric properties.

The number of arcs of each kind or the ordinality of the matrices obeys transitive ordering. If the number of XY arcs for a given pair X,Y for a is N and for b N-1 and c N-2, then both a>b, b>c and a>c. Similarly if for XY a>b and a=c and for ZQ b>c (but a = b), then a>c. Adding the distances to the number terms does not break this transitive ordering since the distances are all positive real numbers. Therefore most of the functions we have defined obey a metric structure. Interestingly, the normalized functions do not obey this ordering which may partly explain they do not perform as well as the unnormalized functions.

4.10 Conclusion

Triangulation-based encodings are an effective approach to reducing large complicated three-dimensional objects, like protein structures, to small and constant-sized representations suited for machine learning. With protein structures, simply knowing the kinds of residues which are adjacent in the triangulation is sufficient for accurate classification and regression analysis. Adding information about distances along the arcs of the triangulation increased the accuracy for classification, but was less important. It was surprising how small the effects of distance information were. Normalizing the data derived from the triangulation degraded the quality of the results. While the Delaunay triangulation performed the best of the three triangulations examined, the exact details of the triangulation algorithm are probably not critical as long as the triangulation uses local geometry to remove redundant or irrelevant features.

5 HIV DRUG RESISTANCE PREDICTION USING MULTIPLE REGRESSION: AN APPLICATION OF A NEW SEQUENCE/STRUCTURE HYBRID PROTEIN ENCODING METHOD[119]

5.1 Abstract

Drug resistance is commonly encountered during treatment for HIV/AIDS, and decreases the efficacy of the antiviral drugs. Genotyping the infecting virus gives sequence data for computational prediction of resistance, which is more efficient than performing experimental assays for resistance. Current predictions rely on simple rules with modest accuracy; therefore, a prediction method with high accuracy is needed to improve drug selection for therapy. Here, we apply a hybrid sequence/structure protein representation in conjunction with multiple regression for predicting resistance to drugs. The algorithm was tested on genotype-phenotype data for HIV-1 protease (PR) and HIV-1 reverse transcriptase (RT). The overall cross-validated regression R^2-values were 0.51-0.72 for predicting resistance to four PR inhibitors; and 0.76-0.91 for three RT inhibitors demonstrating successful predictions.

5.2 Introduction

HIV-infections have spread all over the world in the three decades since the first case of AIDS was found. Current treatment is highly active antiretroviral therapy (HAART), which combines at least three drugs. Drugs inhibiting HIV-1 reverse transcriptase (RT) or protease (PR) target two important viral enzymes. Both enzymes play an essential role for effective replication of the virus. However, mutations in drug targets causing resistance to the drugs rise commonly causing a challenge in therapy[120]. Multiple mutations accumulate over time, resulting in a huge number of possible combinations of mutations. Accurate and fast computational prediction of resistance is needed urgently for better drug selection instead of expensive experimental assays.

Many machine learning methods have been tested for predictions of drug resistance: linear regression, decision trees[62], neural networks[72], support vector regression[70, 121], and Bayesian networks[122].

We have introduced a hybrid sequence/structure representation using Delaunay triangulation for efficient encoding of inter-residue contacts within 3-dimensional structural data[123]. Previous application of this encoding to PR genotype-phenotype data gave superior accuracies over other methods for prediction of drug resistance[124]. Results for predicting sequences with resistance to 4 PR inhibitors gave a high classification accuracy of >0.95 with 5-fold cross-validation using either support vector machine (SVM) or artificial neural networks and >0.97 using the sparse dictionary. This accuracy is significantly higher than values of 0.60-0.87 obtained for the same set of sequences and inhibitors using other prediction methods [125, 126]. We have applied this hybrid sequence/structure representation to encode the HIV PR and RT protein structures, and used multiple regression to predict the relative resistance for selected drugs: SQV, TPV, IDV and LPV inhibiting HIV PR; and AZT, Delavirdine (DLV) and Efavirenz (EFV) inhibiting HIV RT.

5.3 Methods

5.3.1 Datasets

Genotype-Phenotype Data are from the Stanford HIV drug resistance database[36] (http://hivdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi). Data were used for 4 PR inhibitors SQV, TPV, IDV and LPV, and 3 HIV RT inhibitors AZT, DLV and EFV. All the genotypes were expanded to produce individual unique amino acid sequences because more than one possible amino acid was shown at some positions in the sequence.

5.3.2 Hybrid sequence/structure protein representation using Delaunay triangulation

A hybrid sequence/structure protein representation method was used[105, 124]. Only the sequences of the mutated proteins are needed and only one protein structure is necessary. Hence, all mutants are represented as vectors of the same dimensionality, which is a desired property for most of the pattern recognition algorithms.

Two structure templates were used: 3OXC for HIV-1 PR, and 2WOM for HIV-1 RT (from www.pdb.org). The amino acid residues in each structure were represented by their alpha carbon positions. Delaunay triangulation was performed as described[124] resulting in a vector of 210 independent values, which is used as a feature vector to represent the protein structure in learning and classification.

5.3.3 Regression analysis for drug resistance prediction and cross validation

The 210-dimensional vector representing each mutant is used in regression analysis. The drug resistance value from the Phenosense assay for each genotype is given in the datasets. The mutations relative to a standard sequence are analyzed with the assayed resistance value to find a linear model. Then, a *k*-fold regression test was performed. The training set of size N is randomly divided into k groups. Among them, $k-1$ groups are utilized for constructing the linear model. Then, the linear model is used to predict the drug resistance for the remaining group with N/k mutations. The predicted resistances are compared with the measured ones and the R^2 values are recorded. Finally, the average and standard deviation of the k R^2 values are computed.

5.4 Results

5.4.1 Predicting HIV protease inhibitor resistance

We performed *k*-fold (*k*=5) regression analysis on the sequence and resistance data. The real relative resistance values were included for the multiple regression. The regression gave R^2 values of

0.5141-0.7212 for four different drugs as shown in Table I, which demonstrates that resistance can be predicted successfully by the hybrid sequence/structure encoding method.

Table 5.4.1.1 Multiple Regression On Predicted Relative Resistance FOR PR INHIBITORS

	R^2 values, mean	R^2 values, stddev
IDV	0.5141	0.0306
LPV	0.7212	0.0158
TPV	0.5208	0.0543
SQV	0.5758	0.0254

5.4.2 Predicting HIV reverse transcriptase inhibitor resistance

Multiple regression analysis was performed similarly for HIV RT and its inhibitors AZT, DLV and EFV. The regression results gave very high R^2 values of 0.7622-0.9164 for the three different inhibitors, as shown in Table II. Therefore, the hybrid sequence/structure method gave excellent success in predicting resistance to RT inhibitors.

Table 5.4.2.1 Multiple regression on predicted relative resistance FOR RT INHIBITORS

	R^2 values, mean	R^2 values, stddev
AZT	0.7622	0.0237
DLV	0.9088	0.0073
EFV	0.9164	0.0079

5.5 Discussion

We have evaluated a new method to predict the drug resistance for both HIV-1 PR and HIV-1 RT antiviral inhibitors from genotype data using a hybrid sequence and structure protein representation and multi-regression analysis. This method was tested on four HIV PR inhibitors and three HIV RT inhibitors and produced high accuracy. Regression analysis, determined from existing mutational data, can then be used to estimate the relative resistance value of novel mutants to drugs. In contrast, more standard methods only assess the presence of known resistance mutations in the sequence. The overall cross-validation regression R^2 was 0.51-0.72 for four PR inhibitors; while even higher values of 0.76-0.92 were obtained for three RT inhibitors. Therefore, this new method is able to predict drug resistance with

high accuracy and has promise for selecting the most effective drugs when resistance arises during AIDS therapy.

5.6 Acknowledgment

Xiaxia Yu was supported by the Georgia State University Research Molecular Basis of Disease Program. This research was supported, in part, by the National Institutes of Health grant GM062920.

AIM 2: Developing a new classification algorithm to distinguishing between the drug resistant and the none drug resistant mutants

6 SPARSE REPRESENTATION FOR PREDICTION OF HIV-1 PROTEASE DRUG RESISTANCE[124]

6.1 Abstract

HIV rapidly evolves drug resistance in response to antiviral drugs used in AIDS therapy. Estimating the specific resistance of a given strain of HIV to individual drugs from sequence data has important benefits for both the therapy of individual patients and the development of novel drugs. We have developed an accurate classification method based on the sparse representation theory, and demonstrate that this method is highly effective with HIV-1 protease. The protease structure is represented using our newly proposed encoding method based on Delaunay triangulation, and combined with the mutated amino acid sequences of known drug-resistant strains to train a machine-learning algorithm both for classification and regression of drug-resistant mutations. An overall cross-validated classification accuracy of 97% is obtained when trained on a publically available data base of approximately 1.5×10^4 known sequences (Stanford HIV database http://hivdb.stanford.edu/cgi-bin/GenoPhenoDS. cgi). Resistance to four FDA approved drugs is computed and comparisons with other algorithms demonstrate that our method shows significant improvements in classification accuracy.

6.2 Introduction

Since the disease of AIDS (Acquired Immunodeficiency Syndrome) was first recognized in the US in the early 1980s, it has become a severe worldwide epidemic[127]. Based on the life cycle of the infectious agent human immunodeficiency virus (HIV), many inhibitors were constructed to treat AIDS. These inhibitors can retard the entry, replication or maturation of the virus. Therefore all of them are effective as anti-AIDS drugs.

The inhibitors of HIV protease have proved to be potent anti-viral drugs[128], since the protease plays an important role in the maturation of the virus[129]. Up till now, nine HIV protease inhibitors have been approved by the FDA (Food and Drug Administration): amprenavir (APV), indinavir (IDV), lopinavir (LPV), nelfinavir (NFV), ritonavir (RTV), saquinavir (SQV), atazanavir (ATV), tipranavir (TPV) and darunavir (DRV).

The structure of the HIV-1 protease is shown in Figure 1. HIV protease is a homodimer, and each monomer has 99 residues. The inhibitors bind inside the active site in the center of the dimer by hydrogen bonds and van der Waals interactions and prevent the cleavage of viral precursor proteins. Therefore, the virus cannot form mature particles and thus cannot infect other host cells[130, 131].

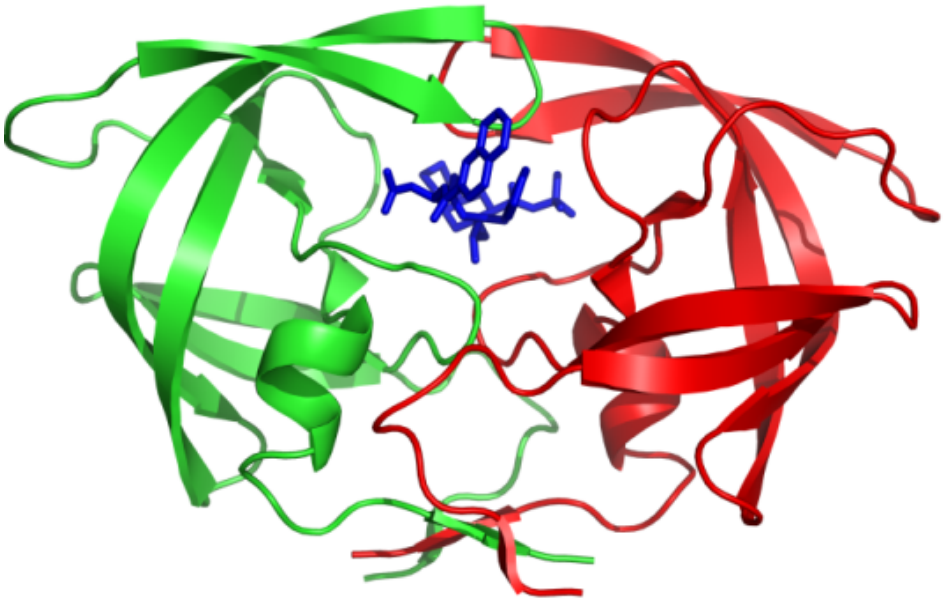


Figure 5.4.2.1 The structure of HIV-1 protease with Saquinavir. Two monomers are shown in red and green. Saquinavir is shown in blue.

However, because HIV has deficient proofreading[33] and a high rate of replication[35], mutations evolve rapidly in its genome. Such mutations lead to drug resistance or decreased susceptibility to certain drugs, though in some rare cases the drug efficacy was observed to increase for certain mutations[132]. Hence, resistance testing is recommended for AIDS patients due to the decreased suscepti-

bility for certain drugs[133] Mutations associated with resistance are found in almost half the protease residues. They are located around the active site of the protease where they can alter the interactions with inhibitors and throughout the structure[134]. Multiple mutations accumulate over time. Due to the huge number of possible combinations of mutations, it is a challenge to predict which protease sequences will cause resistance to specific inhibitors. Accurate predictions would be valuable for prescribing the most effective drugs for infections with resistant HIV.

Most existing approaches to predict HIV drug resistance from sequence data use only the sequence data and often only selected sets of mutation sites, such as geno2pheno[135], REGA[136], Stanford HIVdb[137], ANRS[138], and HIV-GRADE[125]. In this paper we incorporate structural data into the predictions. The structural information improves the quality of the predictions by representing interactions between physically adjacent mutation sites that are not adjacent in sequence unlike other methods.

The resistance can be assessed for HIV strains by experiments growing the infected cells in the presence of different drugs. However, even minimal wet lab experiments to measure the antiviral efficacy of individual inhibitors are time consuming and expensive. Therefore, it would be valuable develop computer methods to predict whether a mutant is drug-resistant or not.

In the field of extracting information, the sparse signal representation has emerged in recent years as a promising research area. Indeed, the sparsity is a hidden prior information for most of the signals in the physical world and the related philosophy and algorithms have been applied in a diverse areas[82, 139]. Sparse signal representation can be visualized as a technique for extracting the essential features from the data while simultaneously minimizing the effects of the noise in the data. For example, a sparse signal representation of audio data would extract the continuous sound waves while suppressing the uncorrelated and non-continuous background noise.

Therefore, in this paper, we apply the sparse signal technique in the prediction of HIV-1 protease drug resistance from sequences. In the BACKGROUND section, a brief background of the sparse signal representation is presented; in PREVIOUS WORK, we briefly review the area; in METHODS section, the details of our proposed classification algorithm are introduced. Following that, the RESULTS and DISCUSSION sections describe the outcomes and related discussions.

6.3 Background

Compressive sensing uses sparse signal representations to eliminate noise and non-critical features from the data[82, 139]. The data are expanded in terms of an orthogonal basis – often a Fourier or wavelet basis for conventional signals – and the critical features extracted based on the magnitudes of the coefficients of the expansion. A classical expansion, like the Fourier transform, is not always the optimal basis for expansion and therefore the choice of an optimal basis is done using optimization[139]. The optimal basis for machine learning with protein sequence and structure data is defined in terms of a dictionary of exemplars which are determined with the singular value decomposition KSVD[140] as described in the methods below.

The idea of the above compressive sensing and sparse signal representation has achieved very exciting results in many areas such as signal acquisition[141], signal representation[140], pattern classification[142], and image processing[86]. In this work the idea of dictionary learning and classification is extended and applied in the problem of predicting drug resistance from HIV-1 protease sequence data.

6.4 Methods

In this section, we first provide a vector representation for the protein structure, and then the sparse dictionary is used to perform the classification task.

6.4.1 Data sets

A total of 11731 phenotype results from 1727 isolates were obtained from Genotype-Phenotype Datasets on the Stanford HIV drug resistance database[137] (http://hivdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi).

In this experiment, four protease inhibitors, SQV, TPV, IDV and LPV, were tested.

For SQV, IDV and LPV, among all these genotype sequences, those mutants with the relative resistant fold < 3.0 were classified as non- resistant, denoted as 0; while those with the relative resistant fold ≥ 3.0 were classified as resistant, denoted as 1[143].

For TPV, those mutants with the relative resistant fold < 2.0 were classified as non-TPV resistant, denoted as 0; while those with the relative resistant fold ≥ 2.0 were classified as TPV resistant, denoted as 1[144].

6.4.2 Preprocessing of the datasets

In order to unify the data in the original datasets, those sequences with an insertion, deletion, or containing a stop codon relative to the consensus have been removed so that the data represent proteases of 99 amino acids.

Due to the limitations of the sequencing assay or presence of multiple viral sequences in the same sample, many of the sequences in the dataset have multiple mutations at the same sites yet share the same drug-resistance characteristics. An individual protein molecule can only have one type of amino acid at one location. Therefore, we need to expand the data to multiple sequences with single amino acids at each location. For instance, among the 99 letters of a sequence, 97 of them have a single amino-acid. However, at one site there are two different types of amino-acids, and another site has three. In this case, this record must be expanded to a total of $6 = (2 \times 3)$ different sequences, each of which has only one amino-acid for each of its 99 residues, sharing the same drug resistance of the original se-

quence. In this work, we designed a fast way to perform this expansion, which significantly enriches the test data.

Without loss of generality, for a sequence in the original data set, we denote the number of variations on each of its 99 sites to be $J_i, i = 1,2,...,99$. Therefore, this sequence can be expanded to a total number of $P = \prod_{i=1}^{99} J_i$ different sequences, each of which has only one type of amino acid at each position. In order to generate them all, equivalently, for any $p \in \{1,2,...,P\}$, we need to pick a unique combination among the 99 positions.

This choice can be done with a simple recursive implementation. Unfortunately, it has so high a complexity that in practice, we only obtain roughly 5k sequences within 24 hours on an Intel Core i7 workstation. In order to improve this speed, we designed a new method for this expansion by analogy to the base-conversion problem. For a simple example, assume J_i=2 for all $i = 1,2,...,99$. Then, the task of listing all the 2^{99} sequences, though a huge number, can be done by simply finding the representation of each $p \in \{1,2,...,2^{99}\}$ under base 2 and picking the 1st (resp. 2nd) amino acid on each site if a 0 (resp. 1) is encountered on that digit. By analogy, in this task, we need to convert a decimal number *p* to a mixed-base number: its *i*-th digit is a J_i-based number.

This can be done, similarly to the decimal-binary conversion, by successive short division. However, the difference is that instead of dividing by 2, here J_i should be used for the *i*-th division. The short division is repeated and the remainders are recorded in a reversed order, which finally gives a 99-digit mixed-base representation of *p*, denoted as π. Then, for each site, we just pick the amino acid according to the *i*-th digit of π.

With this new scheme, we generated a total of 1.5×10^5 sequences in less than 10 seconds on the same machine. This significantly enriches the available data for the subsequent analysis.

6.4.3 Protease structure representation

It is necessary to use a representation of the structure that is invariant with respect to the arbitrary choice of origin and orientation of the molecule. Therefore, the procedure in[145] was used to convert the HIV-1 protease structure into a 210-dimensional vector.

The structure of wild type (consensus) HIV protease with SQV (PDBID: 3OXC[146]) was obtained from the Protein Structure Database at www.pdb.org. Then, the position of each residue was represented by its alpha carbon position. Because the wild-type HIV-1 protease has 198 residues in the dimer, the α-carbon positions consist of 198 three-dimensional vectors, $\vec{C} = \{C_1, C_2, ..., C_{198} : C_i \in \Re^3\}$. The Delaunay triangulation is then performed on the $\vec{C}$ and a graph $G = < \vec{C}, E >$ is obtained. Then, for the edge $e \in E$, the two residues it connects are denoted as A_i and A_j where $A_i, A_j \in \mathrm{A}$ being the set of all the 20 amino-acids. We then recode the distance between C_i and C_j as $d(A_i, A_j)$. This process is repeated for all the edges in G and the distances computed for the same pair of amino-acids are averaged. Finally, the averaged values are filled into the corresponding positions of a matrix $D \in \Re^{20 \times 20}$. For example: $D(1, 2)$ and $D(2, 1)$ contains the average distance between the amino-acids A_1 and A_2 appearing in the graph G.

Evidently, the matrix D is symmetric. Therefore, it has a total of 210 degrees of freedom (upper triangular part plus the diagonal). Those 210 values are concatenated in a row-wise manner to form a 210-dimensional vector, which will be termed “structure vector” for short. The subsequent learning and classification are based on such structure vectors.

6.4.4 Sparse dictionary classification

From the brief introduction of the compressive sensing/sparse representation, it can be seen that for a more accurate signal reconstruction, rather than using some existing fixed basis/frames such as the Fourier basis, it is very important to find a suitable basis/frame Ψ, so that the signals of interest have sparse representations in Ψ. In the signal processing community, such a frame is also called a dic-

tionary. Given a group of signals, the task of finding a dictionary that can represent the group of signals sparsely is called the dictionary construction.

The use of the signal dependent frame, as opposed to the generic frames/basis such as Fourier, wavelet, etc., gives us a new approach to the signal reconstruction problem. Indeed, one can view the construction of the signal dependent frame (dictionary) as a process of building a sparse, nonlinear model for the signals at hand. As a result, the fidelity of reconstructing a new signal from the dictionary can then be considered as a measure of how the new signal fits the model represented by the dictionary. Therefore, this can be used under a classification framework: Assume we have *n* groups of signals, for example (but not limited to) *n*=2 in our drug-resistant/non-resistant case. Then, we can construction two dictionaries as the models for the resistant/non-resistant groups, respectively. After that, a new signal (the "structure vector" described in the above section), is fit to the two models by reconstructing it using the two dictionaries. The reconstruction errors using different dictionaries are compared and the smaller error indicates that the signal fits to that specific dictionary better than to the other. As can be observed, there is no limitation on n being 2 and therefore the proposed method can be viewed as a nonlinear multi-group classification scheme. In addition, the sparsity of the representation makes the classification more efficient. In what follows, we present the details of the proposed algorithm.

