RHODES UNIVERSITY

Structural analysis of proteases from South African HIV-1 (subtype C) patients undergoing Lopinavir treatment, using comparative modeling, ligand-docking and molecular dynamics

A Thesis

by

Olivier SHEIK AMAMUDDY

Department of Biochemistry and Microbiology

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science,

Bioinformatics and Computational Molecular Biology

December 2015

ABSTRACT

HIV is regarded as one of the most devastating infectious diseases of the last few decades, and has a high prevalence in South Africa, subtype C being the most common. Palliative measures used to fight HIV involve the use various types of inhibitors, including the use of HIV protease inhibitors. Representatives from this class of inhibitors are gradually losing their efficacy due to development of resistance mutations from HIV-1. In this study, compounds from the South African Natural Compound Database (SANCDB) were screened against HIV-1 protease models generated from protease protein sequences belonging to 11 South African HIV patients before and after treatment with Lopinavir. The effect of Lopinavir on the alteration of drug-binding affinity before and after treatment is investigated by molecular docking of the protease against other FDA-approved drugs and detection of mutation types using the HIVdb tool. A network representation of hydrogen bonding between docked ligands and their receptor proteases has been developed and a profiling method of visualizing receptor-ligand docking energies at the local level is presented.

Four potential HIV-1 protease inhibitors were identified from the list of 599 natural compounds on the basis of receptor conformation and binding free energy. Ligand stabilities were monitored by 20ns molecular dynamics runs using the GROMACS software.

DECLARATION

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from 15 July 2015 to 15 December 2015 under the supervisions of Dr Kevin Lobb and Prof Özlem Taştan Bishop.

I, Olivier Sheik Amamuddy, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature .....

Date .....

ACKNOWLEDGMENTS

I am very thankful to my supervisors Dr Kevin Lobb and Professor Özlem Tastan Bishop for the motivation, support and guidance they gave me all throughout the course of the research project. I am also very thankful to Professor Jaufeerally-Fakim for giving me the chance to pursue this course and H3ABioNet for fully funding it.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF USEFUL WEB LINKS

ANOLEA	http://melolab.org/anolea/
CATH	http://www.cathdb.info/
HHpred	http://toolkit.tuebingen.mpg.de/hhpred
MODBASE	http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
MODELLER	https://salilab.org/modeller/
NCBI	http://ncbi.nlm.nih.gov/
PDB	http://pdb.org/
PDB-BLAST, BLASTP	http://blast.ncbi.nlm.nih.gov/
SANCDB	https://sancdb.rubi.ru.ac.za/
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
Stanford HIVdb	http://hivdb.stanford.edu/
Stanford HIVdb web tool	http://sierra2.stanford.edu/sierra/servlet/JSierra?action=sequenceInput
SWISS-MODEL	http://swissmodel.expasy.org/

CHAPTER 1: Background on HIV-1 biology and current therapies

This chapter gives an overview of the economic importance, classification and general molecular biology of the Human Immunodeficiency Virus type 1 (HIV-1) and elaborates on some of the successful drugs that help control the spread of the virus in the human body. The resilient nature of the pathogen is also described.

1.1. HIV-1 biology and importance

HIV-1 is one of the two AIDS-causing lentiviruses in humans (HIV-1 and HIV-2) (Sharp, P. M. and Hahn 2011). They are a result of cross-species transmissions of simian immunodeficiency viruses that naturally affect African primates (Sharp, P. M. and Hahn 2011). Only to be formally recognized as a disease in 1981 in the United States, AIDS is now deemed one of the most devastating infectious diseases of the last few decades (Sharp & Hahn 2010), causing 2.6 million new infections yearly (Messiaen et al. 2013). According to the UNAIDS, South Africa has the highest prevalence of HIV/AIDS sufferers compared to any country worldwide, with an estimated 6.8 million people living with HIV, the majority consisting of people aged 15 and above (AIDS Foundation of South Africa 2014; UNAIDS 2014). The reasons behind the high prevalence are mainly of socio-economic nature, including poverty, uneven access to medical aid, social instability and high mobility of migrant labor, compounded with illiteracy, despite the high levels of knowledge about the transmission and prevention of the disease (AIDS Foundation of South Africa 2014). One big challenge is that there can be a gap of 8 to 10 years between the initial HIV infection and the development of AIDS, which makes prediction of the course of an epidemic difficult (Botes et al. 2007).