Denote $u_1, u_2, \ldots, u_M, v_1, v_2, \ldots, v_M$ as the training sets and $u_{M+1}, u_{M+2}, \ldots, u_{M+N}, v_{M+1}, v_{M+2}, \ldots, v_{M+N}$ as the testing sets. In order to learn and encode the information of the vectors belonging to SQV group (resistant to SQV), we construct an over complete dictionary *J* from $u_1, u_2, \ldots, u_M$. To that end, the K-SVD algorithm is employed and shown in Algorithm 1.

The dictionary *J* records the information of the SQV group and similarly, the other over-complete dictionary *K*, which learns and encodes the information of non-SQV group, is constructed from $v_1, v_2, \ldots, v_m$ also with the K-SVD algorithm.

Algorithm 1 **K-SVD Dictionary Construction**[140]

1: Initialize J by the discrete cosine transformation matrix
2: **repeat**
3: Find sparse coefficients $\Lambda(\lambda_i's)$ using any pursuit algorithm.
4: **for** j=1, 2, ..., update j_i, the j-th column of J, by the following process **do**
5: Find the group of vectors that use this atom: $\zeta_i \coloneqq \{i : 1 \le i \le M, \lambda_i(j) \ne 0\}$
6: Compute $E_j \coloneqq Q - \Sigma_{i \ne j} j_i \Lambda_T^i$ where Λ_T^i is the i-th row of Ë
7: Extract the i-th columns in E_j, where $i \in \zeta_j$, to form E_j^R
8: Apply SVD to get $E_j^R = U \Delta V$
9: j_i is updated with the first column of U
10: The non-zeros elements in Λ_T^j is updated with the first column of $V \times \Delta(1,1)$
11: **end for**
12: **until** Convergence criteria is met

In this work, we used the orthogonal matching pursuit algorithm to find the sparse coefficients[147]. The two dictionaries encode the information in either group of vectors. Therefore, intuitively, a vector belonging to the SQV group could be represented by *J* with high fidelity and vice versa for the non-SQV group. Formally, a new vector $\vec{w} \in \Re^{210}$ with unknown category, is reconstructed by both dictionaries *J* and *K*. To that end, the orthogonal match pursuit algorithm is used to find a sparse coefficient Λ and Γ, such that

$$\vec{w} \approx J\Lambda \text{ s.t. } \Lambda \in \Re^{210}, \| \Lambda \|_0 < k$$

$$\vec{w} \approx K\Gamma \text{ s.t. } \Gamma \in \Re^{210}, \| \Gamma \|_0 < k$$

However, the two dictionaries could represent $\vec{w}$ with different accuracy. The representation errors are recorded as:

$$e_{SQV} = \| \vec{w} - J\Lambda \|_2$$

$$e_{non-SQV} = \| \vec{w} - K\Gamma \|_2$$

and finally

$$e = e_{SQV} - e_{non-SQV}$$

Therefore, if $e > 0$, the new vector $\vec{w}$ could be represented better by the dictionary constructed from the vectors of the SQV group. Hence, it is classified to be resistant to the SQV. The overall algorithm is listed in Algorithm 2

Algorithm 2 Drug resistance classification algorithm
1: repeat
2: Randomly choose m vectors from SQV group, the rest n being training data
3: Construct dictionary J using Algorithm 1
4: Randomly choose m vectors from none group, the rest n being training data
5: Construct dictionary K using Algorithm 1
6: for each vector v in testing data do
7: computing the sparse representation of v using both dictionaries J and K
8: computing the representation errors using the two dictionaries
9: if the error of using J is larger then
10: v is resistant to SQV
11: else
12: v is NOT resistant to SQV
13: end if
14: end for
15: Compute the confusion matrix
16: until For 9 times

6.5 Experiments and results

6.5.1 k-fold validation

In order to fully use all the data, a k-fold cross-validation was performed in all the experiments for all the four drugs. Specifically, $k-1/k$ of all the sequences are used for training the classifier and the remaining $1/k$ data are used for testing. We pick k to be 5 for all the tests. For each of the four types of the drugs, we then have approximately 10k "structure vectors", half are resistant and the other half are non-resistant. Accordingly, there is about 2k testing vectors for each drug.

6.5.2 Support vector machine

The support vector machine (SVM) is a framework for the supervised learning and classifying task. After its proposal by Vapnik[118], the SVM has been used widely in the machine learning/pattern classification filed.

When feeding the encoding result into SVM, 5-fold cross validation tests were performed implemented in MATLAB SVM toolbox[148, 149]. We tested several choices for the SVM kernel and the linear kernel has the best performance, as reported in Table 5 (choice of kernel is further discussed in Section 2.4.8). Care was taken to insure that all positive and negative instances of a given protein were removed from either training or testing dataset when generating a set for cross-validation. This avoided the potential problem of having negative instances associated with a positive test item or positive instances associated with a negative test item and thus generating systematically optimistic (and incorrect) assessments of the training accuracy.

Table 6.5.2.1 Mean accuracy, specificity and sensitivity using SVM

	IDV	LPV	SQV	TPV
Accuracy	0.961	0.959	0.950	0.961
stddev $_{(\times 10^2)}$	0. 233	0. 251	0. 249	0. 402
Sensitivity	0.951	0.947	0.947	0.958
stddev $_{(\times 10^2)}$	0. 469	0. 348	0. 424	0. 463
Specificity	0.971	0.973	0.953	0.964
stddev $_{(\times 10^2)}$	0. 368	0.341	0. 325	0. 369

6.5.3 Artificial Neural Networks

The same testing strategy was applied with the Artificial Neural Networks (ANN) to classify data. Specifically, the three-layer feedforward network was used in Matlab[149-151]. The network had one hidden layer of 20 nodes and was trained with backpropagation with a maximum of 50 training epochs. Similar to SVM, 5-fold cross validation was also used for ANN and the result is shown in Table 6.

Table 6.5.3.1 Mean accuracy, specificity and sensitivity using ANN

	IDV	LPV	SQV	TPV
Accuracy	0.961	0.963	0.957	0.951
stddev$_{(\times 10^2)}$	0.857	0.641	0. 723	1.27
Sensitivity	0.960	0.965	0.958	0.953
stddev$_{(\times 10^2)}$	1.16	0.741	0.483	1.89
Specificity	0.963	0.961	0.956	0.950
stddev$_{(\times 10^2)}$	0.981	0.598	1.06	0. 672

6.5.4 Proposed sparse dictionary classifier

Following the approach described in METHODS, the sparse representation was also implemented and 5-fold cross validation was performed. The result is shown in Table 7.

Table 6.5.4.1 Mean accuracy, specificity, and sensitivity using sparse representation

	IDV	LPV	SQV	TPV
Accuracy	0.969	0.974	0.970	0.990
stddev$_{(\times 10^2)}$	0.151	0. 292	0.139	0.277
Sensitivity	0.951	0.957	0.959	0.984
stddev$_{(\times 10^2)}$	0.529	0.494	0.604	0.423
Specificity	0.989	0.992	0.981	0.995
stddev$_{(\times 10^2)}$	0.297	0.361	0.692	0.199

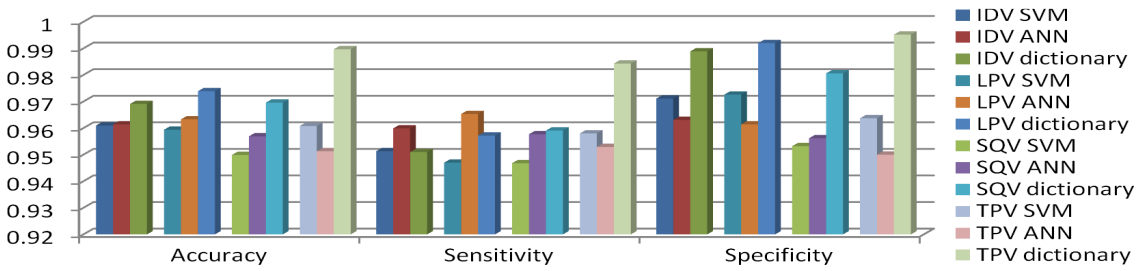


Figure 6.5.4.1 Comparison of accuracy, specificity and sensitivity of sparse dictionary, SVM, and ANN

For clarity, the mean accuracy of all the above methods is compared in Figure 16. From it we can observe that the mean accuracy of the proposed dictionary classifier is higher than for other methods.

While Figure 16 visualizes the comparison among the mean accuracies, sensitivities and specificities, we further conducted statistical tests for all the 5-fold cross validation results. At the significant level of 0.01, the accuracy, sensitivity and specificity of the proposed method are higher than for both SVM and ANN.

6.5.5 Comparison with other methods

Furthermore, we have tested several state-of-the-art methods including HIV-GRADE (Version 12-2009), ANRS-rules (Version 7/2009), Stanford HIVdb (Version 6.0.6), Rega (Version 8.0.2), and geno2pheno (version December 13, 2000), which are available at http://www.hiv-grade.de/cms/grade/,

using the same datasets described above. Since the original dataset obtained from Stanford HIVdb are all protein sequences, and all these servers take nucleotide sequence, the sequence manipulation suite[152] was used to convert the protein sequences into nucleotide ones. When parsing the output of these methods, the output term with “susceptibility”, is considered as non-resistant, whereas output of “resistance” is considered as being resistant. Accuracies are presented in Table 4. For the HIV-grade, there are outputs termed “Intermediate”. When calculating the accuracies, "Intermediate" is considered as resistant, and the result is shown in the table 4. In the table, N/A indicates that there is no output for this method-inhibitor.

Table 6.5.5.1 Accuracy compared to other methods

	IDV	LPV	SQV	TPV
HIV-grade	0.851	0.805	0.802	0.728
ANRS	0.851	0.870	N/A	0.597
HIVdb	N/A	0.839	N/A	0.768
Rega	0.856	0.840	0.693	N/A
Sparse	0.969	0.974	0.970	0.990

From the comparison we can observe the high accuracy achieved in our proposed sparse method. The consistent high level of accuracy demonstrates that including structural information and sparse encoding is a promising new alternative approach to only using sequence information for this important task of predicting drug resistance.

6.5.6 Mean accuracy with respect to different sparsity

The parameters of the algorithm, in particular the sparsity and the dictionary size, affect the final classification outcome. The sparsity controls how many atoms are used to re-construct a given vector. If it is large, then both dictionaries would give smaller representation errors. Therefore, it is a parameter that can be tuned. By varying from 7 to 12, we repeated the learning and classification steps. Then the mean accuracy was measured and plotted in Figure 17. It is noted that for all the tests here, the dictionary size is fixed at 250.

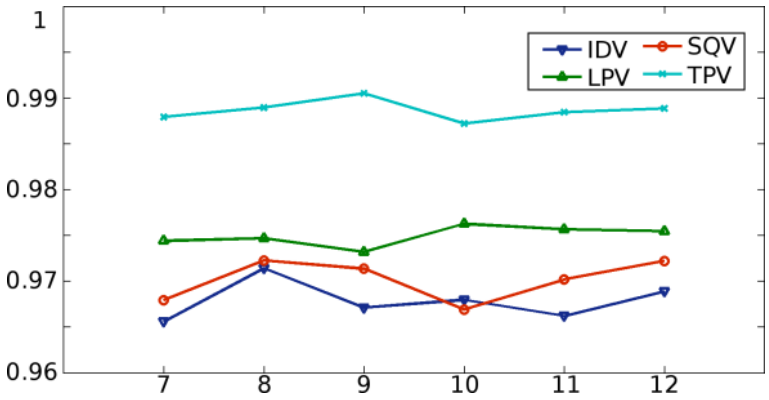


Figure 6.5.6.1 The accuracy changes with respect to the change of the sparsity. The lines are the mean accuracies of the k tests with different sparsity. The dictionary size is fixed at 250.

6.5.7 Mean accuracy with respect to dictionary size

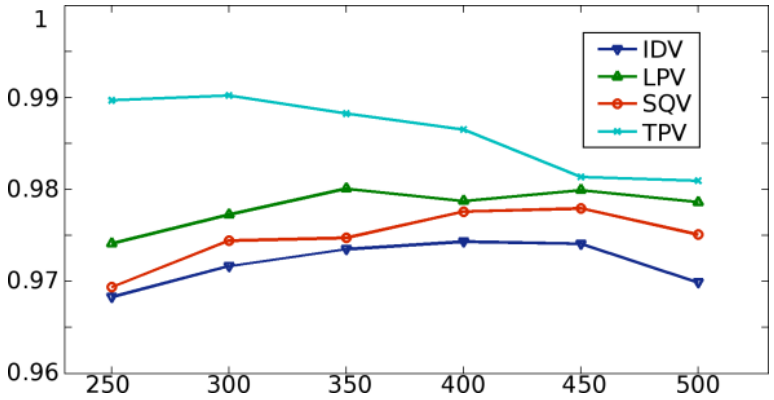


Figure 6.5.7.1 The accuracy changes with respect to the change of the dictionary size. The lines are the mean accuracies of the k tests with different dictionary sizes. The sparsity is fixed at 9.

The dictionary is an over-complete set of vectors (atoms) and the number of atoms in it is also a parameter that affects the learning and classification performance. Therefore, similar to the tests for the sparsity above, tests with different dictionary sizes were conducted (varying from 250 to 500) and the resulting accuracies are recorded in Figure 4. Moreover, for these tests, the sparsity value was fixed at 9.

From the tests we can observe that with further parameter tuning, the proposed algorithm has the potential of reaching even higher accuracy.

6.5.8 Computational Performance

As mentioned in Section 4.1, we have approximately 10k training "structure vectors" and 2k testing vectors for each single classification task. As can be seen in Table 9, although the proposed algorithm achieves better classification accuracy, it also takes longer to finish. For the SVM, any choice of kernel other than the linear one does not lead to convergence within 104 seconds.

Table 6.5.8.1 Running times for training

Method	SVM (linear)	SVM (non linear)	ANN	proposed
Training Time (Sec)	20.6	no convergence (>10^4)	21.9	358
Testing Time (Sec)	0.4	N/A	0.1	2

6.6 Discussion

Given a mutant strain of HIV-1, in order to establish whether it is resistant to certain drugs, wet lab biological experiments are conducted. However, this process is both time and resource consuming. Therefore, performing such experiment in silico will save much time and resources. Hence, in this work we propose an algorithm to predict the drug resistance property of the mutant HIV-1 protease from its sequence. It is based on the signal sparse representation theory. Essentially, we learn the characteristics of resistant and non-resistant mutants of the HIV-1 protease by constructing two over-complete dictionaries. Then, given the sequence of a new mutant, we measure how accurately this new sequence can be represented by the two dictionaries. The category of the dictionary with smaller error is assigned to the new mutant. The algorithm is tested on different sequences, and the result was compared with the common classification tools SVM and ANN. The result shows that the proposed sparse dictionary classifier can distinguish between drug resistant and non-resistant sequences significantly better than the other methods. Moreover, this new method outperforms existing approaches in terms of accuracy. This method for in silico prediction of resistance may be a promising way to select effective drugs in AIDS therapy without performing the actual biological experiments. In our on-going and future research, we will be extending the bi-partition algorithm to multiple class classification. This would enable grouping the proteins in finer divided and more accurate sub-categories.

6.7 Acknowledgments

Xiaxia Yu was supported by the Georgia State University Research Molecular Basis of Disease Program. This research was supported, in part, by the National Institutes of Health grant GM062920.

7 PREDICTION OF HIV DRUG RESISTANCE FROM GENOTYPE WITH ENCODED THREE-DIMENSIONAL PROTEIN STRUCTURE[153]

7.1 Abstract

Background: Drug resistance has become a severe challenge for treatment of HIV infections. Mutations accumulate in the HIV genome and make certain drugs ineffective. Prediction of resistance from genotype data is a valuable guide in choice of drugs for effective therapy.

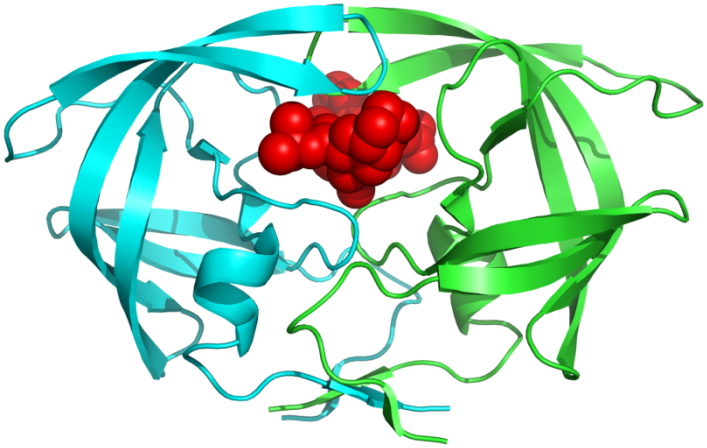
Results: In order to improve the computational prediction of resistance from genotype data we have developed a unified encoding of the protein sequence and three-dimensional protein structure of the drug target for classification and regression analysis. The method was tested on genotype-resistance data for mutants of HIV protease and reverse transcriptase. Our graph based sequence-structure approach gives high accuracy with a new sparse dictionary classification method, as well as support vector machine and artificial neural networks classifiers. Cross-validated regression analysis with the sparse dictionary gave excellent correlation between predicted and observed resistance.

Conclusion: The approach of encoding the protein structure and sequence as a 210-dimensional vector, based on Delaunay triangulation, has promise as an accurate method for predicting resistance from sequence for drugs inhibiting HIV protease and reverse transcriptase.

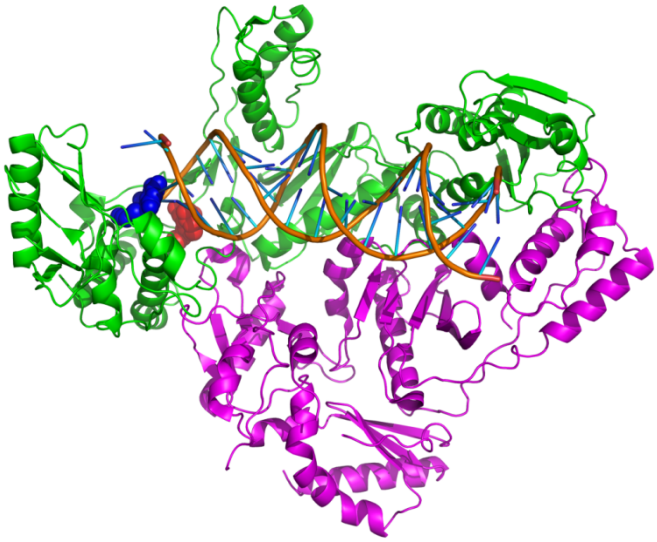
7.2 Background

HIV/AIDS is a pandemic disease and more than 35 million people are infected worldwide[154]. There is no effective vaccine; however, the long-term survival of many patients has been enabled by drug therapy. Highly Active Antiretroviral Therapy (HAART) using three or four different drugs with different viral targets is very effective in stabilizing the infection[155]. These antiviral drugs target different stages in the viral life-cycle. Two important drug targets are the HIV protease (PR) and reverse transcrip-

tase (RT), which have essential roles in viral replication. HIV RT converts the viral RNA genome into DNA, which is translated by the host cell machinery into the viral precursor proteins. HIV PR functions to cleave the large viral precursor proteins into individual enzymes and structural proteins, which produces infectious viral particles. Among the 23 approved drugs in current clinical use, there are seven nucleoside RT inhibitors (NRTIs), four non-nucleoside RT inhibitors (NNRTIs), and eight PR inhibitors (PIs)[156]. The approved PIs were designed to bind in the active site of HIV PR, and prevent the processing of viral precursor proteins (Figure 1A). NRTIs are chemical analogs of the natural nucleoside substrates of the HIV RT that bind to the protein active site and block its activity in synthesizing DNA from viral RNA. The inhibitors in the NNRTI class also decrease the enzymatic activities of RT, however, they bind in an allosteric site in the palm domain of the p66 subunit instead of the active site of RT (Figure 1B).



(A)



(B)

Figure 6.5.86.5.8.1 Structures of HIV-1 PR and RT. (A) The structure of HIV-1 PR dimer in complex with the inhibitor darunavir [157]. The two subunits of HIV-1 PR are shown in green and red, and the PI darunavir is colored blue. (B) The structure of HIV-1 RT dimer in complex with DNA and bound NNRTI and NRTI[26, 27]. The p66 subunit is shown in green and the p51 subunit is shown in purple. NRTI is colored blue, NNRTI is red, and double stranded DNA is orange.

Despite the success of HAART, current therapy is limited by the rapid emergence of drug resistance[156]. The virus can mutate to acquire resistance during therapy due to the lack of proofreading by RT[33] and high replication rate[35]. These resistance mutations alter the drug targets such as PR and RT[158]. Some of the 35 mutations associated with resistance to PIs alter amino acids located in the active site of PR while the majority alter residues in distal regions of the enzyme structure[134]. Similarly

for RT, several of the mutations associated with resistance to NRTIs alter amino acids in the active site of the enzyme while others are located in more distal regions. The amino acid mutations occurring in association with resistance to the NNRTIs tend to cluster around the inhibitor binding site[42, 43]. The molecular mechanisms for these antiviral drugs are described in the review[10].

The resistance mutations lower the effectiveness of specific drugs and may cause failure of the treatment. Infections with resistant HIV are prevalent; surveys in North America and Europe show that 8-20% of HIV infections in untreated people contain primary drug resistance mutations[10]. Over time, multiple mutations can accumulate giving a huge number of possible combinations of mutations in each protein. This persistent problem led to the recommendation for resistance testing to guide the choice of drugs in AIDS therapy [133, 159, 160]. Fast sequencing of the genome of the infecting virus can be combined with computational predictions of resistance to guide the choice of effective antiviral drugs[160]. Accurate and fast computational predictions are desirable to avoid the expense, limited availability and time involved for performing an experimental cell-based assay for resistance where results can take four weeks.

Accurate predictions can be valuable for prescribing the most effective drugs for infections with resistant HIV. Most genotype interpretation algorithms in clinical use are knowledge based[161]. These interpretation algorithms apply a set of rules or scores for each mutation and drug. The performance of several commonly used interpretation algorithms: Stanford HIVdb[126], HIV-grade[125], REGA and ANRS (www.hivfrenchresistance.org/) has been compared[125]. In addition, many computational classification techniques have been evaluated for predicting drug resistance from the genotype data. The standard classification techniques of artificial neural networks (ANN)[63, 72, 73, 162, 163], decision tree[62, 63], random forests[163], support vector machine (SVM)[163] [70] and regression analysis[63] have been applied in HIV drug resistance predictions. Statistical methods can also be applied to analyze the relationship between genotype and phenotype. The association of mutations with resistance to the

PIs saquinavir (SQV) and indinavir (IDV) was determined using cluster analysis, recursive partitioning, and linear discriminant analysis[65]. These methods are limited by the high dimensionality of the genotype data, hence non-parametric methods were proposed and tested on resistance data for the PI amprenavir (APV)[68, 69]. Protein structural information has also been used to generate four-body statistical potentials of mutants for training with classification and regression statistical learning algorithms and tested in predicting resistance to RT and PR inhibitors[79].

We have evaluated an efficient encoding of information from the three-dimensional protein structure for the prediction of resistance from genotype. The structural encoding via Delaunay triangulation improves the quality of the predictions by representing interactions between amino acid neighbours in the three-dimensional structure unlike the linear sequence representation of other methods. This unified sequence-structure representation was used in supervised training with SVM, ANN, and a new sparse dictionary classification method. The compressive sensing/sparse dictionary representation[81] [82] has been applied successfully in image analysis to enhance learning capacity and efficiency. Sparse representation has been employed for image restoration[84, 164], denoising[85], deblurring[86], signal processing[165], and face detection[88]. Initial tests of this procedure for classifying resistance to 4 PIs was presented in[124]. Here, the structural encoding has been expanded to include regression analysis and classification of genotype-phenotype data for seven PIs, six NRTIs and three NNRTIs.

7.3 Results

We combined structural information with genotype for regression analysis and supervised learning on resistance data. The new graph based sequence-structure encoding was tested with the Genotype-Phenotype Data from the Stanford HIV drug resistance database[137] (http://hivdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi). Data were available for two different protein targets: HIV-1 PR and HIV-1 RT. For HIV-1 PR, eight PR inhibitors atazanavir (ATV), IDV, nelfinavir (NFV), ritonavir (RTV), lopinavir (LPV), tipranavir (TPV) and SQV were tested. While for the study of HIV RT inhibitor resistance, NNRTIs

nevirapine (NPV), delaviridine (DLV), efavirenz (EFV), and NRTIs lamivudine (3TC), abacavir (ABC), zidovudine (AZT), stavudine (D4T), didanosine (DDI) and tenofovir (TDF) were tested. The data include the protein sequence and resistance value from the Phenosense assay for each virus isolate. Genotype-phenotype data were available for 744 to 1674 isolates for different inhibitors of HIV PR, while RT was represented by 353 to 746 records for the 9 different NRTIs and NNRTIs. The preprocessing of the sequence and resistance data are detailed in Methods. Genotypes were expanded to unique protein sequences due to the presence of more than one amino acid at some positions. This expansion resulted in a total of 10,228 to 17,545 unique sequences of HIV PR mutants and 2,004 to 11,367 RT mutants for the various inhibitor resistance values.

7.3.1 Graph based protein sequence/structure representation using Delaunay triangulation

The sequences were combined with information from the three-dimensional protein structures by employing a graph generated by Delaunay triangulation as described in[105]. Two structure templates were used: 3OXC[146] for HIV-1 PR, and 2WOM[166] (from www.pdb.org). Only one structure vector is needed for each protein. In other words, all PR mutant sequences are combined with a single 210-dimensional vector derived from one PR structure, and similarly, a single structure vector is used for the RT mutants in subsequent regression and classification of resistance data. As a result, all mutants are represented as vectors of constant dimensionality, which is a desirable property for most of the pattern recognition algorithms. This structure vector was combined with sequences in regression analysis and classification for resistance.

7.3.2 Multiple regression on HIV protease inhibitor resistance

After each of the mutated sequences was represented by a 210-dimensional vector, we performed the regression analysis for the drug resistance data. We performed k-fold (k=5) regression analysis on the sequence and resistance data. The predicted values for relative resistance were plotted

against the experimental values as shown in (Figure 2) for the PR inhibitors ATV, NFV, RTV, IDV, LPV, TPV and SQV.

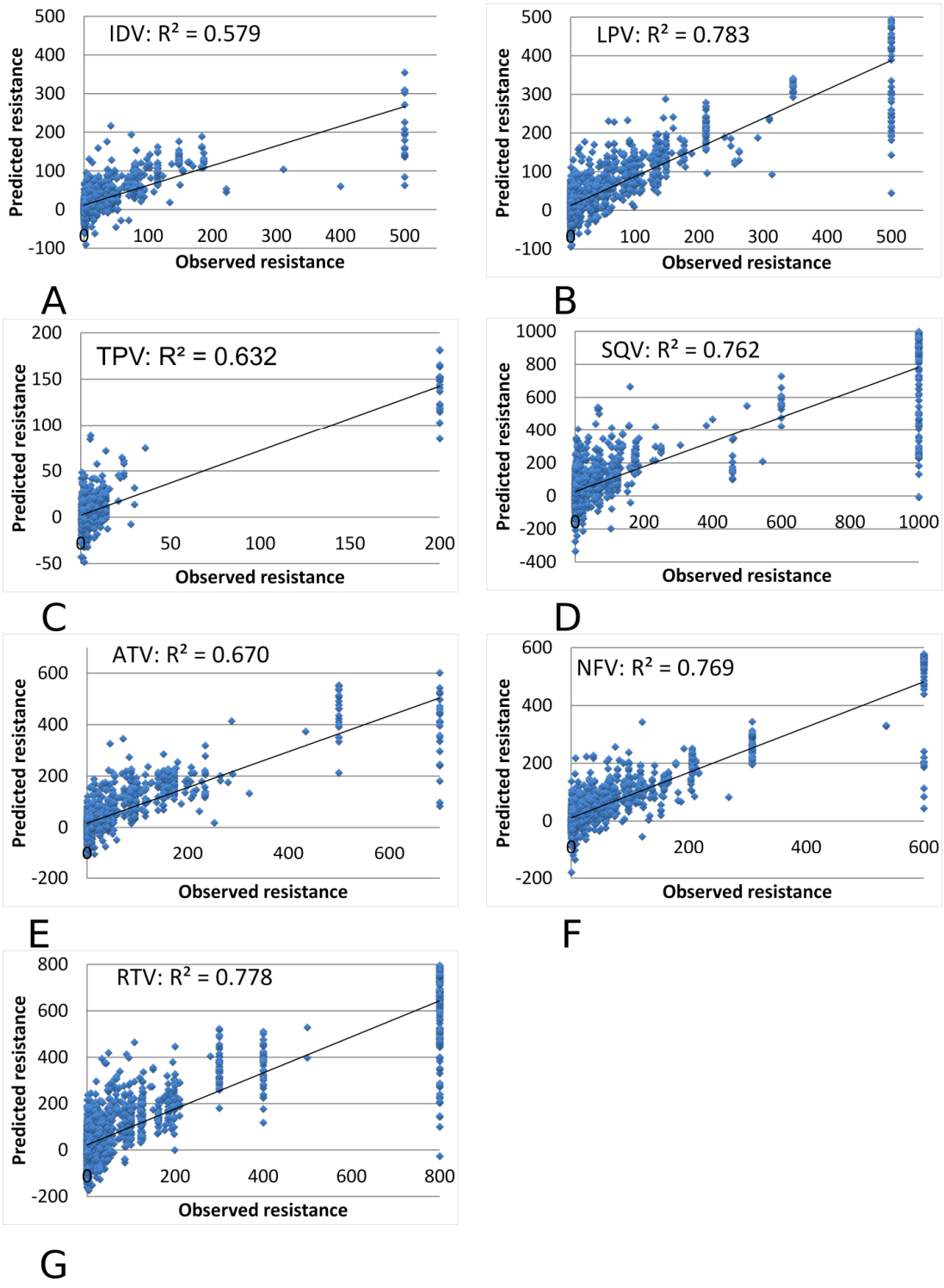


Figure 7.3.2.1 Multiple regression on the predicted and observed resistance for HIV-1 PR inhibitors. The predicted resistance is plotted against the observed value as blue dots. The trend line is shown. Plots show regression for drug resistance: (A) IDV, (B) LPV, (C) TPV, (D) SQV, (E) ATV, (F) NFV, (G) RTV

The multiple regression gave high R^2 values of 0.579-0.783 and very low standard deviations as listed in Table 1. The values are the average of all the R^2 values from k-fold regression. The excellent correlations demonstrate that relative resistance to PIs can be predicted successfully from genotype by the new sequence/structure encoding method. In order to avoid training to an "optimal" n-fold set for cross validation, cross validation sets are chosen independently for each training run. Therefore, there is always a small variation in the results.

Table 7.3.2.1 Multiple regression on predicted relative resistance to HIV-1 PR inhibitors

	IDV	LPV	TPV	SQV	ATV	NFV	RTV
R^2 values, mean	0.579	0.783	0.632	0.762	0.670	0.769	0.778
R^2 values, stddev	0.037	0.014	0.045	0.018	0.035	0.029	0.016

7.3.3 Multiple regression on HIV reverse transcriptase inhibitor resistance

Multiple regression analysis was performed similarly on genotype-phenotype data for drugs inhibiting HIV-1 RT. The predicted and observed values are compared for resistance to the RT inhibitors including NRTIs 3TC, ABC, D4T, DDI, TDF and AZT (Figure 3), and NNRTIs NPV, DLV and EFV for NNRTIs (Figure 4).

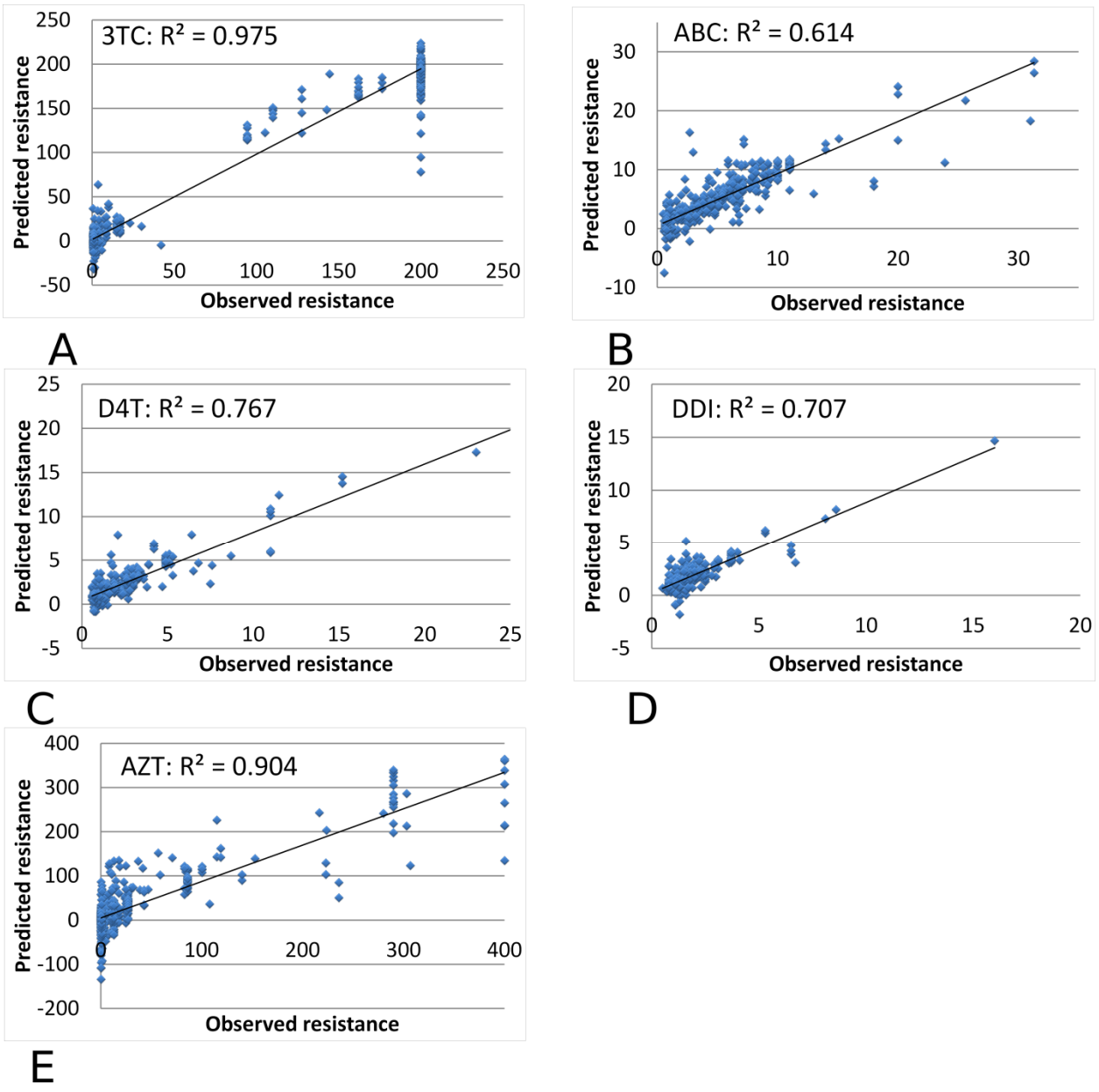


Figure 7.3.3.1 Multiple regression on the predicted and observed resistance for HIV-1 NRTIs. The predicted resistance is plotted against the observed value as blue dots. The trend line is shown for (A) 3TC, (B) ABC, (C) D4T, (D) DDI, and (E) AZT.

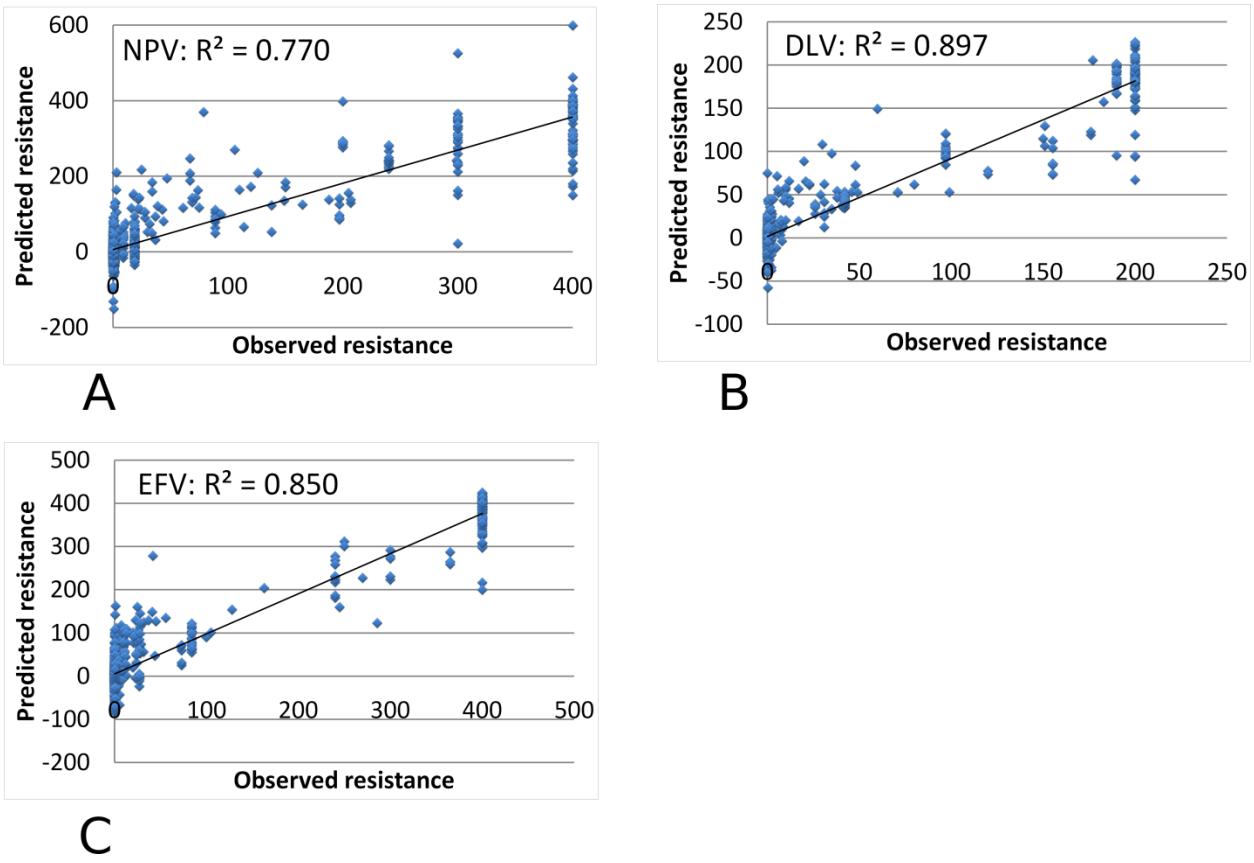


Figure 7.3.3.2 Multiple regression on the predicted and observed resistance for HIV-1 NNRTIs. The predicted resistance is plotted against the observed value as blue dots. The trend line is shown for (A) NPV, (B) DLV and (C) EFV.

The regression results gave high R^2 values of 0.614-0.975 for the different RT inhibitors, as shown in Tables 2 and 3. The resistance to NRTIs was predicted with excellent R^2 values of 0.85-0.90 and very low standard deviations, while resistance predictions for NRTIs gave R^2 values in the larger range of 0.61-0.98. Larger standard deviations were obtained for analysis of ABC and DDI, possibly because the range of values in the dataset was smaller than for the others. Therefore, graph based encoding had excellent success in predicting resistance to RT inhibitors.

Table 7.3.3.1 Multiple regression on predicted relative resistance for NNRTIs

	DLV	EFV	NPV
R^2 values, mean	0.904	0.897	0.850
R^2 values, stddev	0.015	0.012	0.015

Table 7.3.3.2 Multiple regression on predicted relative resistance for NRTIs

	AZT	3TC	ABC	D4T	DDI
R^2 values, mean	0.770	0.975	0.614	0.767	0.707
R^2 values, stddev	0.023	0.004	0.253	0.061	0.146

7.3.4 Classification of Resistance with Support vector machine

The support vector machine (SVM) was proposed by Vapnik[118], and is widely used as a supervised learning classifier. In this experiment, 5-fold cross validation tests were performed by implementing in MATLAB SVM toolbox[148, 149] and the linear kernel was used. The results are shown in Tables 3-5 for HIV-1 PR inhibitors (PIs), HIV-1 RT inhibitors NRTIs and NNRTIs. This classification shows high accuracy, sensitivity and specificity for all inhibitors. For PIs the accuracy values range from a low of 0.93 to a high of 0.96, while sensitivity and specificity range from 0.92-0.96 and 0.94-0.98, respectively. Resistance to NRTIs is classified with even higher accuracies of 0.97-0.99, sensitivities of greater than 0.98 and specificities of 0.95-0.99, while for NNRTIs the classification performance was superior with all values of over 0.97 for accuracy, sensitivity and specificity. The excellent performance with the linear SVM kernel demonstrates conclusively that the novel encoding using Delaunay triangulation separates the resistant and non-resistant data into two distinct categories.

Table 7.3.4.1 Classification using SVM for Resistance to PIs

	ATV	IDV	NFV	RTV	LPV	SQV	TPV
Accuracy	0.955	0.960	0.933	0.946	0.962	0.946	0.961
Stddev ($\times 10^{-2}$)	0.400	0.510	0.350	0.580	0.220	0.580	0.290
Sensitivity	0.943	0.951	0.923	0.945	0.952	0.945	0.957
Stddev ($\times 10^{-2}$)	0.600	1.00	0.400	0.910	0.270	0.910	0.410
Specificity	0.968	0.970	0.943	0.947	0.972	0.947	0.965
Stddev ($\times 10^{-2}$)	0.450	0.290	0.820	0.890	0.280	0.890	0.410

Table 7.3.4.2 Classification using SVM for Resistance to NRTIs

	3TC	ABC	AZT	D4T	DDI	TDF
Accuracy	0.987	0.981	0.984	0.992	0.965	0.975
Stddev ($\times 10^{-2}$)	0.484	0.234	0.390	0.371	0.289	0.914
Sensitivity	0.984	0.981	0.984	0.991	0.977	0.979
Stddev ($\times 10^{-2}$)	0.613	0.379	0.627	0.417	0.436	1.21
Specificity	0.991	0.982	0.984	0.993	0.954	0.970
Stddev ($\times 10^{-2}$)	0.510	0.397	0.470	0.505	0.625	1.76

Table 7.3.4.3 Classification using SVM for Resistance to NNRTIs

	NPV	DLV	EFV
Accuracy	0.982	0.983	0.991
Stddev ($\times 10^{?}$)	0.254	0.473	0.316
Sensitivity	0.972	0.976	0.986
Stddev ($\times 10^{?}$)	0.490	0.600	0.618
Specificity	0.992	0.991	0.996
Stddev ($\times 10^{?}$)	0.397	0.787	0.301

7.3.5 Classification with Artificial Neural Networks

As in the SVM experiment, the 5-cross validation test was applied to the Artificial Neural Networks (ANN) to classify genotype-phenotype data for resistance. Specifically, the three-layer feedforward network was used in Matlab[149-151]. The network had one hidden layer of 20 nodes and was trained with backpropagation with a maximum of 50 training epochs. The results are shown in Tables 6-8 for HIV-1 PR inhibitors, and RT inhibitors NRTIs and NNRTIs. The values calculated for accuracy, sensitivity and specificity for resistance to PIs have a low of 0.91 and reach 0.97. Improved performance was achieved for classifying resistance to RT inhibitors compared with PIs. Results for NRTIs gave values of accuracy, sensitivity and specificity of 0.96-0.99, while for NNRTIs all values were greater than 0.98.