In addition to the long incubation period, the HIV-1 retrovirus has an extensive genetic diversity due to its continued mutations and recombinations (Lihana et al. 2012). Simple classification into M (pandemic), N (new group, Not-M or Not-O) and O (outlier) groups (Lihana et al. 2012) is insufficient to encompass the complexity of the HIV-1 and as such, further subtyping into subtypes (A-D, F-H and J-K) (Lihana et al. 2012) have been employed. Additional classifications of the subtypes have also been defined, such as A1-A5 (Lihana et al. 2012). Yet another level of complexity is the occurrence of inter-subtype recombinant genomes, which can occur in dually-infected patients (Lihana et al. 2012). Owing to the fact that the HIV genetic material consists of two copies of single-stranded RNA (ssRNA), if a single host cell is infected by two (or more)

strains of the virus, chances are that the virions packaged inside the cell may contain an RNA copy from each strain, giving rise to recombinant virus particles (Los Alamos National Laboratory 2012). These strains may be termed as unique recombinant forms (URFs) if they are found only in one individual or as circulating recombinant forms (CRFs) if they are transmitted to other people – the CRFs are labeled with numbers in the order they were described (Los Alamos National Laboratory 2012).

Southern African countries have the highest number of people infected with the HIV-1 of subtype C (Hemelaar 2012). Unfortunately, protease inhibitors (PIs) have mainly been developed for the B subtype, while non-B subtype polymorphisms are less well-documented in terms of drug efficacy despite their lower drug binding affinities (Velázquez-Campoy et al. 2003; Wensing et al. 2010).

1.2. Structural biology of HIV

The mature icosahedral-shaped virion is composed of a glycoprotein-decorated envelope (consisting of gp120 surface proteins, also referred to as SU proteins) anchored to a lipid bilayer by the transmembrane gp41 proteins (also termed as TM proteins), with the bilayer itself being lined by an inner matrix (Turner & Summers 1999; Botes et al. 2007). It should be noted that the lipid bilayer is also composed of host cell proteins, such as MHC (Major Histocompatibility Complex) antigens, actin and ubiquitin (Turner & Summers 1999).

The genetic material is stored as a pair of single-stranded, unspliced RNA molecules, stabilized as a ribonucleoprotein complex by about 2000 nucleocapsid (NC) proteins (Turner & Summers 1999). Also present inside and outside of the capsid are the enzymes integrase, reverse transcriptase and protease (Turner & Summers 1999). In addition to the former proteins, the viral particles also comprise four accessory proteins, namely Nef, Vpr, Vpu and Vif (Botes et al. 2007) and two regulatory proteins, namely Rev and Tat (Adamson & Freed 2010). Accessory proteins are only termed as such because they are dispensable for viral replication in some conditions of culture (Adamson & Freed 2010). The viral proteins are summarized in the Table 1.1:

Table 1.1: Overview of main HIV proteins

Structural proteins	MA (Matrix proteins)
	CA (Capsid proteins)
	Envelope glycoproteins: gp120 (surface) and gp41 (transmembrane)
	NC (nucleocapsid proteins): p24, p17, p9 and p7
Viral enzymes	RT (reverse transcriptase)
	IN (integrase)
	PR (HIV protease)
Accessory proteins	Vpu (viral protein u)
	Nef (negative regulator factor)
	Vif (viral infectivity factor)
	Vpr (viral protein r)
Regulatory proteins	Rev (regulator of virion)
	Tat (trans-activator of transcription)