Table 7.3.5.1 Classification using ANN for Resistance to PIs

	ATV	IDV	NFV	RTV	LPV	SQV	TPV
Accuracy	0.958	0.944	0.917	0.934	0.963	0.957	0.951
Stddev ($\times 10^{?}$)	0.320	1.25	1.38	1.44	0.641	0. 723	1.27
Sensitivity	0.959	0.940	0.913	0.935	0.965	0.958	0.953
Stddev ($\times 10^{?}$)	0.460	1.56	2.46	1.13	0.741	0.483	1.89
Specificity	0.957	0.947	0.922	0.933	0.961	0.956	0.950
Stddev ($\times 10^{2}$)	0.440	0.944	1.05	1.97	0.598	1.06	0. 672

Table 7.3.5.2 Classification using ANN for Resistance to NRTIs

	3TC	ABC	AZT	D4T	DDI	TDF
Accuracy	0.982	0.984	0.987	0.983	0.965	0.970
Stddev ($\times 10^{2}$)	0.469	0.525	0.164	0.452	0.176	1.21
Sensitivity	0.984	0.978	0.988	0.980	0.973	0.965
Stddev ($\times 10^{2}$)	0.994	0.700	0.428	0.983	0.434	1.67
Specificity	0.980	0.991	0.986	0.986	0.957	0.975
Stddev ($\times 10^{2}$)	0.835	0.474	0.490	0.687	0.168	1.00

Table 7.3.5.3 Classification using ANN for Resistance to NNRTIs

	NPV	DLV	EFV
Accuracy	0.983	0.986	0.986
Stddev ($\times 10^2$)	0.524	0.488	0.503
Sensitivity	0.979	0.985	0.982
Stddev ($\times 10^2$)	0.507	1.24	0.955
Specificity	0.987	0.987	0.990
Stddev ($\times 10^2$)	0.554	0.448	0.462

7.3.6 Classification using sparse dictionary

The sparse dictionary classifier was also implemented using the 5-fold cross validation tests using the approach described in[124]. The results are shown in Tables 7-9 for HIV-1 PR inhibitors, HIV-1 RT NRTIs and NNRTIs. High values were obtained for accuracy, sensitivity, and specificity. Accuracies ranged from 0.95-0.99 for resistance to PIs, 0.82-0.92 for NRTIs and 0.81-0.84 for NNRTIs. The sensitivities were all greater than 0.93 for the calculations on resistance to PIs, and specificities were greater than 0.96. Lower values were obtained for calculations on some of the RT inhibitors where values for sensitivity ranged from 0.75 to 0.96, while high specificity values from 0.86 to 1.00 was calculated. These performance measures are somewhat poorer than for the standard SVM and ANN classifiers. It is not surprising; however, that more development may be necessary for applying the new sparse dictionary as a classifier since previously it has been employed primarily for image processing.

Table 7.3.6.1 Classification using sparse dictionary for resistance to PIs

	ATV	NFV	RTV	IDV	LPV	SQV	TPV
Accuracy	0.973	0.946	0.962	0.969	0.974	0.970	0.990
Stddev ($\times 10^2$)	0.262	0.602	0.269	0.151	0. 292	0.139	0.277
Sensitivity	0.961	0.927	0.968	0.951	0.957	0.959	0.984
Stddev ($\times 10^2$)	0.244	0.635	0.976	0.529	0.494	0.604	0.423
Specificity	0.986	0.967	0.958	0.989	0.992	0.981	0.995
Stddev ($\times 10^2$)	0.661	1.44	1.23	0.297	0.361	0.692	0.199

Table 7.3.6.2 Classification using sparse dictionary for resistance to NRTIs

	3TC	ABC	AZT	D4T	DDI	TDF
Accuracy	0.918	0.915	0.932	0.879	0.816	0.852
Stddev ($\times 10^2$)	3.44	3.14	4.20	5.06	7.63	7.20
Sensitivity	0.963	0.872	0.947	0.814	0.801	0.789
Stddev ($\times 10^2$)	2.60	5.08	4.73	6.81	6.11	8.45
Specificity	0.888	0.973	0.933	0.987	0.860	0.972
Stddev ($\times 10^2$)	6.78	0.185	8.75	1.02	12.1	4.19

Table 7.3.6.3 Classification using sparse dictionary for resistance to NNRTIs

	NPV	DLV	EFV
Accuracy	0.826	0.844	0.811
Stddev ($\times 10^2$)	2.46	2.49	6.43
Sensitivity	0.761	0.773	0.753
Stddev ($\times 10^2$)	3.48	3.82	8.43
Specificity	0.938	0.973	0.935
Stddev ($\times 10^2$)	2.87	2.11	3.55

7.3.7 Comparison with standard genotype interpretation methods

Finally, we compared our methods with the standard drug resistance prediction methods HIV-GRADE, ANRS-rules, Stanford HIVdb, and Rega, which are available at http://www.hiv-grade.de/cms/grade/, using the same genotype-phenotype datasets described in Methods. The procedure discussed in[124] was used to convert the protein sequences into nucleotide sequences. Other methods usually give resistance interpretations in three categories of “resistance, “intermediate” and “susceptible”. Since multiple classification is difficult with SVM and ANN, only two classes were considered for calculating the accuracy. Both "resistant" and "intermediate" are considered as "resistant"; while "susceptible" is considered as "non-resistant". The results are shown in Tables 10-12 for HIV-1 PR inhibitors, HIV-1 RT NRTIs and NNRTIs. N/A means that no output was obtained from the server for this dataset.

Table 7.3.7.1 Accuracy (%) compared to other methods for HIV-1 PR inhibitors

	ATV	NFV	IDV	LPV	SQV	TPV
HIV-grade	84.7	81.2	85.1	80.5	80.2	72.8
ANRS	N/A	78.1	85.1	87.0	N/A	59.7
HIVdb	N/A	83.4	N/A	83.9	N/A	76.8
Rega	84.4	82.2	85.6	84.0	69.3	N/A
SVM	95.5	96.0	94.6	96.2	94.6	96.1
ANN	95.8	94.4	93.4	96.3	95.7	95.1
Sparse dictionary	97.3	94.6	96.9	97.4	97.0	99.0

Table 7.3.7.2 Accuracy (%) compared to other methods for HIV-1 RT NRTIs

	3TC	ABC	AZT	D4T	DDI	TDF
HIV-grade	91.5	89.7	94.6	88.1	89.7	80.7
ANRS	92.0	83.9	94.4	87.7	73.3	72.7
HIVdb	94.3	95.0	94.5	86.2	87.6	79.7
Rega	95.9	86.0	94.0	92.2	88.3	73.8
SVM	98.7	98.1	98.4	99.2	96.5	97.5
ANN	98.2	98.4	98.7	98.3	96.5	97.0
Sparse dictionary	91.8	91.5	93.2	87.9	81.6	85.2

Table 7.3.7.3 Accuracy (%) compared to other methods for NNRTIs

	NPV	DLV	EFV
HIV-grade	98.7	N/A	98.1
ANRS	94.8	N/A	97.9
HIVdb	98.4	N/A	98.7
Rega	98.6	96.8	98.7
SVM	98.2	98.3	99.1
ANN	98.3	98.6	98.6
Sparse dictionary	82.6	84.4	81.1

The accuracies demonstrate that classification with our structural encoding significantly outperforms other state of the art methods for predicting resistance to PIs for the three tested classifiers SVM, ANN and the sparse dictionary. Accuracies of 93.4-99.0% were obtained with structural encoding compared to 59.7-87.0% for the standard methods. The highest accuracies of greater than 95% were achieved with the sparse dictionary method. The prediction accuracy for resistance to the NRTI class of RT inhibitors also showed the advantages of our structural encoding with values of 81.6-99.2% compared with 72.7-95.9% for standard methods. In this case, the SVM and ANN classifiers performed better than the new sparse dictionary giving accuracies of at least 97%. For the NNRTIs, the structural en-

coding with SVM or ANN gave higher accuracies of 98.3-99.1% compared with 94.8-98.7% for standard methods. The sparse dictionary, however, showed lower performance with accuracies of 81.1-84.4% for NNRTI resistance, indicating some improvements may be needed for the new classifier.

7.4 Discussion

The serious problem of drug resistance arising during therapy of HIV-infected individuals can be tackled by sequencing the HIV drug targets to identify mutations followed by computational prediction of resistance to guide the choice of effective therapy. Computational predictions of the most effective drugs for the mutated HIV provide a major advantage of low cost and speed relative to experimental assays for resistance. Most standard prediction methods are knowledge based methods, such as the genotype interpretation algorithms. These algorithms either use a set of rules, for example, the Visible Genetics/Bayer Diagnostics genotype interpretation rules[167], to generate the susceptibility of the infecting virus for each drug; or apply a score or 'penalty' for each drug such as the Stanford HIV database[168] and mutation rate based score[56]. Also, a combined rule-based and penalty-based method has been proposed and applied to both HIV-1 PR and RT inhibitors[57]. Although these methods are fast, they suffer from the major disadvantage of relying on specific known mutations strongly associated with resistance and cannot identify newly appearing resistance mutations, or assess the effects of many mutations more weakly associated with resistance.

Various machine learning and statistical methods have been applied to this problem, including the widely used classifiers, ANN[72, 73], decision tree[62], and SVM[70]. Statistical methods such as cluster analysis, recursive partitioning, and linear discriminant analysis have been evaluated[65], and non-parametric methods proposed for high dimensionality data[68, 69]. Most of these methods are based on the linear protein sequence and omit potentially valuable information from the three-dimensional protein structure. Additional information has been introduced in the form of 544 physico-chemical descriptors for the amino acid mutations leading to correlation coefficients of 0.75-0.94[162].

Other groups have included structural features such as PR-drug contacts in the binding site with majority voting [18]. In another example, Delaunay triangulation of RT and PR structures was used as input to a four-body statistical potential to predict resistance to inhibitors. The four-body statistical potential was derived from 1375 non-redundant structures in the PDB to assess protein structural quality. This procedure gave mean accuracies of 0.68-0.83 for PIs, 0.70-0.89 for NRTIs and 0.75-0.82 for NNRTIs[79]. These accuracies are significantly lower than we obtain with a single structure vector (Tables 13-15). Our procedure uses Delaunay triangulation to directly encode the structure and sequence for machine learning without the extra step of calculating a potential. Our direct encoding is likely responsible for the higher accuracy in our results.

Another energy based approach uses molecular mechanics calculations on the PR-drug structure have been used to predict resistance of mutants, and high correlation (R^2 of 0.76-0.85) was reported between calculated value and IC_{50} from the experimental assay[51]. However, these calculations must be performed for each individual mutant-drug combination and will be slow for assessing large numbers of mutants for resistance.

We have developed a simple graph representation of protein structure for fast classification. The protein structure is a three-dimensional object that has many physical and chemical factors potentially effecting stability and activity. Previously, we showed that Delaunay triangulation was the best of several graph-based encodings of protein structure and sequence[37]. The graph-based encoding algorithm condenses a complicated three-dimensional object, a protein structure, into a relatively small hash function with 210 unique values per sequence and structure. One critical outcome is that the graph-based encoding results in a linearly separable data set that can be used readily by several different machine learning algorithms. Similarly, the encoding is sufficiently linear that straightforward multiple linear regression can be performed on the training data. The hash value maintains enough information about the complicated object to provide useful information for machine learning and regression.

This unified sequence-structure encoding gave high accuracy in initial tests on four PRIs[124]. Here, we demonstrate successful application of the structure vector in multiple regression analysis and classification on resistance data for seven inhibitors of HIV PR and nine inhibitors of RT. The 5-fold validated regression analysis gave excellent correlation between predicted and observed resistance with excellent R2 values of 0.58-0.78 for PIs, 0.61-0.98 for NRTIs and 0.85-0.90 for NNRTIs. Classification with SVM, ANN or a new sparse dictionary method gave high accuracies for predicting the resistance for PR and RT inhibitors. The structure vector encoding had superior accuracy to predictions on the same sequences using standard interpretation algorithms. The sparse dictionary classifier was the best of tested classifiers for prediction of resistance to PIs, whereas SVM classification gave the best performance on resistance prediction for RT inhibitors. This structure vector encoding of genotype data has the advantage of using a single 210-dimensional vector for each protein target. The algorithm has one slow step for preparing the encoding from a single protein structure that can be applied to all genotypes in a fast calculation, in contrast to molecular mechanics calculations that must be set up in a non-trivial manner for each individual protein sequence. The entire protein sequence is combined with the structure vector, so there is the potential for accommodating new mutations or combinations of mutations with weak but concerted effects on resistance. The procedure can be extended easily in future calculations for resistant mutants with insertions in the protein sequence, which occur commonly in RT[156]. The new sparse dictionary classification approach can be extended to multiple classifiers by using more than two dictionaries, which is a significant advantage over the tested standard SVM or ANN classifiers, and may permit accurate predictions for different levels of resistance.

7.5 Conclusions

The simple unified encoding of structural information with genotype gives high accuracy for prediction of resistance to HIV PR and RT inhibitors as well as excellent correlation coefficients in regression analysis. The improvement over algorithms using only linear sequence information suggests the

importance of local interactions between mutated residues in the protein structure, which is consistent with the correlated local changes observed in the crystal structures of a highly resistant PR mutant with 20 substitutions[169]. Graph-based encoding of sequence and structure holds promise for fast and accurate predictions of resistance from sequence in order to guide the choice of effective drugs for treatment of HIV infections. In future, this approach can be expanded to predict resistance for other drugs and more diverse types of data.

7.6 Materials and Methods

7.6.1 Data sets and data preparation

All the datasets were retrieved from Genotype-Phenotype Data on the Stanford HIV drug resistance database[106] (http://hivdb.stanford.edu/cgi-bin/GenoPhenoDS.cgi). In this experiment, the proposed algorithm was tested on two different systems: HIV-1 PR and HIV-1 RT resistance data. For HIV-1 PR, eight PR inhibitors atazanavir (ATV), nelfinavir (NFV), ritonavir (RTV), IDV, lopinavir (LPV), tipranvir (TPV) and SQV were tested. While for the study of HIV RT inhibitor resistance, NNRTIs NPV, delaviridine (DLV), efavirenz (EFV), and NRTIs lamivudine (3TC), abacavir (ABC), zidovudine (AZT), stavudine (D4T), didanosine (DDI) and tenofovir (TDF) were tested.

All positive and negative instances of a given mutant were removed from either training or testing dataset before the cross-validation. This may avoid the potential problem of having negative instances associated with a positive test item or positive instances associated with a negative test item, and thus assure the training accuracy.

7.6.2 Pre-processing of the datasets

In order to unify the data in the original datasets, those sequences with an insertion, deletion, or containing a stop codon relative to the consensus have been removed so that the data represent proteins of identical size.

Many of the sequence records in the dataset have multiple residues at the same sites yet share the same drug-resistance value, which may be due to sequencing limitations or to the existence of multiple viral strains in the same isolate. In order to represent a single amino acids sequence for each mutant protein, we need to expand the data to multiple sequences with single amino acids at each location. For instance, in one 99-amino acid mutant of HIV PR, at one site there are two different types of amino-acids, and another site has three. In this case, this record must be expanded to a total of $6 = (2 \times 3)$ different sequences, each of which has only one amino-acid for each of its 99 residues, sharing the same drug resistance. We designed a fast method to perform this expansion as detailed in[124], which significantly enriches the test data.

The results of the expansion for each of the HIV-1 PR inhibitors were: a total of 16846 sequences were obtained from 1622 isolates with assays for IDV resistance; a total of 16269 sequences from 1322 isolates for LPV; a total of 10228 sequences from 744 isolates for TPV; a total of 17118 sequences from 1640 isolates for SQV; a total of 12084 sequences from 1012 isolates for ATV; a total of 17545 sequences from 1674 isolates for NFV; and a total of 16652 sequences from 1589 isolates for RTV.

For each of the HIV-1 RT inhibitors the expansion resulted in: a total of 11367 sequences were obtained from 746 isolates with assays for NPV resistance; a total of 11299 sequences from 732 isolates for DLV; a total of 11354 sequences from 734 isolates for EFV; a total of 4850 sequences from 633 isolates for 3TC; a total of 4846 sequences from 628 isolates for ABC; a total of 4847 sequences from 630 isolates for AZT; a total of 4845 sequences from 630 isolates for D4T; a total of 4849 sequences from 632 isolates for DDI; and a total of 2004 sequences from 353 isolates for inhibitor TDF.

7.6.3 Cutoffs for resistance/susceptibility for each drug

For the HIV-1 PR inhibitors: ATV, IDV, NFV, and RTV, the genotype sequences giving the relative resistance fold < 3.0 were classified as non-resistant (susceptible), denoted as 0; while those with the relative resistance fold ≥ 3.0 were classified as resistant, denoted as 1[63].

With the HIV-1 RT inhibitors: for ABC and TPV, those mutants with the relative resistant fold < 2.0 were classified as non-resistant, denoted as 0; while those with the relative resistant fold ≥ 2.0 were classified as resistant, denoted as 1; for 3TC, AZT, NPV, DLV, EFV, SQV, IDV and LPV those mutants with the relative resistant fold < 3.0 were classified as non-resistant, denoted as 0; while those with the relative resistant fold ≥ 3.0 were classified as resistant, denoted as 1; for D4T, DDI and TDF, those mutants with the relative resistant fold < 1.5 were classified as non-resistant, denoted as 0; while those with the relative resistant fold ≥ 1.5 were classified as resistant, denoted as 1[63].

7.6.4 Encoding structure and sequence with Delaunay triangulation

The sequence and structure of the protein were represented using a graph-based encoding as described in[124]. Delaunay triangulation was used to define a graph which spanned the protein structure and defined adjacent pairs of amino acid residues. Adjacent pairs of amino acids were summarized into a vector of the 210 unique kinds of amino acid pairs by calculating the distance for each adjacent pair in the structure and tabulating by the types of amino acids in that adjacent pair. Only the sequences of the mutated proteins are needed and only one protein structure is necessary. As a result, all mutants are represented as vectors of the same dimensionality, which is a desired property for most of the pattern recognition algorithms. The structures 3OXC[146] for HIV-1 PR, and 2WOM[166] for HIV-1 RT (from www.pdb.org) were used as templates for Delaunay triangulation.

7.6.5 k-fold validation

In order to fully use all the data, a k-fold cross-validation was performed in all the experiments for all the drugs. Specifically, we randomly choose (k-1)/k of all the sequences (some are drug resistant, while others are non-drug resistant) for training the classifier and the remaining 1/k data are used for testing. These tests used k=5. Independent randomly selected k-folds were chosen throughout the study to avoid bias in the results. The apparent polymorphism in the original sequence data requires extra

care when generating k-fold data sets for testing or training. When a sequence was removed from a k-fold in generating a testing or training dataset, all derived instances of that sequence were removed as well. This ensures that the individual k-fold datasets are truly independent from each other and thus ensures that the estimated accuracies are meaningful.

7.6.6 Regression analysis for drug resistance prediction

The Genotype-Phenotype Datasets provide a drug resistance value, with respect to a certain type of drug, with each genotype. The mutations relative to a standard sequence are denoted as $x_1, x_2, \ldots x_N ; x_i \in \Re^{210}$ where N is the total number of mutations and R^{210} is the structure vector. Also the corresponding drug resistance values are denoted as the real numbers $y_1, y_2, \ldots, y_N ; y \in \Re$ including 0 for the resistance value of the wild type virus. We then seek a linear model between the x_i's and y_i's by minimizing the cost function E :

$$E := \sum_{i=1}^{N} (y_i - A \cdot x_i - b)^2 \qquad (1)$$

with respect to the 210 dimensional vector A and scalar b.

Furthermore, in order to better utilize the available data set, we performed a k -fold cross-validation (in this work, k=5). Specifically, the training set of size N is randomly divided into k groups. Among them, $k-1$ groups are utilized for constructing the linear model as in Equation (1). Then, the linear model is used to predict the drug resistance for the remaining group with N/k mutations. The predicted resistances are compared with the measured ones and the R^2 values are recorded. Finally, the average and standard deviation of the k R^2 values are computed.

7.6.7 Sparse dictionary classification

In this experiment, we applied our newly proposed method described in[124] on both HIV-1 PR and HIV-1 RT data sets. In this case, the sequences of the mutants are considered as the group of signals, and given these signals, we would like to construct a dictionary to represent them sparsely.

The construction of a dictionary can be considered as finding a suitable over-complete basis (frame), in which the signals of interest would be represented with far fewer non-zero coefficients, than in an arbitrary fixed basis such as a Fourier basis. The newly constructed basis is also called a dictionary. This dictionary can be used to assess how well the new signal fits the model represented by the dictionary, and therefore, it can be used as a new classification method.

In our experiment, we assume there are two groups of signals: one for drug resistant mutants, while the other group is non-drug resistant mutants. We construct two dictionaries which could be considered as the models for the resistant and non-resistant groups, respectively. Then, given a new signal (mutant, in our case), both dictionaries are used to represent this signal. By calculating and comparing the reconstruction error, the dictionary with the smaller error indicates that the signal belongs to this category. Theoretically, more than two groups of signals could be treated by defining more than two dictionaries, and such a procedure could be used as a multi-group classification method. The two dictionaries for each set of drug resistance data were constructed and the classification performed as described in[124].

7.7 Acknowledgements

This research was supported, in part, by the National Institutes of Health grant GM062920 (ITW, RWH), and by a fellowship from the Georgia State University Molecular Basis of Disease Program (XY).

AIM 3: Retrieving essential features which might determine whether a mutant is resistant or not to certain drugs

8 IDENTIFYING ESSENTIAL FEATURES FOR THE REPRESENTATIVE MUTANTS FROM DRUG RESISTANT DATA

8.1 ABSTRACT

Drug resistance is one of the most important reasons causing the failure of anti-AIDS treatment. Since the first case of AIDS was found in US in early 1980s, it has been almost three decades now and many scientists and researchers are working on discover its mechanisms. Currently, X-ray crystallography and NMR are two most widely used methods for biologists/chemists to study the structures of the protein-inhibitor complexes. However, due to the HIV virus' rapid replication rate and the lack of proof-reading mechanisms, and moreover, the mutations could be accumulated, there are a large number of different kinds of mutants to study on. Furthermore, since the minimal wet experiment are time and labor consuming, it's necessary to discover a better method to guide biologist/chemists choosing the most potential mutants to research on. In order to solve this problem, we have developed a new algorithm to reveal the most potential mutants from the whole drug resistant mutant database based on our newly proposed unified protein sequence and 3D structure encoding method. This algorithm was tested on genotype-resistance data for mutants of HIV protease and reverse transcriptase and successfully chooses around 200 mutants out of 10K from the whole database.

8.2 Introduction

AIDS (Acquired Immunodeficiency Syndrome) is one of the most severe diseases all over the world and approximately 35.5 million people are living with it by the year 2012[170]. It has been almost three decades since the first case of AIDS was found in US and it's known that the cause of AIDS is HIV

(Human Immunodeficiency Virus)[171]. With thirty years' study, the biological mechanism of the disease is better understood, and more efficient treatment could be offered during the anti-AIDS therapy.

Currently, total of 26 licensed drugs are used in anti-AIDS therapy[9]. All these drugs target to different steps during the HIV life cycle, including entry, reverse transcription, integration and maturation. During the life cycle, HIV protease is the enzyme that is essential for the production of infectious virus[172]. Its inhibitors help blocking the proteolytic activity of the protease, preventing the maturation of the virus[173, 174]. HIV RT functions as converting the viral RNA genome into DNA during the HIV life cycle. It was the first drug target, and the nucleoside analog zidovudine (AZT) was the first FDA approved anti-AIDS drug [175, 176]. During the anti-AIDS treatment, which is often referred as highly active antiretroviral therapy (HAART), three or more antiretroviral drugs choosing from different categories are given to patients during the treatment. The study shows that such treatment could extend the lifespan of the patients[177].

However, since HIV is a member of retrovirus family[178], it has all the characteristics of the retroviruses, and RNA carries its genomic information[178, 179]. Due to the lack of proofreading by reverse transcriptase[33] and high replication rate as many as 10^9 daily[35], drug resistance is one of the most severe problems during the treatment of the AIDS[120, 180]. Moreover, during the anti-viral treatment, drug pressure could cause the selection of the drug resistant strains, and replacement of the wild-type virus[181, 182]. This might cause the failure of the treatment. In this case, understanding the mechanism of the drug resistant is important and could help improve the current anti-AIDS therapy.

Nowadays, several possible mechanisms are studied to explain the drug resistant [183-188]. Most commonly used methods to study the mechanisms are X-ray crystallography and NMR. After obtaining the 3D structure of the protein, scientists study and compare the mutant structure with the wild type to reveal the possible drug resistant mechanisms. However, since HIV has a high mutation rate at about 10^{-4} to 10^{-5} mutations per nucleotide and cycle of replication[189] and the polymorphous gene of

HIV protease and reverse transcriptase, a huge number of mutants are exists. Take HIV protease as an example, even without the inhibitors, mutations of more than thirty different residues associated with protease inhibitors has been reported[120]. Moreover, mutations could be accumulated[190, 191]: single site mutations could combine together to contribute to more drug resistant. For instances, PR20 is the mutant with 20 substitutions of Q7K, L10F, I13V, I15V, D30N, V32I, L33F, E35D, M36I, S37N, I47V, I54L, Q58E, I62V, L63P, A71V, I84V, N88D, L89T and L90M exhibits. It is resistant 1000 folds more than wild type protease to darunavir (DRV) and saquinavir (SQV)[192]. Therefore, a natural request would be: having all those mutants, which ones are the most meaningful mutants for biologists/chemists to study with? By answering this question, it could save both time and money, and faster the process of the study of the drug resistant mechanism.

Mean shift clustering is first introduced in 1975 by Fukunaga and Hostetler[93] in the purpose of seeking the mode of a density function in the given sample set. Fukunaga and Hostetler[93] also suggested that mean shift clustering is an instance of gradient ascent by using decreasing distance functions, which often referred as kernel, from a given point to a point in the sample set. This algorithm started widely used until 1995 when Cheng[94] developing a more generalized formulation of the algorithm. By clarifying the relationship between mean shift and the optimization, the algorithm could potentially be applied on clustering and global optimization problems. Applications of the mean shift algorithm range from image/video segmentation, image representation/retrieval, discontinuity-preserving smoothing[95, 96], higher level tasks like appearance-based clustering[97, 98], tracking including blob tracking[99] and face tracking[100], shape detection and recognition[101], so on and so forth. Afterwards, applications extend to other fields like biology. These applications include analysis of structural variation in genome[102], DNA microarray analysis[103], time-warped gene expression analysis[104], with many other implementations.

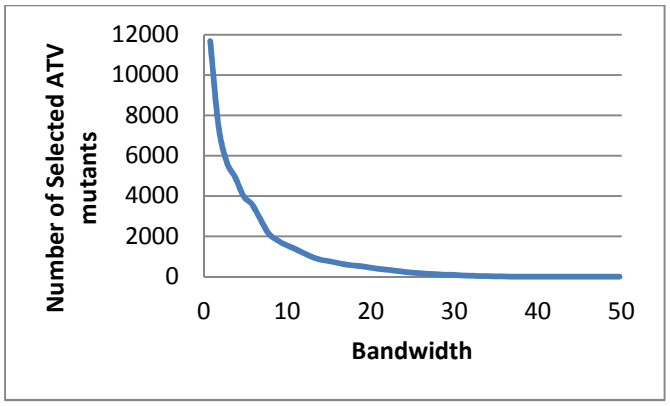
In this paper, we proposed a new algorithm based on the non-parametric iterative mean shift and our recently proposed protein encoding method to extract the most representative drug resistant mutants from the database.

8.3 Experiments and results

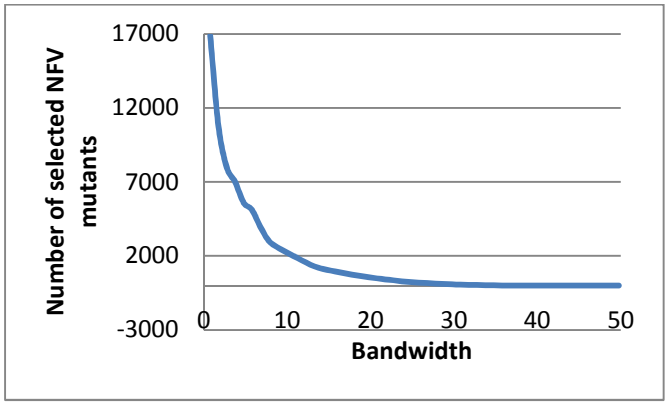
Mean shift clustering, multiple regression and quantile regression were performed on the data for both HIV-1 protease and reverse transcriptase whose sequences and structures were encoded by Delaunay triangulation.

8.3.1 Mean shift clustering on HIV protease inhibitor resistance

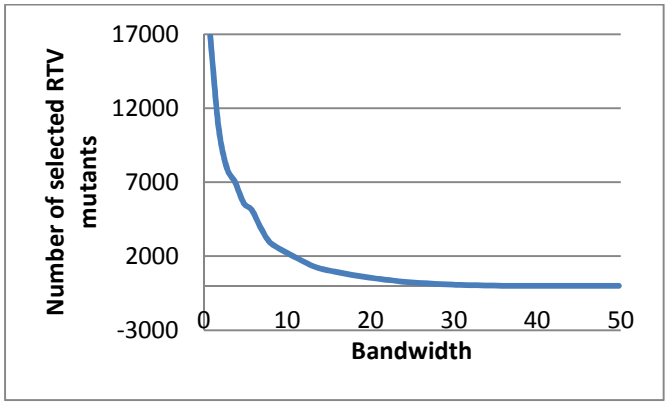
After each of the mutated sequences was represented by a 210-dimensional vector, we performed the mean shift clustering on the drug resistance data to select the most representative mutants. The result shows that the larger the bandwidths set, the smaller number of mutants is selected (Figure 23).


Number of Selected ATV mutants
12000
10000
8000
6000
4000
2000
0
0
10
20
30
40
50
Bandwidth

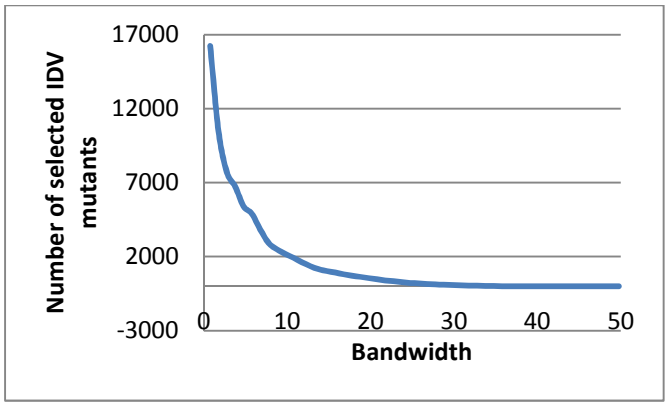
(A)


Number of selected NFV mutants
17000
12000
7000
2000
-3000
0
10
20
30
40
50
Bandwidth

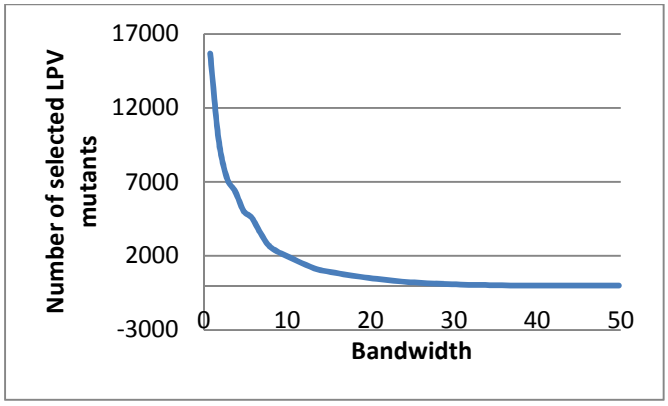
(B)


Number of selected RTV mutants
17000
12000
7000
2000
-3000
0
10
20
30
40
50
Bandwidth

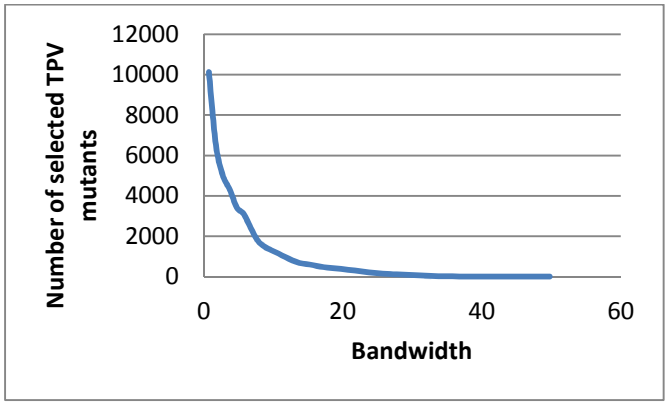
(C)


Number of selected IDV mutants
17000
12000
7000
2000
-3000
0
10
20
30
40
50
Bandwidth

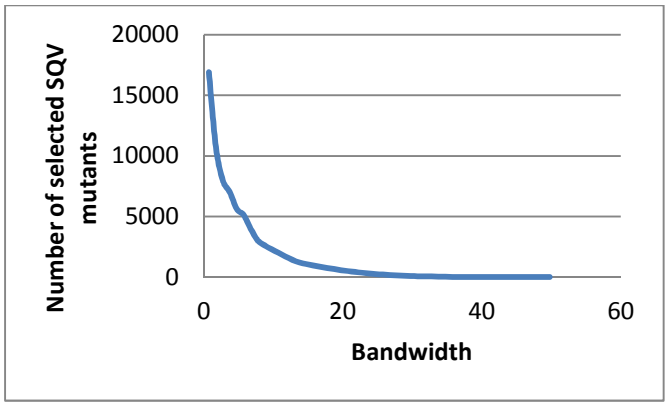
(D)


Number of selected LPV mutants
17000
12000
7000
2000
-3000
0
10
20
30
40
50
Bandwidth

(E)


Number of selected TPV mutants
12000
10000
8000
6000
4000
2000
0
0
20
40
60
Bandwidth

(F)

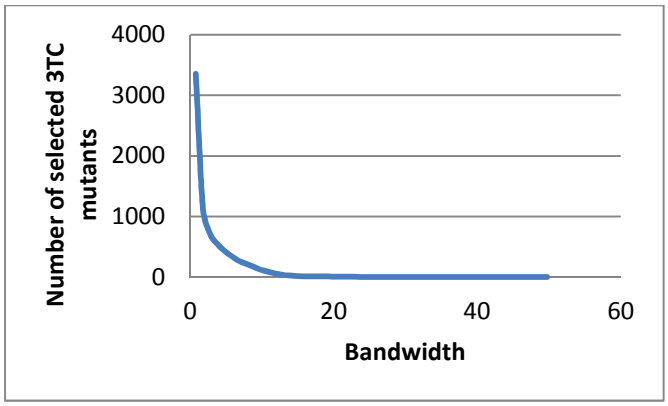


(G)

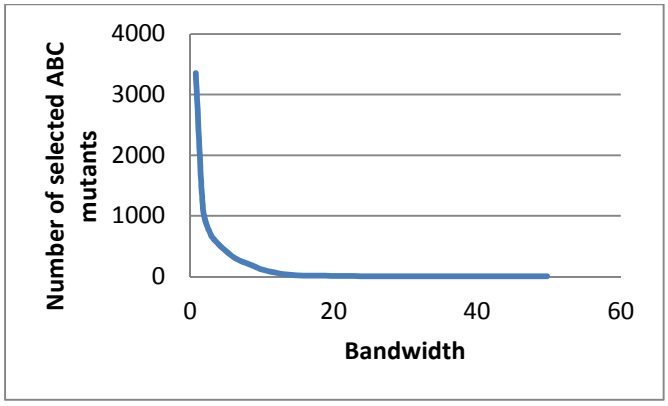
Figure 8.3.1.1 The relationship between the bandwidths and the number of selected mutants. The bandwidth is plotted against the number of selected mutants. The trend line is shown in blue. Plots show regression for drug resistance: (A) ATV, (B) NFV, (C) RTV, (D) IDV, (E) LPV, (F) TPV, and (G) SQV.

8.3.2 Mean shift clustering on HIV reverse transcriptase inhibitors resistance

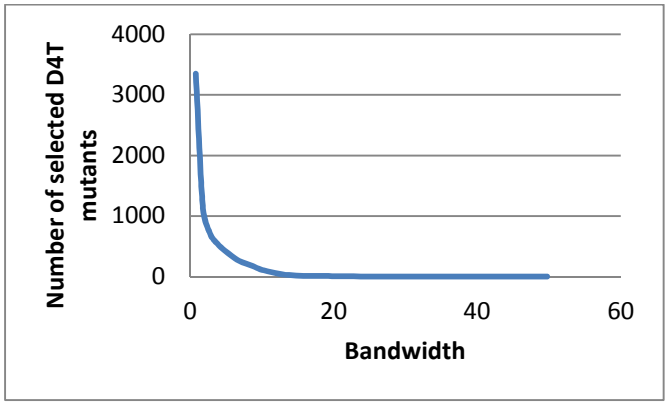
Similarly, mean shift clustering was performed on the drug resistance data for HIV-1 reverse transcriptase inhibitors. The bandwidth and the selected mutants numbers are compared to the reverse transcriptase inhibitors including NRTIs 3TC, ABC, D4T, DDI, TDF and AZT (Figure 24), and NPV, DLV and EFV for NNRTIs (Figure 25).


Number of selected 3TC mutants
4000
3000
2000
1000
0
0
20
40
60
Bandwidth

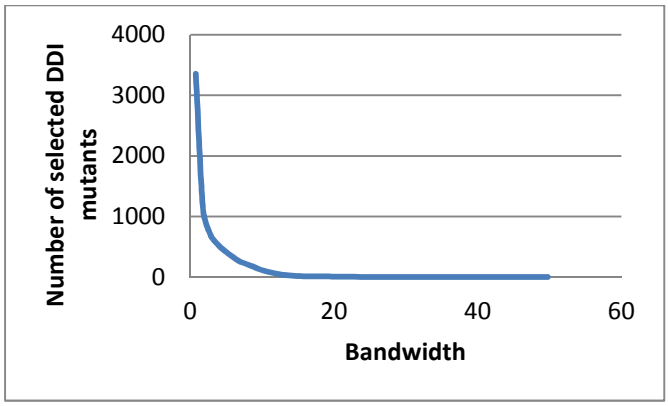
(A)


Number of selected ABC mutants
4000
3000
2000
1000
0
0
20
40
60
Bandwidth

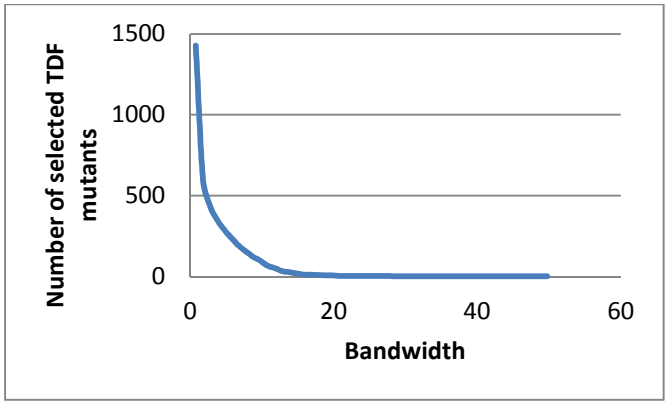
(B)


Number of selected D4T mutants
4000
3000
2000
1000
0
0
20
40
60
Bandwidth

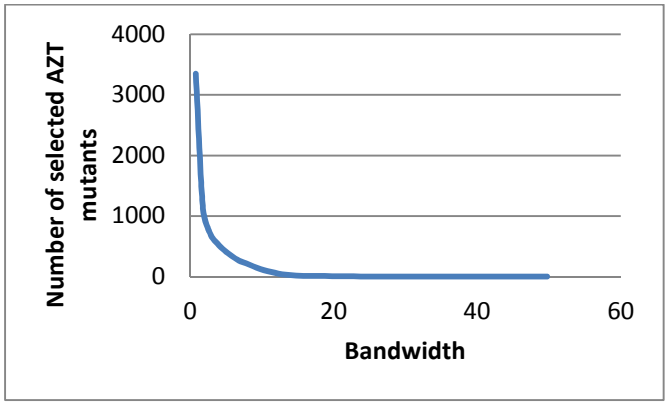
(C)



(D)

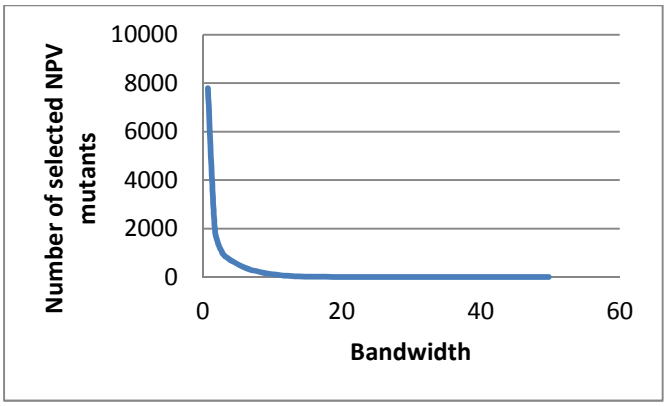


(E)

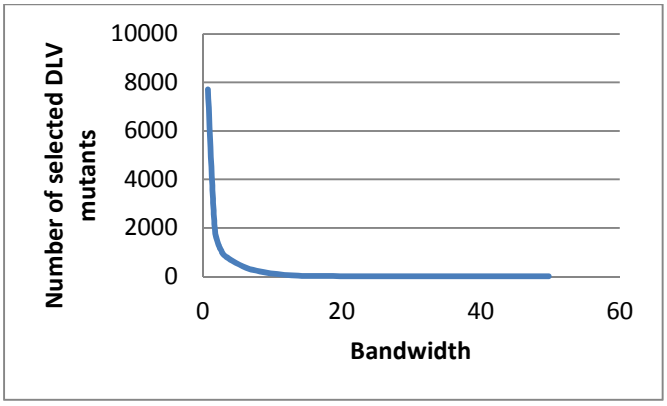


(F)

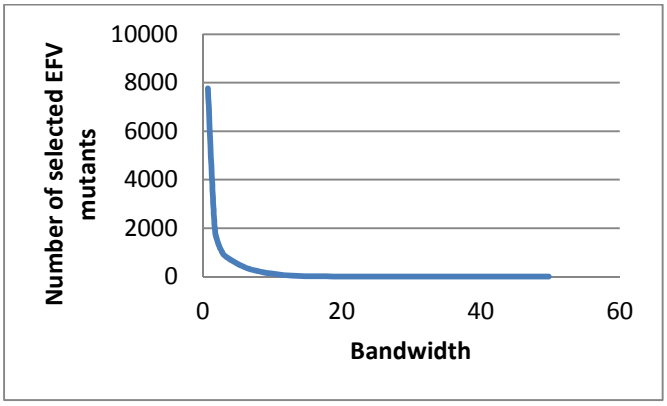
Figure 8.3.2.1 The relationship between the bandwidths and the number of selected mutants. The bandwidth is plotted against the number of selected mutants. The trend line is shown in blue. Plots show regression for drug resistance: (A) 3TC, (B) ABC, (C) D4T, (D) DDI, (E) TDF and (F) AZT.