Adapted from Botes et al. 2007 and Adamson & Freed 2010

1.3. HIV genomic organization

A typical representation of the HIV-1genome (Figure 1.1) is based on the reference HIV-1 strain HXB2CG, showing the genes observable across three frames along its 9719 bases long genome. The RNA is flanked by Long Terminal Repeats (LTRs), which contain regulatory regions for the initiation of transcription at the 5' region and polyadenylation-enabling region at the 3' end (Krebs et al. 2001). The Gag polyprotein, the Nef and Vif accessory proteins are coded from the first frame while Tat and Vpu are coded from the second frame; *pol, vpr, rev* and the *env* sequences are encoded by the third frame of the viral genome (Los Alamos National Laboratory 2014). Other genomic features not shown in Figure 1.1 are summarized in Table 1.2:

Table1. 2: Summary of some HIV-1 structural elements (Los Alamos National Laboratory 2014)

TAR	45 nucleotide-long binding site for the Tat and cellular proteins. Essential for Tat function.
RRE	Approximately 200 nucleotide-long region binding site required for Rev functioning.
PE	Region recognized by the cysteine histidine box motif of the NC protein.
SLIP	TTTTTT slippery region regulating the frame-shift between Gag and Gag-Pol polyproteins.
CRS	Proposed to inhibit expression of structural proteins in the absence of Rev. Exact function is not known.
INS	Multiple independent copies may present in a genome. Inhibit gene expression at the post-translational level.

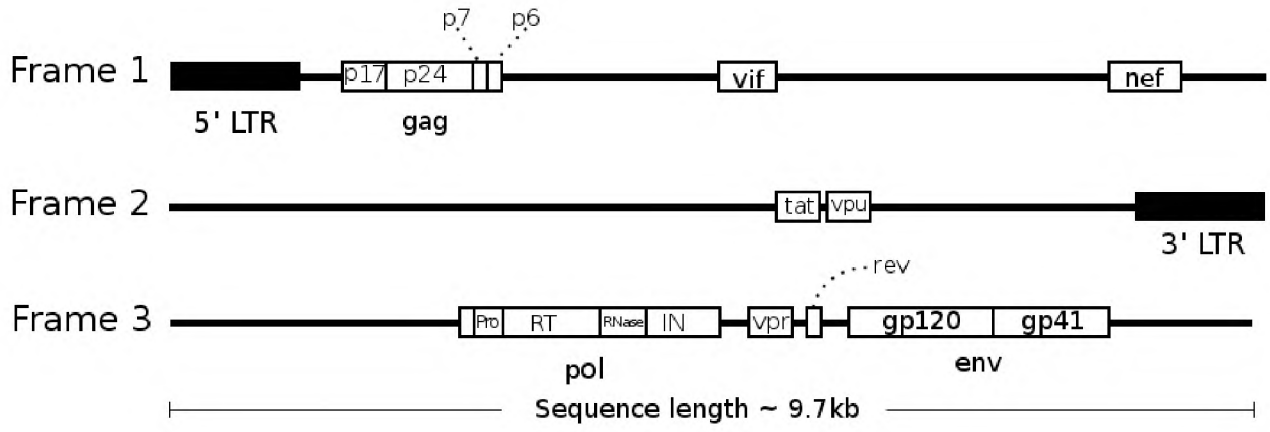


Figure 1.1: Genome organization of HIV-1, showing the open reading frames from the reference HIV-1 strain HXB2CG.