(A)



(B)

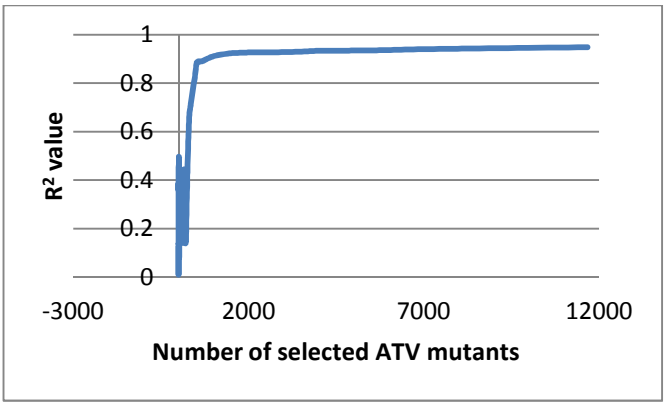


(C)

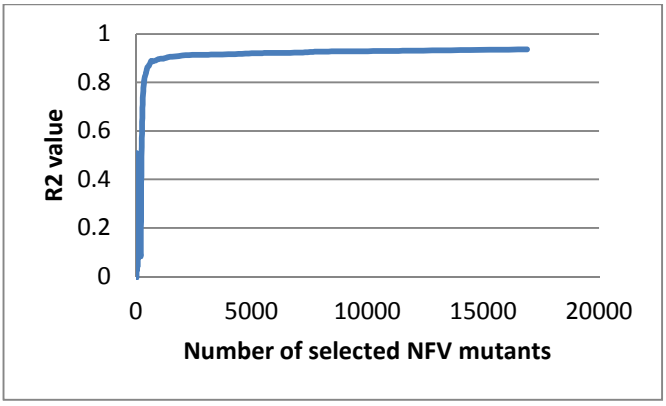
Figure 8.3.2.2 The relationship between the bandwidths and the number of selected mutants. The bandwidth is plotted against the number of selected mutants. The trend line is shown in blue. Plots show regression for drug resistance: (A) NPV, (B) DLV and (C) EFV.

8.3.3 Multiple regression on HIV protease inhibitor resistance

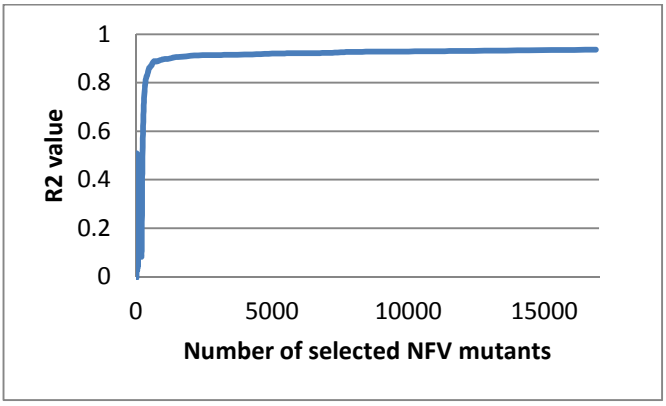
Afterwards, a multiple regression was applied to the selected mutants to evaluate the selected results. The R^2 values for relative resistance were plotted against the number of selected mutants as shown in (Figure 26) for the PR inhibitors ATV, NFV, RTV, IDV, LPV, TPV and SQV.


1
0.8
0.6
0.4
0.2
0
R² value
-3000
2000
7000
12000
Number of selected ATV mutants

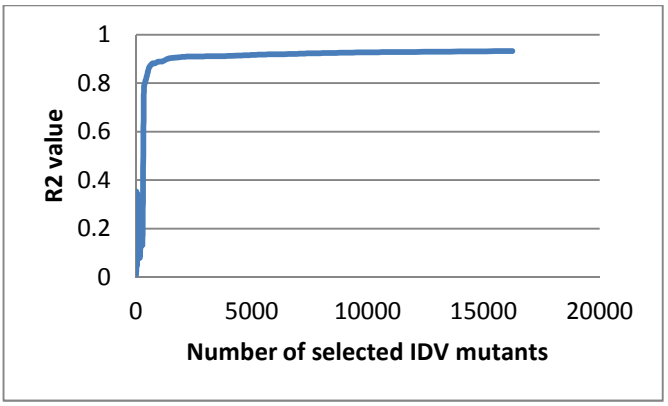
(A)


1
0.8
0.6
0.4
0.2
0
R2 value
0
5000
10000
15000
20000
Number of selected NFV mutants

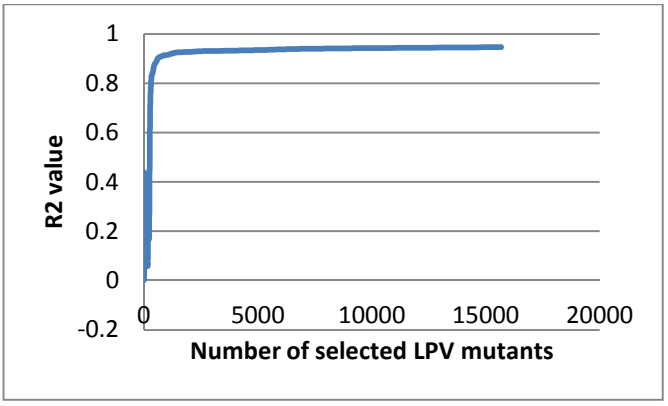
(B)


1
0.8
0.6
0.4
0.2
0
R2 value
0
5000
10000
15000
Number of selected NFV mutants

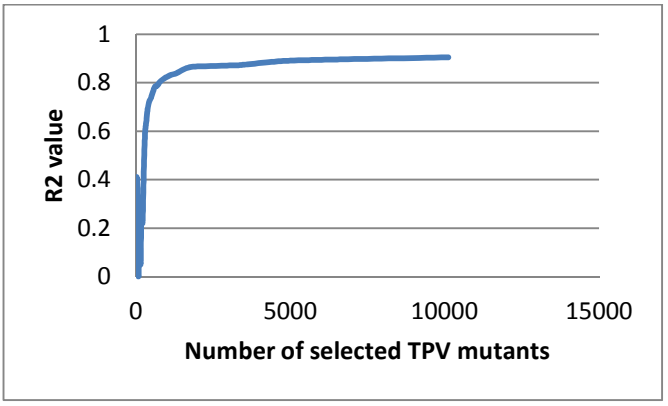
(C)


1
0.8
0.6
0.4
0.2
0
R2 value
0
5000
10000
15000
20000
Number of selected IDV mutants

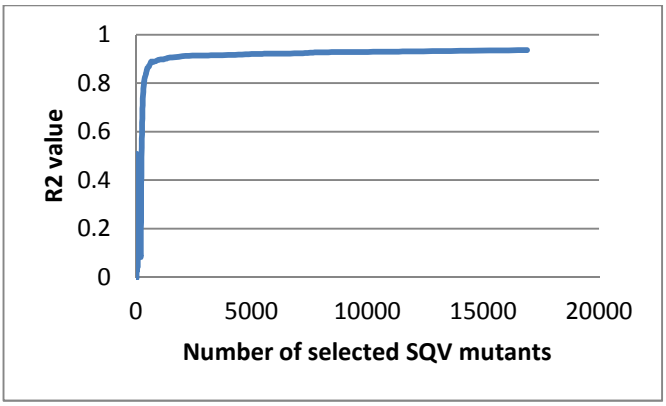
(D)


1
0.8
0.6
0.4
0.2
0
-0.2
R2 value
0
5000
10000
15000
20000
Number of selected LPV mutants

(E)


1
0.8
0.6
0.4
0.2
0
R2 value
0
5000
10000
15000
Number of selected TPV mutants

(F)

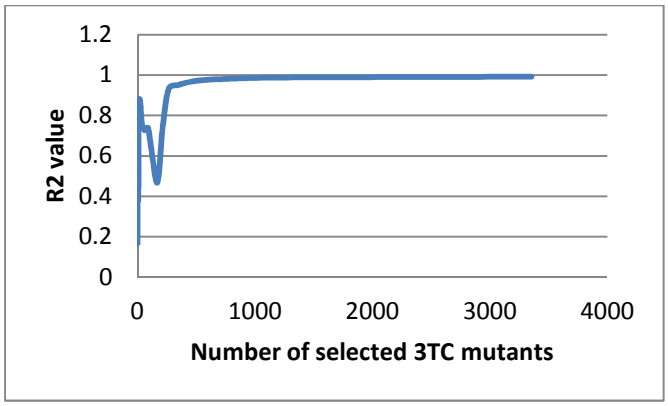


(G)

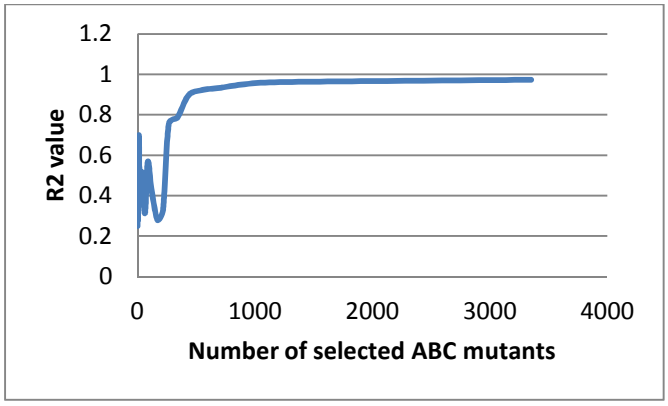
Figure 8.3.3.1 The relationship between the multiple regression results and the number of selected mutants. The R^2 is plotted against the number of selected mutants. The trend line is shown in blue. Plots show regression for drug resistance: (A) ATV, (B) NFV, (C) RTV, (D) IDV, (E) LPV, (F) TPV, and (G) SQV.

8.3.4 Multiple regression on HIV reverse transcriptase inhibitor resistance

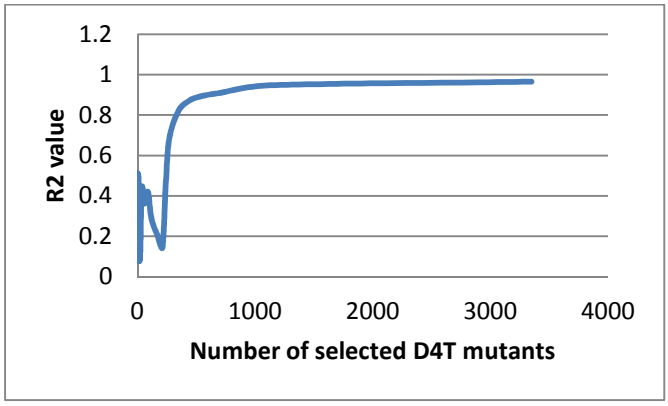
Multiple regression analysis was performed similarly on genotype-phenotype data for drugs inhibiting HIV-1 RT. The R^2 values for relative resistance were plotted against the number of selected mutants as shown in for the RT inhibitors including NRTIs 3TC, ABC, D4T, DDI, TDF and AZT (Figure 27), and NPV, DLV and EFV for NNRTIs (Figure 28).


1.2
1
0.8
0.6
0.4
0.2
0
R2 value
0
1000
2000
3000
4000
Number of selected 3TC mutants

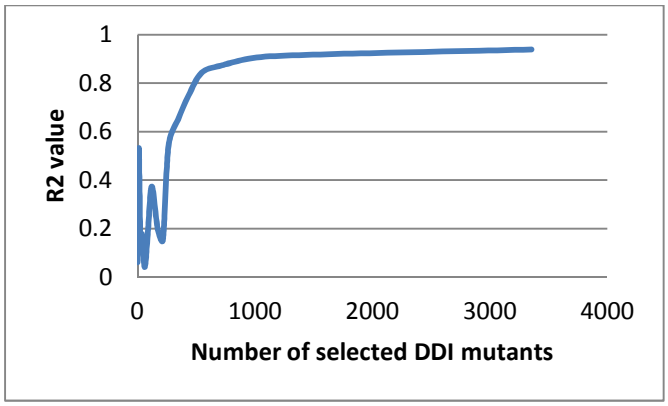
(A)


1.2
1
0.8
0.6
0.4
0.2
0
R2 value
0
1000
2000
3000
4000
Number of selected ABC mutants

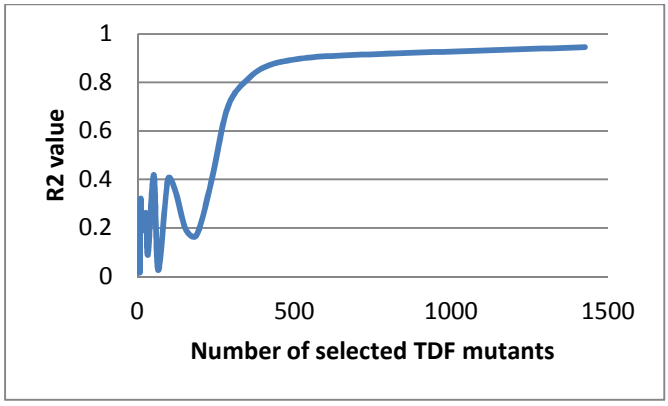
(B)


1.2
1
0.8
0.6
0.4
0.2
0
R2 value
0
1000
2000
3000
4000
Number of selected D4T mutants

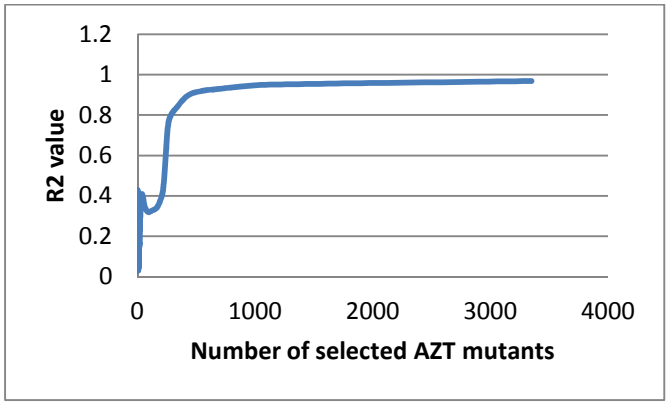
(C)



(D)

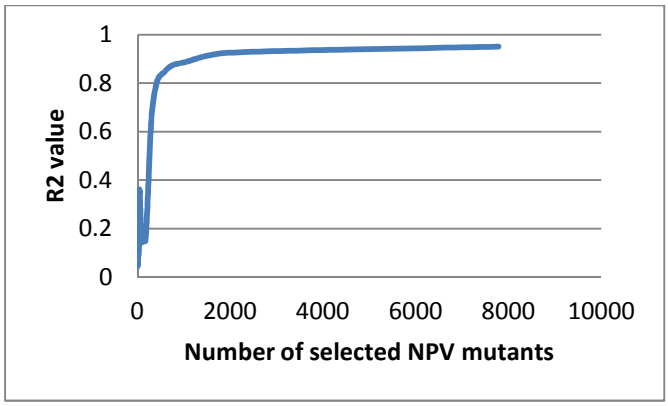


(E)

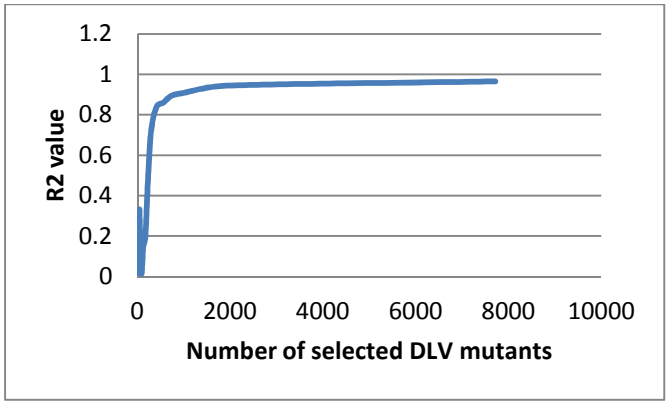


(F)

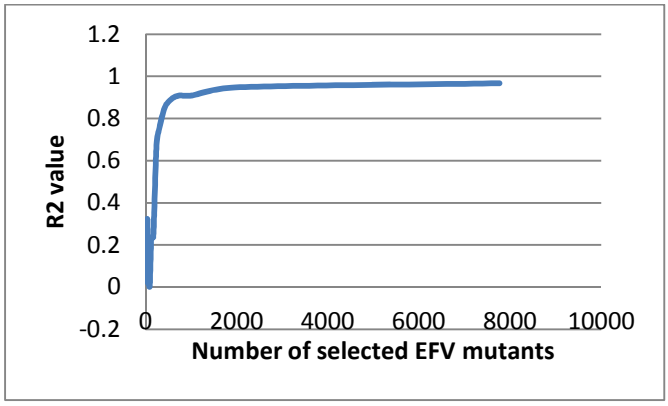
Figure 8.3.4.1 The relationship between the multiple regression results and the number of selected mutants. The R^2 is plotted against the number of selected mutants. The trend line is shown in blue. Plots show regression for drug resistance: (A) 3TC, (B) ABC, (C) D4T, (D) DDI, (E) TDF and (F) AZT.



(A)



(B)



(C)

Figure 8.3.4.2 The relationship between the multiple regression results and the number of selected mutants. The R^2 is plotted against the number of selected mutants. The trend line is shown in blue. Plots show regression for drug resistance: (A) NPV, (B) DLV and (C) EFV.

8.3.5 Bandwidth selection and multiple regression on HIV-1 PR inhibitor resistance

Based on the above experiments, the relationships between the bandwidth, the number of selected mutants and the multiple regression results are shown. Following experiments were performed to find the balance of the number of selected mutants and the R^2 results. The results show that for both HIV-1 PR and HIV-1 RT, about 200~300 mutants are needed to represent all the drug resistance data (Table 25~27).

Table 8.3.5.1 The bandwidth, number of selected mutants and R2 on HIV-1 PR

	Bandwidth	Number of selected mutants	R^2
ATV	22	344	0.7284
NFV	23.75	288	0.6993
RTV	23.75	270	0.7922
IDV	23	321	0.7791
LPV	23.50	284	0.7623
TPV	19	412	0.7114
SQV	23.75	278	0.7391

Table 8.3.5.2 The bandwidth, number of selected mutants and R2 on HIV-1 RT NRTIs

	Bandwidth	Number of selected mutants	R^2
3TC	13.75	26	0.8317
ABC	7	255	0.7507
D4T	6.75	266	0.6534
DDI	5.75	343	0.6651
TDF	4.75	286	0.6914
AZT	6.75	254	0.7809

Table 8.3.5.3 The bandwidth, number of selected mutants and R2 on HIV-1 RT NNRTIs

	Bandwidth	Number of selected mutants	R^2
NPV	6.75	307	0.6693
DLV	6.75	298	0.7276
EFV	7.75	242	0.6733

8.3.6 Quantile information analysis on HIV-1 PR inhibitor resistance

In order to further analysis the mutants selected by mean shift, all the drug resistant mutants were grouped and separated into 10 bins based on their drug resistance value. Both the total number of mutants and the selected number of mutants are counted and recorded in each corresponding table.

For ATV, their resistant values are ranging from 0 to 700. Therefore, those mutants with resistant value between 0 and 70 were put into bin I, those with resistant value between above 70 and below 140 were put into bin II, and so on.

For NFV, their resistant values are ranging from 0 to 600. Therefore, those mutants with resistant value between 0 and 60 were put into bin I, those with resistant value between above 60 and below 120 were put into bin II, and so on.

For RTV, their resistant values are ranging from 0 to 800. Therefore, those mutants with resistant value between 0 and 80 were put into bin I, those with resistant value between above 80 and below 160 were put into bin II, and so on.

For IDV, their resistant values are ranging from 0 to 500. Therefore, those mutants with resistant value between 0 and 50 were put into bin I, those with resistant value between above 50 and below 100 were put into bin II, and so on.

For LPV, their resistant values are ranging from 0 to 500. Therefore, those mutants with resistant value between 0 and 50 were put into bin I, those with resistant value between above 50 and below 100 were put into bin II, and so on.

For TPV, their resistant values are ranging from 0 to 200. Therefore, those mutants with resistant value between 0 and 20 were put into bin I, those with resistant value between above 20 and below 40 were put into bin II, and so on.

For SQV, their resistant values are ranging from 0 to 1000. Therefore, those mutants with resistant value between 0 and 100 were put into bin I, those with resistant value between above 100 and below 200 were put into bin II, and so on.

The table 4-10 shows the total number of mutants in the bin before and after selection.

Table 8.3.6.1 Comparison of number of selected ATV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	9454	189
II	1179	36
III	844	18
IV	200	9
V	39	10
VI	3	1
VII	34	3
VIII	129	1
IX	0	0
X	202	24

Table 8.3.6.2 Comparison of number of selected NFV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	13711	183
II	2126	55
III	540	22
IV	357	7
V	21	4
VI	256	1
VII	2	0
VIII	0	0
IX	9	1
X	523	15

Table 8.3.6.3 Comparison of number of selected RTV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	12220	151
II	1589	34
III	918	11
IV	300	6
V	304	7
VI	0	0
VII	22	2
VIII	0	0
IX	0	0
X	1299	59

Table 8.3.6.4 Comparison of number of selected IDV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	14885	246
II	1101	35
III	511	10
IV	216	14
V	14	4
VI	0	0
VII	8	1
VIII	12	1
IX	0	0
X	99	10

Table 8.3.6.5 Comparison of number of selected LPV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	11630	152
II	2087	62
III	1393	31
IV	200	13
V	333	8
VI	26	1
VII	153	3
VIII	3	2
IX	0	0
X	444	12

Table 8.3.6.6 Comparison of number of selected TPV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	9921	366
II	87	16
III	0	0
IV	0	0
V	1	1
VI	0	0
VII	0	0
VIII	0	0
IX	0	0
X	219	29

Table 8.3.6.7 Comparison of number of selected SQV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	14746	223
II	910	19
III	107	0
IV	28	3
V	94	2
VI	132	2
VII	0	0
VIII	1	1
IX	0	0
X	1100	28

8.3.7 Quantile information analysis on HIV-1 reverse transcriptase inhibitor resistance (NRTIs)

In order to further analysis the mutants selected by mean shift, all the drug resistant mutants were grouped and separated into 10 bins based on their drug resistance value. Both the total number of mutants and the selected number of mutants are counted and recorded in each corresponding table.

For 3TC, their resistant values are ranging from 0 to 200. Therefore, those mutants with resistant value between 0 and 20 were put into bin I, those with resistant value between above 20 and below 40 were put into bin II, and so on.

For ABC, their resistant values are ranging from 0 to 170. Therefore, those mutants with resistant value between 0 and 17 were put into bin I, those with resistant value between above 17 and below 34 were put into bin II, and so on.

For D4T, their resistant values are ranging from 0 to 26. Therefore, those mutants with resistant value between 0 and 2.6 were put into bin I, those with resistant value between above 2.6 and below 5.2 were put into bin II, and so on.

For DDI, their resistant values are ranging from 0 to 28. Therefore, those mutants with resistant value between 0 and 2.8 were put into bin I, those with resistant value between above 2.8 and below 5.6 were put into bin II, and so on.

For TDF, their resistant values are ranging from 0 to 500. Therefore, those mutants with resistant value between 0 and 50 were put into bin I, those with resistant value between above 50 and below 100 were put into bin II, and so on.

For AZT, their resistant values are ranging from 0 to 400. Therefore, those mutants with resistant value between 0 and 40 were put into bin I, those with resistant value between above 40 and below 80 were put into bin II, and so on.

The table 11-16 shows the total number of mutants in the bin before and after selection.

Table 8.3.7.1 Comparison of number of selected 3TC mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	2711	11
II	14	0
III	1	0
IV	1	0
V	73	0
VI	57	0
VII	54	1
VIII	45	0
IX	88	0
X	1806	14

Table 8.3.7.2 Comparison of number of selected ABC mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	4780	241
II	65	13
III	0	0
IV	0	0
V	0	0
VI	0	0
VII	0	0
VIII	0	0
IX	0	0
X	1	1

Table 8.3.7.3 Comparison of number of selected D4T mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	3791	188
II	948	51
III	23	10
IV	14	7
V	37	4
VI	17	2
VII	1	1
VIII	4	2
IX	8	1
X	2	0

Table 8.3.7.4 Comparison of number of selected DDI mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	4603	314
II	194	17
III	25	5
IV	4	1
V	7	2
VI	9	2
VII	2	0
VIII	1	0
IX	3	1
X	1	1

Table 8.3.7.5 Comparison of number of selected TDF mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	2001	265
II	1	1
III	0	0
IV	0	0
V	0	0
VI	0	0
VII	0	0
VIII	0	0
IX	0	0
X	2	1

Table 8.3.7.6 Comparison of number of selected AZT mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	4079	142
II	94	19
III	253	13
IV	27	4
V	7	3
VI	30	5
VII	19	5
VIII	164	3
IX	6	0
X	168	35

8.3.8 Quantile information analysis on HIV-1 reverse transcriptase inhibitor resistance (NNRTIs)

In order to further analysis the mutants selected by mean shift, all the drug resistant mutants were grouped and separated into 10 bins based on their drug resistance value. Both the total number of mutants and the selected number of mutants are counted and recorded in each corresponding table.

For NPV, their resistant values are ranging from 0 to 400. Therefore, those mutants with resistant value between 0 and 40 were put into bin I, those with resistant value between above 40 and below 80 were put into bin II, and so on.

For DLV, their resistant values are ranging from 0 to 200. Therefore, those mutants with resistant value between 0 and 20 were put into bin I, those with resistant value between above 20 and below 40 were put into bin II, and so on.

For EFV, their resistant values are ranging from 0 to 400. Therefore, those mutants with resistant value between 0 and 40 were put into bin I, those with resistant value between above 40 and below 80 were put into bin II, and so on.

The table 17-19 shows the total number of mutants in the bin before and after selection.

Table 8.3.8.1 Comparison of number of selected NPV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	9898	157
II	157	17
III	114	9
IV	56	7
V	94	9
VI	169	7
VII	1	1
VIII	293	30
IX	1	0
X	584	66

Table 8.3.8.2 Comparison of number of selected DLV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	9476	198
II	241	24
III	587	12
IV	35	7
V	155	4
VI	20	3
VII	0	0
VIII	73	3
IX	9	3
X	703	43

Table 8.3.8.3 Comparison of number of selected EFV mutants in each bin

Bin	Number of total mutants	Number of selected mutants
I	9907	172
II	116	14
III	166	10
IV	24	1
V	42	2
VI	132	2
VII	26	4
VIII	48	6
IX	2	1
X	891	32

8.4 DISCUSSION

The serious problem of drug resistance arising during therapy of HIV-infected individuals can caused the failure of the treatment. Many scientists are working on revealing the drug resistant mecha-

nism using X-ray crystallography or NMR. However, since there are a large number of mutants, it is difficult to choose which mutant to research on.

In this experiment, we have developed new selection algorithm based on a simple graph representation of protein structure to solve this problem. The protein structure is 3-D and could be efficiently represented by Delaunay triangulation[37]. Based on this encoding method, a mean shift was applied to select the most representative mutants. Multiple linear regression was performed to evaluate the selection results.

This selection algorithm works well on selecting drug resistant mutants from both HIV protease and reverse transcriptase inhibitors drug resistant mutants. Among all the mutants, around 200 most potential mutants were selected. The multiple linear regression was applied on these selected mutants' drug resistant value, and most of the R2 were above 0.70.

9 Future work and summaries

9.1 Future work

This research work presented the study of two interesting questions: first, is there a possible way to predict drug resistance directly from protein sequence; second, are there more important drug resistant mutants among all the datasets? In order to solve these two questions, Delaunay triangulation, sparse representation as well as mean shift were applied and the according results were demonstrated in the above sections. In the future, more possible directions could be further studied to continue this research work. More details would be discussed in the following.

Speed Up Of Sparse Representation Based Classifier: In this study, the sparse dictionary algorithm was coded using Matlab. As we already known that, Matlab has the strength of visualization and matrix calculation. However, it has some limitation on processing speed improvement. As we've already discussed in Chapter 6, the running time of sparse dictionary is about 300 seconds, comparing to around 20 seconds for those of SVM linear and ANN. In order to speed up the algorithms, C++ could be used. Moreover, parallel computing methods could also be included to further speed up the algorithm, CUDA or OpenMP for instances.

Multi-Class Classifier: Based on the concepts of the sparse dictionary based classifier, multi-class classifier could be further developed. In order to solving this problem, more dictionaries could be trained based on the training data. After obtaining these dictionaries, given one testing data, the error could be calculated for each dictionary. The testing data might belong to the dictionary with the smallest the error.

Sparse based mutant selection algorithm: When representing the encoded mutants in sparse basis, only very few items in this basis are non-zero. For those non-zero items in the basis, we could consider them are important features of the encoded mutants. If we further calculate the most frequent

appeared items in the basis, the items with the most frequency would be the most important mutants. Therefore, if we trace back to the original mutants, those might be the potential ones which could be further research on.

9.2 Summaries

Drug resistance is one of the most important reasons causing the failure of anti-AIDS treatment. Normally during the treatment, the patients are tested to choose the more efficient combination of drugs. However, such experiments need two weeks before the results could be obtained. In order to shorten patients' waiting time and money, computational methods could be used to predict the drug resistance to certain inhibitors.

In this work, we first proposed an effective triangulation-based encoding method. By applying this method, three-dimensional protein structures could be reduced to small constant-sized representations which are suitable for most machine learning algorithms. When using this method to encode protein structures, the information of the kinds of adjacent residues in the triangulation is sufficient for accurate classification and regression analysis. This encoding method was applied to predict of resistance to HIV PR and RT inhibitors and gave high accuracy. Moreover, the results of correlation coefficients in regression analysis are also impressive.

Following that, incorporated with the triangulation-based encoding method, we proposed and evaluated a new classification method to predict the drug resistance for both HIV-1 PR and HIV-1 RT antiviral inhibitors from genotype data. This classification algorithm is based on the sparse representation theory. In this algorithm, we learn the characteristics of resistant and non-resistant mutants of the HIV-1 protease by constructing two over-complete dictionaries. Then, given the sequence of a new mutant, we measure how accurately this new sequence can be represented by the two dictionaries. The category of the dictionary with smaller error is assigned to the new mutant. This classification method was tested on four HIV PR inhibitors and three HIV RT inhibitors and produced high accuracy. According to

the results, this new method is able to predict drug resistance with high accuracy and can distinguish between drug resistant and non-resistant sequences significantly better than the other methods.

After that, among all the drug resistant mutants, mean shift algorithm was applied to retrieve most important drug resistant mutants. The result successfully selected around 300 mutants out of 10k. Furthermore, when quantify all the mutants, for the most drug resistant group, around 30 mutants were selected. These most important drug resistant mutants are more interesting for biologists/chemists to further research on.

In this work, we presented some evidence obtained by our experimental study. This may indicate that by using more efficient protein encoding algorithm and more accurate predicting methods, drug resistance could be determined directly from protein sequences. Moreover, more important drug resistance mutants could be retrieved from the drug resistant database.

10 REFERENCES

1. Croft, W.B., D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. 2010: Addison-Wesley Reading.
2. Palmer, R., et al., *Artificial economic life: a simple model of a stockmarket.* Physica D: Nonlinear Phenomena, 1994. **75**(1): p. 264-274.
3. Kononenko, I., *Machine learning for medical diagnosis: history, state of the art and perspective.* Artificial Intelligence in medicine, 2001. **23**(1): p. 89-109.
4. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-96.
5. Schumann, R.R., et al., *Structure and function of lipopolysaccharide binding protein.* Science, 1990. **249**(4975): p. 1429-1431.
6. HIV/AIDS, J.U.N.P.o., *AIDS Scorecards: Overview: UNAIDS Report on the Global AIDS Epidemic 2010*. 2010: UNAIDS.
7. Organization, W.H., *Global HIV/AIDS response: epidemic update and health sector progress towards universal access: progress report 2011.* Geneva, Switzerland: World Health Organization, 2011.
8. Henkel, J., *Attacking AIDS with a'Cocktail'Therapy.* FDA consumer, 1999. **33**(4).
9. De Clercq, E., *Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV.* International journal of antimicrobial agents, 2009. **33**(4): p. 307-320.
10. Menéndez-Arias, L., *Molecular basis of human immunodeficiency virus drug resistance: an update.* Antiviral research, 2010. **85**(1): p. 210-231.
11. Turner, B.G. and M.F. Summers, *Structural biology of HIV.* Journal of molecular biology, 1999. **285**(1): p. 1-32.
12. Miller, M., et al., *Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution.* Science, 1989. **246**(4934): p. 1149.
13. Navia, M.A.a.F., P.M.D. and McKeever, B.M. and Leu, C.T. and Heimbach, J.C. and Herber, W.K. and Sigal, I.S. and Darke, P.L. and Springer, J.P, *Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1.* Nature, 1989. **337**(6208).
14. Louis, J., et al., *HIV-1 protease: structure, dynamics, and inhibition.* Advances in pharmacology (San Diego, Calif.), 2007. **55**: p. 261.
15. Darke, P., et al., *HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins.* Biochemical and biophysical research communications(Print), 1988. **156**(1): p. 297-303.
16. Louis, J.a.W., IT and Tozser, J. and Clore, GM and Gronenborn, AM, *HIV-1 protease: maturation, enzyme specificity, and drug resistance.* Advances in pharmacology (San Diego, Calif.), 2000. **49**: p. 111.
17. Wlodawer, A. and J. Vondrasek, *INHIBITORS OF HIV-1 PROTEASE: A Major Success of Structure-Assisted Drug Design 1.* Annual review of biophysics and biomolecular structure, 1998. **27**(1): p. 249-284.
18. Karacostas, V., et al., *Human immunodeficiency virus-like particles produced by a vaccinia virus expression vector.* Proceedings of the National Academy of Sciences, 1989. **86**(22): p. 8964.
19. Roberts, N., et al., *Rational design of peptide-based HIV proteinase inhibitors.* Science, 1990. **248**(4953): p. 358.

20. Esnouf, R., et al., *Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors.* Nature Structural & Molecular Biology, 1995. **2**(4): p. 303-308.
21. Hsiou, Y., et al., *Structure of unliganded HIV-1 reverse transcriptase at 2.7 Å resolution: implications of conformational changes for polymerization and inhibition mechanisms.* Structure, 1996. **4**(7): p. 853-860.
22. Rodgers, D., et al., *The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1.* Proceedings of the National Academy of Sciences, 1995. **92**(4): p. 1222-1226.
23. Kohlstaedt, L., et al., *Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor.* Science, 1992. **256**(5065): p. 1783-1790.
24. Huang, H., et al., *Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance.* Science, 1998. **282**(5394): p. 1669-1675.
25. Sarafianos, S.G., et al., *Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition.* Journal of molecular biology, 2009. **385**(3): p. 693-713.
26. Ren, J., et al., *Structural mechanisms of drug resistance for mutations at codons 181 and 188 in HIV-1 reverse transcriptase and the improved resilience of second generation non-nucleoside inhibitors.* Journal of molecular biology, 2001. **312**(4): p. 795-805.
27. Tu, X., et al., *Structural basis of HIV-1 resistance to AZT by excision.* Nature structural & molecular biology, 2010. **17**(10): p. 1202-1209.
28. Balzarini, J. and E. De Clercq, *Analysis of inhibition of retroviral reverse transcriptase.* Methods in enzymology, 1996. **275**: p. 472.
29. Balzarini, J. and E. De Clercq, *Biochemical pharmacology of nucleoside and non-nucleoside analogues active against HIV reverse transcriptase.* status: published, 1998.
30. Sluis-Cremer, N., N.A. Temiz, and I. Bahar, *Conformational changes in HIV-1 reverse transcriptase induced by nonnucleoside reverse transcriptase inhibitor binding.* Current HIV research, 2004. **2**(4): p. 323.
31. de Béthune, M.-P., *Non-nucleoside reverse transcriptase inhibitors (NNRTIs), their discovery, development, and use in the treatment of HIV-1 infection: a review of the last 20 years (1989–2009).* Antiviral research, 2010. **85**(1): p. 75-90.
32. Roberts, J.D., K. Bebenek, and T.A. Kunkel, *The accuracy of reverse transcriptase from HIV-1.* Science, 1988. **242**(4882): p. 1171-1173.
33. Ji, J. and L.A. Loeb, *Fidelity of HIV-1 reverse transcriptase copying RNA in vitro.* Biochemistry, 1992. **31**(4): p. 954-958.
34. Perelson, A.S., et al., *HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time.* Science, 1996. **271**(5255): p. 1582-1586.
35. Ho, D.D., et al., *Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection.* Nature, 1995. **373**(6510): p. 123-126.
36. database, S.H. *HIV Stanford drug resistance database*. March 8th, 2009; Available from: http://hivdb.stanford.edu/cgi-bin/PIResiNote.cgi.
37. Tie, Y., et al., *High resolution crystal structures of HIV-1 protease with a potent non-peptide inhibitor (UIC-94017) active against multi-drug-resistant clinical strains.* Journal of molecular biology, 2004. **338**(2): p. 341-352.
38. Liu, F., et al., *Kinetic, stability, and structural changes in high-resolution crystal structures of HIV-1 protease with drug-resistant mutations L24I, I50V, and G73S.* Journal of molecular biology, 2005. **354**(4): p. 789-800.

39. Kovalevsky, A., et al., *Effectiveness of nonpeptide clinical inhibitor TMC-114 on HIV-1 protease with highly drug resistant mutations D30N, I50V, and L90M.* J. Med. Chem, 2006. **49**(4): p. 1379-1387.
40. Liu, F., *Kinetic and crystallographic studies of drug-resistant mutants of HIV-1 protease: Insights into the drug resistance mechanisms*. 2006.
41. Nikolenko, G., et al., *Mechanisms of HIV-1 drug resistance to nucleoside and nonnucleoside reverse transcriptase inhibitors.* Molecular Biology, 2011. **45**(1): p. 93-109.
42. Ceccherini-Silberstein, F., et al., *Characterization and structural analysis of novel mutations in human immunodeficiency virus type 1 reverse transcriptase involved in the regulation of resistance to nonnucleoside inhibitors.* Journal of virology, 2007. **81**(20): p. 11507-11519.
43. Tambuyzer, L., et al., *Short communication Compilation and prevalence of mutations associated with resistance to non-nucleoside reverse transcriptase inhibitors.* Antiviral therapy, 2009. **14**: p. 103-109.
44. Cheung, P., B. Wynhoven, and P. Harrigan, *2004: which HIV-1 drug resistance mutations are common in clinical practice?* AIDS reviews, 2004. **6**(2): p. 107.
45. Jonckheere, H., et al., *Resistance of HIV-1 reverse transcriptase against [2', 5'-bis-O-(tert-butyldimethylsilyl)-3'-spiro-5''-(4''-amino-1'', 2''-oxathiole-2'', 2''-dioxide)](TSAO) derivatives is determined by the mutation Glu138--> Lys on the p51 subunit.* Journal of Biological Chemistry, 1994. **269**(41): p. 25255-25258.
46. Nikolenko, G.N., et al., *Mutations in the connection domain of HIV-1 reverse transcriptase increase 3′-azido-3′-deoxythymidine resistance.* Proceedings of the National Academy of Sciences, 2007. **104**(1): p. 317-322.
47. Santos, A.F., et al., *Conservation patterns of HIV-1 RT connection and RNase H domains: identification of new mutations in NRTI-treated patients.* PLoS One, 2008. **3**(3): p. e1781.
48. Waters, J.M., et al., *Mutations in the thumb-connection and RNase H domain of HIV type-1 reverse transcriptase of antiretroviral treatment-experienced patients.* Antivir Ther, 2009. **14**(2): p. 231-239.
49. Archer, R.H., et al., *Mutants of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase resistant to nonnucleoside reverse transcriptase inhibitors demonstrate altered rates of RNase H cleavage that correlate with HIV-1 replication fitness in cell culture.* Journal of virology, 2000. **74**(18): p. 8390-8401.
50. Zhang, J., et al., *Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance.* Proceedings of the National Academy of Sciences, 2010. **107**(4): p. 1321-1326.
51. Sebastian, J. and H. Faruki, *Update on HIV resistance and resistance testing.* Medicinal research reviews, 2004. **24**(1): p. 115-125.
52. Ravela, J., et al., *HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms.* Journal of acquired immune deficiency syndromes (1999), 2003. **33**(1): p. 8.
53. Meynard, J.-L., et al., *Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial.* Aids, 2002. **16**(5): p. 727-736.
54. Van Laethem, K., et al., *A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients.* Antiviral therapy, 2002. **7**(2): p. 123-9.
55. Rhee, S.-Y., et al., *Human immunodeficiency virus reverse transcriptase and protease sequence database.* Nucleic acids research, 2003. **31**(1): p. 298-303.
56. Schmidt, B., et al., *Simple algorithm derived from a geno-/phenotypic database to predict HIV-1 protease inhibitor resistance.* Aids, 2000. **14**(12): p. 1731-1738.

57. Zazzi, M., et al., *Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype.* Journal of Antimicrobial Chemotherapy, 2004. **53**(2): p. 356-360.
58. Liu, T.F. and R.W. Shafer, *Web resources for HIV type 1 genotypic-resistance test interpretation.* Clinical infectious diseases, 2006. **42**(11): p. 1608-1618.
59. Tural, C., et al., *Clinical utility of HIV-1 genotyping and expert advice: the Havana trial.* Aids, 2002. **16**(2): p. 209-18.
60. Lengauer, T. and T. Sing, *Bioinformatics-assisted anti-HIV therapy.* Nature Reviews Microbiology, 2006. **4**(10): p. 790-797.
61. Shafer, R.W., *Genotypic testing for human immunodeficiency virus type 1 drug resistance.* Clinical Microbiology Reviews, 2002. **15**(2): p. 247-277.
62. Beerenwinkel, N., et al., *Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype.* Proceedings of the National Academy of Sciences, 2002. **99**(12): p. 8271-8276.
63. Rhee, S.-Y., et al., *Genotypic predictors of human immunodeficiency virus type 1 drug resistance.* Proceedings of the National Academy of Sciences, 2006. **103**(46): p. 17355-17360.
64. Vermeiren, H., et al., *Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling.* Journal of virological methods, 2007. **145**(1): p. 47-55.
65. Sevin, A.D., et al., *Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications to AIDS clinical trials group 333.* Journal of Infectious Diseases, 2000. **182**(1): p. 59-67.
66. Foulkes, A. and V.D. Gruttola, *Characterizing the Relationship Between HIV-1 Genotype and Phenotype: Prediction-Based Classification.* Biometrics, 2002. **58**(1): p. 145-156.
67. Foulkes, A. and V. DeGruttola, *Characterizing classes of antiretroviral drugs by genotype.* Statistics in medicine, 2003. **22**(16): p. 2637-2655.
68. DiRienzo, G., *Nonparametric methods to predict HIV drug susceptibility phenotype from genotype.* 2003.
69. DiRienzo, G. and V. DeGruttola, *Collaborative HIV resistance-response database initiatives: sample size for detection of relationships between HIV-1 genotype and HIV-1 RNA response using a non-parametric approach.* Antivir Ther, 2002. **7**: p. S71.
70. Beerenwinkel, N., et al., *Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes.* Nucleic acids research, 2003. **31**(13): p. 3850-3855.
71. Beerenwinkel, N., et al., *Computational methods for the design of effective therapies against drug resistant HIV strains.* Bioinformatics, 2005. **21**(21): p. 3943-3950.
72. Wang, D. and B. Larder, *Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks.* Journal of Infectious Diseases, 2003. **188**(5): p. 653-660.
73. Drăghici, S. and R.B. Potter, *Predicting HIV drug resistance with neural networks.* Bioinformatics, 2003. **19**(1): p. 98-107.
74. Shenderovich, M.D., et al., *Structure-based phenotyping predicts HIV-1 protease inhibitor resistance.* Protein science, 2003. **12**(8): p. 1706-1718.
75. Jenwitheesuk, E. and R. Samudrala, *Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations.* BMC structural biology, 2003. **3**(1): p. 2.
76. Cao, Z.W., et al., *Computer prediction of drug resistance mutations in proteins.* Drug discovery today, 2005. **10**(7): p. 521-529.
77. Jenwitheesuk, E. and R. Samudrala, *Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach.* Antivir. Ther, 2005. **10**(1): p. 157-166.

78. Ravich, V.L., M. Masso, and I.I. Vaisman, *A combined sequence–structure approach for predicting resistance to the non-nucleoside HIV-1 reverse transcriptase inhibitor Nevirapine.* Biophysical chemistry, 2011. **153**(2): p. 168-172.
79. Masso, M. and I.I. Vaisman, *Sequence and structure based models of HIV-1 protease and reverse transcriptase drug resistance.* BMC genomics, 2013. **14**(Suppl 4): p. S3.
80. Maggio, E.T., et al., *Structural pharmacogenomics, drug resistance and the design of anti-infective super-drugs.* Drug discovery today, 2002. **7**(24): p. 1214-1220.
81. Donoho, D.L. and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ1 minimization.* Proceedings of the National Academy of Sciences, 2003. **100**(5): p. 2197-2202.
82. Donoho, D.L., *Compressed sensing.* Information Theory, IEEE Transactions on, 2006. **52**(4): p. 1289-1306.
83. Aharon, M., M. Elad, and A. Bruckstein, *K-SVD: Design of dictionaries for sparse representation.* Proceedings of SPARS, 2005. **5**: p. 9-12.
84. Mairal, J., M. Elad, and G. Sapiro, *Sparse representation for color image restoration.* Image Processing, IEEE Transactions on, 2008. **17**(1): p. 53-69.
85. Elad, M. and M. Aharon, *Image denoising via sparse and redundant representations over learned dictionaries.* Image Processing, IEEE Transactions on, 2006. **15**(12): p. 3736-3745.
86. Lou, Y., A.L. Bertozzi, and S. Soatto, *Direct sparse deblurring.* Journal of Mathematical Imaging and Vision, 2011. **39**(1): p. 1-12.
87. Sprechmann, P. and G. Sapiro. *Dictionary learning and sparse coding for unsupervised clustering*. in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. 2010. IEEE.
88. Wright, J., et al., *Robust face recognition via sparse representation.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009. **31**(2): p. 210-227.
89. Mairal, J., et al. *Discriminative learned dictionaries for local image analysis*. in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. 2008. IEEE.
90. Peyré, G., *Sparse modeling of textures.* Journal of Mathematical Imaging and Vision, 2009. **34**(1): p. 17-31.
91. Cong, Z., W. Xiaogang, and C. Wai-Kuen, *Background subtraction via robust dictionary learning.* EURASIP Journal on Image and Video Processing, 2011. **2011**.
92. Gou, S., G. an Zhuang, and L. Jiao, *Transfer clustering based on dictionary learning for images segmentation.* 2011.
93. Hostetler, L., *The estimation of the gradient of a density function, with applications in pattern recognition.* IEEE Transactions on information theory, 1975. **21**(1): p. 32-40.
94. Cheng, Y., *Mean shift, mode seeking, and clustering.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1995. **17**(8): p. 790-799.
95. Comaniciu, D. and P. Meer. *Mean shift analysis and applications*. in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*. 1999. IEEE.
96. Comaniciu, D. and P. Meer, *Mean shift: A robust approach toward feature space analysis.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002. **24**(5): p. 603-619.
97. Ramanan, D. and D.A. Forsyth. *Finding and tracking people from the bottom up*. in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. 2003. IEEE.
98. Ramanan, D. and D.A. Forsyth. *Using temporal coherence to build models of animals*. in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. 2003. IEEE.
99. Collins, R.T. *Mean-shift blob tracking through scale space*. in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. 2003. IEEE.

100. Bradski, G.R., *Computer vision face tracking for use in a perceptual user interface.* 1998.
101. Sclaroff, S. and L. Liu, *Deformable shape detection and description via model-based region grouping.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2001. **23**(5): p. 475-489.
102. Wang, L.-y., et al., *MSB: a mean-shift-based approach for the analysis of structural variation in the genome.* Genome research, 2009. **19**(1): p. 106-117.
103. Barash, D. and D. Comaniciu. *Meanshift clustering for DNA microarray analysis*. in *Computational Systems Bioinformatics Conference, 2004. CSB 2004. Proceedings. 2004 IEEE*. 2004. IEEE.
104. Liu, X. and H.-G. Müller, *Modes and clustering for time-warped gene expression profile data.* Bioinformatics, 2003. **19**(15): p. 1937-1944.
105. Bose, P., X. Yu, and R.W. Harrison, *Encoding protein structure with functions on graphs.* Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on, 2011: p. 338-344.
106. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.
107. Rangwala, H. and G. Karypis, *frmsdpred: Predicting local rmsd between structural fragments using sequence information.* Proteins: Structure, Function, and Bioinformatics, 2008. **72**(3): p. 1005-1018.
108. Reyaz-Ahmed, A., et al. *Protein Model Assessment via Improved Fuzzy Decision Tree*. in *BIOCOMP*. 2010.
109. Richards, F.M., *The interpretation of protein structures: total volume, group volume distributions and packing density.* Journal of molecular biology, 1974. **82**(1): p. 1-14.
110. Zimmer, R. and R. Thiele, *New scoring schemes for protein fold recognition based on Voronoi contacts.* Bioinformatics, 1998. **14**(3): p. 295-308.
111. Joachims, T., *Making large scale SVM learning practical.* 1999.
112. Wang, G. and R.L. Dunbrack, *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-1591.
113. Kim, C. and B. Lee, *Accuracy of structure-based sequence alignment of automatic methods.* BMC bioinformatics, 2007. **8**(1): p. 355.
114. John, B. and A. Sali, *Comparative protein structure modeling by iterative alignment, model building and model assessment.* Nucleic acids research, 2003. **31**(14): p. 3982-3992.
115. Zemla, A., *LGA: a method for finding 3D similarities in protein structures.* Nucleic acids research, 2003. **31**(13): p. 3370-3374.
116. Poupon, A., *Voronoi and Voronoi-related tessellations in studies of protein structure and interaction.* Current opinion in structural biology, 2004. **14**(2): p. 233-241.
117. Zienkiewicz, O., R. Taylor, and J. Zhu, *The finite element method: its basis and fundamentals.* . 2005, Butterworth-Heinemann, Oxford.
118. Vapnik, V.N., *The nature of statistical learning theory*. 2000: Springer-Verlag New York Inc.
119. Yu, X., R.W. Harrison, and I.T. Weber. *HIV drug resistance prediction using multiple regression*. in *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*. 2013. IEEE.
120. Johnson, V.A., et al., *Update of the drug resistance mutations in HIV-1: March 2013.* Top Antivir Med, 2013. **21**(1): p. 6-14.
121. Sing, T. and N. Beerenwinkel, *Mutagenetic tree Fisher kernel improves prediction of HIV drug resistance from viral genotype.* Advances in Neural Information Processing Systems, 2007. **19**: p. 1297.
122. Deforche, K., et al., *Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance.* Bioinformatics, 2006. **22**(24): p. 2975-2979.

123. Bose, P., X. Yu, and R.W. Harrison. *Encoding protein structure with functions on graphs*. in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*. 2011. IEEE.
124. Yu, X., I. Weber, and R. Harrison, *Sparse Representation for HIV-1 Protease Drug Resistance Prediction*, in *2013 SIAM International Conference on Data mining*. 2013: Austin, TX, USA. p. 342-349.
125. Obermeier, M., et al., *HIV-GRADE: A Publicly Available, Rules-Based Drug Resistance Interpretation Algorithm Integrating Bioinformatic Knowledge.* Intervirology, 2012. **55**(2): p. 102-107.
126. Talbot, A., et al., *Predicting tipranavir and darunavir resistance using genotypic, phenotypic, and virtual phenotypic resistance patterns: an independent cohort analysis of clinical isolates highly resistant to all other protease inhibitors.* Antimicrobial agents and chemotherapy, 2010. **54**(6): p. 2473-2479.
127. HIV/AIDS., J.U.N.P.o., *2008 Report on the global AIDS epidemic*. 2009: World Health Organization.
128. Louis, J.M., et al., *HIV-I protease: Maturation, enzyme specificity, and drug resistance.* Advances in pharmacology, 2000. **49**: p. 111-146.
129. Darke, P.L., et al., *HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins.* Biochemical and biophysical research communications, 1988. **156**(1): p. 297-303.
130. Karacostas, V., et al., *Human immunodeficiency virus-like particles produced by a vaccinia virus expression vector.* Proceedings of the National Academy of Sciences, 1989. **86**(22): p. 8964-8967.
131. Roberts, N.A., et al., *Rational design of peptide-based HIV proteinase inhibitors.* Science, 1990. **248**(4953): p. 358-361.
132. Agniswamy, J., et al., *Terminal Interface Conformations Modulate Dimer Stability Prior to Amino Terminal Autoprocessing of HIV-1 Protease.* Biochemistry, 2012.
133. Hirsch, M.S., et al., *Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel.* Clinical Infectious Diseases, 2008. **47**(2): p. 266-285.
134. Weber, I.T. and J. Agniswamy, *HIV-1 protease: structural perspectives on drug resistance.* Viruses, 2009. **1**(3): p. 1110-1136.
135. Prosperi, M.C.F., et al., *Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment.* Antivir Ther, 2009. **14**: p. 433-442.
136. Van Laethem, K., et al., *A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients.* Antiviral therapy, 2002. **7**(2): p. 123-9.
137. Rhee, S.Y., et al., *Human immunodeficiency virus reverse transcriptase and protease sequence database.* Nucleic acids research, 2003. **31**(1): p. 298-303.
138. Meynard, J.L., et al., *Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial.* Aids, 2002. **16**(5): p. 727-736.
139. Candès, E.J. and M.B. Wakin, *An introduction to compressive sampling.* Signal Processing Magazine, IEEE, 2008. **25**(2): p. 21-30.
140. Aharon, M., M. Elad, and A. Bruckstein, *k-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation.* Signal Processing, IEEE Transactions on, 2006. **54**(11): p. 4311-4322.
141. Duarte, M.F., et al., *Single-pixel imaging via compressive sampling.* Signal Processing Magazine, IEEE, 2008. **25**(2): p. 83-91.

142. Wright, J., et al., *Sparse representation for computer vision and pattern recognition.* Proceedings of the IEEE, 2010. **98**(6): p. 1031-1044.
143. Rhee, S.Y., et al., *Genotypic predictors of human immunodeficiency virus type 1 drug resistance.* Proceedings of the National Academy of Sciences, 2006. **103**(46): p. 17355-17360.
144. Petropoulos, C.J., et al., *A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1.* Antimicrobial Agents and Chemotherapy, 2000. **44**(4): p. 920-928.
145. Bose, P., Yu, X., Harrison, R.W., *Encoding protein structure with functions on graphs*, in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*. 2011, IEEE: Atlanta. p. 338-344.
146. Tie, Y., et al., *Atomic resolution crystal structures of HIV-1 protease and mutants V82A and I84V with saquinavir.* Proteins: Structure, Function, and Bioinformatics, 2007. **67**(1): p. 232-242.
147. Pati, Y.C., R. Rezaiifar, and P. Krishnaprasad. *Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition*. in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. 1993. IEEE.
148. Canu, S., et al., *Svm and kernel methods matlab toolbox.* Perception Systemes et Information, INSA de Rouen, Rouen, France, 2005. **2**: p. 2.
149. Guide, M.U., *The MathWorks Inc.* Natick, MA, 1998. **4**.
150. Hornik, K., M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators.* Neural networks, 1989. **2**(5): p. 359-366.
151. Howard, D. and M. Beale, *Neural Network Toolbox, for Use with MATLAB, User's Guide, Version 4, The MathWorks.* Inc. product, 2000: p. 133-205.
152. Plot, C., *The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences.* Biotechniques, 2000. **28**(6).
153. Yu, X., I. Weber, and R. Harrison, *Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure.* BMC genomics, 2014: p. Minor revision.
154. UNAIDS, J., *Global report: UNAIDS report on the global AIDS epidemic 2010.* UNAIDS Geneva, 2010.
155. Mehellou, Y. and E. De Clercq, *Twenty-six years of anti-HIV drug discovery: where do we stand and where do we go.* J Med Chem, 2010. **53**(2): p. 521-538.
156. Menéndez-Arias, L., *Molecular basis of human immunodeficiency virus type 1 drug resistance: Overview and recent developments.* Antiviral research, 2013(98): p. 93-120.
157. Tie, Y., et al., *High resolution crystal structures of HIV-1 protease with a potent non-peptide inhibitor (UIC-94017) active against multi-drug-resistant clinical strains.* Journal of molecular biology, 2004. **338**(2): p. 341-352.
158. Agniswamy, J., et al., *Extreme multidrug resistant HIV-1 protease with 20 mutations is resistant to novel protease inhibitors with P1′-pyrrolidinone or P2-tris-THF.* Journal of medicinal chemistry, 2013.
159. Shafer, R., et al., *HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance.* Aids, 2007. **21**(2): p. 215.
160. Vandamme, A.-M., et al., *European recommendations for the clinical use of HIV drug resistance testing: 2011 update.* AIDS Rev, 2011. **13**(2): p. 77-108.
161. Vercauteren, J. and A.-M. Vandamme, *Algorithms for the interpretation of HIV-1 genotypic drug resistance information.* Antiviral research, 2006. **71**(2): p. 335-342.
162. Kjaer, J., et al., *Prediction of phenotypic susceptibility to antiretroviral drugs using physiochemical properties of the primary enzymatic structure combined with artificial neural networks.* HIV medicine, 2008. **9**(8): p. 642-652.

163. Wang, D., et al., *A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy.* Artificial intelligence in medicine, 2009. **47**(1): p. 63-74.
164. Gao, Y., et al., *Sparse texture active contour.* IEEE Transactions on Image Processing, 2013.
165. Sprechmann, P. and G. Sapiro, *DICTIONARY LEARNING AND SPARSE CODING FOR UNSUPERVISED CLUSTERING.*
166. Corbau, R., et al., *Lersivirine, a nonnucleoside reverse transcriptase inhibitor with activity against drug-resistant human immunodeficiency virus type 1.* Antimicrobial agents and chemotherapy, 2010. **54**(10): p. 4451-4463.
167. Larder, B., *Quantitative prediction of HIV-1 phenotypic drug resistance from genotypes: the virtual phenotype (VirtualPhenotype).* Antiviral therapy, 2000. **5**: p. 63-63.
168. Puchhammer-Stöckl, E., et al., *Comparison of virtual phenotype and HIV-SEQ program (Stanford) interpretation for predicting drug resistance of HIV strains.* HIV medicine, 2002. **3**(3): p. 200-206.
169. Agniswamy, J., et al., *HIV-1 protease with 20 mutations exhibits extreme resistance to clinical inhibitors through coordinated structural rearrangements.* Biochemistry, 2012. **51**(13): p. 2819-2828.
170. Bakamjian, L., et al., *Global report: UNAIDS report on the global AIDS epidemic 2013.* Journal of the American Board of Family Medicine, 2013. **26**(2): p. 187-95.
171. Gallo, R.C., et al., *Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS.* Science, 1984. **224**(4648): p. 500-503.
172. Lendeckel, U. and N.M. Hooper, *Viral proteases and antiviral protease inhibitor therapy*. 2009: Springer.
173. Kohl, N.E., et al., *Active human immunodeficiency virus protease is required for viral infectivity.* Proceedings of the National Academy of Sciences, 1988. **85**(13): p. 4686-4690.
174. Seelmeier, S., et al., *Human immunodeficiency virus has an aspartic-type protease that can be inhibited by pepstatin A.* Proceedings of the National Academy of Sciences, 1988. **85**(18): p. 6612-6616.
175. Mitsuya, H., et al., *3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro.* Proceedings of the National Academy of Sciences of the United States of America, 1985. **82**(20): p. 7096-7100.
176. Fischl, M.A., et al., *The efficacy of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial.* The New England journal of medicine, 1987. **317**(4): p. 185-191.
177. Ahuja, T.S., M. Borucki, and J. Grady, *Highly active antiretroviral therapy improves survival of HIV-infected hemodialysis patients.* American journal of kidney diseases, 2000. **36**(3): p. 574-580.
178. Levy, J.A. and J. Shimabukuro, *Recovery of AIDS-associated retroviruses from patients with AIDS or AIDS-related conditions and from clinically healthy individuals.* Journal of Infectious Diseases, 1985. **152**(4): p. 734-738.
179. Barré-Sinoussi, F., et al., *Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS).* Science, 1983. **220**(4599): p. 868-871.
180. Chaix-Couturier, C., et al., *HIV-1 drug resistance genotyping: a review of clinical and economic issues.* Pharmacoeconomics, 2000. **18**(5): p. 425-433.
181. Drake, J.W., *Rates of spontaneous mutation among RNA viruses.* Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**(9): p. 4171-4175.
182. Wei, X., et al., *Viral dynamics in human immunodeficiency virus type 1 infection.* Nature, 1995. **373**(6510): p. 117-122.

183. Erickson, J.W. and S.K. Burt, *Structural mechanisms of HIV drug resistance.* Annual review of pharmacology and toxicology, 1996. **36**(1): p. 545-571.
184. Weber, I.T. and R.W. Harrison, *Molecular mechanics analysis of drug-resistant mutants of HIV protease.* Protein engineering, 1999. **12**(6): p. 469-474.
185. Sussman, F., M.C. Villaverde, and A. Davis, *Solvation effects are responsible for the reduced inhibitor affinity of some HIV-1 PR mutants.* Protein science, 1997. **6**(5): p. 1024-1030.
186. Rick, S.W., et al., *Molecular mechanisms of resistance: Free energy calculations of mutation effects on inhibitor binding to HIV-1 protease.* Protein science, 1998. **7**(8): p. 1750-1756.
187. Piana, S., P. Carloni, and U. Rothlisberger, *Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations.* Protein science, 2002. **11**(10): p. 2393-2402.
188. Wang, W. and P.A. Kollman, *Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance.* Proceedings of the National Academy of Sciences, 2001. **98**(26): p. 14937.
189. MenÃ©ndez Arias, L., *Molecular basis of human immunodeficiency virus type 1 drug resistance: overview and recent developments.* Antiviral research, 2013. **98**(1): p. 93-120.
190. Ohtaka, H., A. SchÃ¶n, and E. Freire, *Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations.* Biochemistry, 2003. **42**(46): p. 13659-13666.
191. Henderson, G., et al., *Interplay between single resistance-associated mutations in the HIV-1 protease and viral infectivity, protease activity, and inhibitor sensitivity.* Antimicrobial agents and chemotherapy, 2012. **56**(2): p. 623-633.
192. Louis, J., et al., *Inhibition of autoprocessing of natural variants and multidrug resistant mutant precursors of HIV-1 protease by clinical inhibitors.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(22): p. 9072-9077.

11 APPENDICES

Appendix A: List of Publications

- Xiaxia Yu, Irene T. Weber, and Robert W. Harrison. Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. BMC genomics, 15 (Suppl 5), S1, 2014.
- Xiaxia Yu, Irene T. Weber, and Robert W. Harrison. Sparse representation for hiv-1 protease drug resistance prediction. in Data Mining (SDM), 2013 SIAM International Conference on. SIAM, 2013, pp. 342-349.
- Xiaxia Yu, Irene T. Weber, and Robert W. Harrison. HIV drug resistance prediction using multiple regression, an application of a new sequence/structure hybrid protein encoding method. in Computational Advances in Bio and Medical Sciences (ICCABS), 3rd IEEE International Conference on. IEEE, 2013, pp. 1-2.
- Robert W. Harrison, Xiaxia Yu, and Irene T. Weber. Using triangulation to include target structure improves drug resistance prediction accuracy. in Computational Advances in Bio and Medical Sciences (ICCABS), 3rd IEEE International Conference on. IEEE, 2013, pp. 1.
- Chen-Hsiang Shen, Yunfeng Tie, Xiaxia Yu, Yuan-Fang Wang, Andrey Y Kovalevsky, Robert W Harrison, and Irene T Weber. Capturing the reaction pathway in near-atomic-resolution crystal structures of hiv-1 protease. Biochemistry, vol. 51, no. 39, pp. 7726-7732, 2012.
- Yu-Chung E Chang, Xiaxia Yu, Ying Zhang, Yunfeng Tie, Yuan-Fang Wang, Sofiya Yashchuk, Arun K Ghosh, Robert W Harrison, and Irene T Weber. Potent antiviral hiv-1 protease inhibitor grl-02031 adapts to the structures of drug resistant mutants with its p1-pyrrolidinone ring. Journal of medicinal chemistry, vol. 55, no. 7, pp. 3387-3397, 2012.
- Promita Bose, Xiaxia Yu, and Robert W Harrison. Encoding protein structure with functions on graphs. in Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on. IEEE, 2011, pp. 338-344.
- Ying He, Jinzhen Shao, Xiaxia Yu, Qianhe Wang, Bohua Zhu, and Yi Ding. Study on storage protein of high quality hybrid rice in hubei province. Journal of WUHAN Botanical Research, vol. 25, no. 3, pp.270-276, 2007. (In Chinese)
- Kan Hu, Xiaxia Yu, Jinhong Yu, and Yi Ding. Construction of an analysis between three lines of cytoplasmic non-pollen type in yunnan purple rice and their hybrid f1. Chinese Journal of Rice Science, vol. 21, no. 4, pp. 345-349, 2007. (In Chinese)