(Adapted from: http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html)

1.4. The replication mechanism of HIV-1

HIV-1 replication is a multi-step process that begins with entry of the viral particles inside target cells (Adamson & Freed 2010). Fusion of the host cell membrane and the viral lipid envelope precedes entry of the viral particle (Doms 2000), which is accomplished by the binding of the viral glycoproteins (gp41 and gp120) to the host cell surface CD4 receptors and additional interactions with host CD4 co-receptors (CCR5 or CXCR4) (Doms 2000; Berger et al. 1999) . Upon interaction of the glycoproteins with the co-receptors, a series of conformational changes leads to membrane fusion (Doms 2000; Melikyan 2008). The fusion delivers the viral core (composed of the CA protein shell together with the NC, PR and IN proteins and RNA) (Adamson & Freed 2007; Ganser-Pornillos et al. 2008). The core is subsequently uncoated (Warrilow & Harrich 2007) and the viral genome is reverse transcribed by RT to double stranded DNA (Sarafianos et al. 2009), which is transported to the host cell nucleus for stable integration into the host cell genome (Vandegraaff & Engelman 2007; Suzuki & Craigie 2007; Delelis et al. 2008). Integrated DNA can then be transcribed back to the viral RNA, spliced and translated in the host cell cytoplasm (Swanstrom & Wills 1997; Rabson & Graves 1997). The ribosomal frameshift together with the series of proteolytic cleavages (Wensing et al. 2010; Jacks et al. 1988) result in different protein combinations, some of which mentioned in Table 1.1. These proteins and the genomic RNA are then transported to the virus assembly site at the host cell plasma membrane – the assembly itself is orchestrated by the Gag protein, which also fits several host proteins inside the viral packaged particle (Adamson & Freed 2007). Budding of the viral particle off the host cell membrane completes the viral production process (Adamson & Freed 2007). The knowledge of these steps in the HIV-1 replication cycle has enabled the development of multiple antiretroviral (ARV) drugs (Adamson & Freed 2010).

1.5. Controlling HIV-1 using antiretrovirals

Six classes of inhibitors are currently available for treating HIV/AIDS patients (AIDS.gov 2015), the mechanisms of which mainly consist in inhibiting various proteins and/or processes important for HIV replication. These include inhibitors of non-nucleoside reverse transcriptase (NNRTI's), protease inhibitors, integrase inhibitors, fusion inhibitors, entry inhibitors (chemokine co-receptor antagonist) and multi-class combination products (FDA 2015). In order to prevent resistance from any one type of HIV-1 inhibitor, it is recommended that combinations of the drugs be taken as a therapy, which is termed highly active antiretroviral therapy (HAART) (NIAID 2013).

1.6. HIV-1: Escaping PI treatment

In the wait of an effective vaccine, ARV combination therapy has been the main form of treatment against HIV (Ragland et al. 2014). However, despite efforts to develop new drugs and exploit different targets, the virus manages to ward off their therapeutic effect (Ragland et al. 2014). In fact, as at year 2014 resistance has emerged against all 30 ARV's available at that time, as per Ragland et al. 2014 . PIs are currently observed as the most potent ARV's in HIV treatment (Ragland et al. 2014). Unfortunately, the same problems apply as for all the drugs used to treat HIV – as much as 45 residues out of the 99-residue long protease (almost 50% of the enzyme) can mutate and contribute to HIV drug resistance without losing enzymatic activity (Ragland et al. 2014; Wu et al. 2003). Recent work has shown that human enzymes of the apolipoprotein B (A3) family were capable of increasing the HIV-1 mutation rate by viral mRNA editing (Cuevas et al. 2015), thus helping it to evolve and develop faster resistance against antiretrovirals in general. The HIV-1 protein Vif is known to act in tandem with the apolipoprotein by promoting A3-driven degradation (Desimmie et al. 2014).

1.7. Role of HIV-1 protease in the HIV life cycle

In the context of the HIV life cycle, the protease is required for the cleavage of the Gag and Gag-Pol polyprotein precursors for the maturation phase of HIV life cycle (Özen et al. 2011). The protease also cleaves itself (in a process termed autoprocessing) at the N-terminus of the Gag-Pol polyprotein (Louis et al. 2011). Maturation of the virion entails the release of functional proteins in the form of enzymes and structural proteins (Navia et al. 1989; Debouck et al. 1987), which are crucial for the final morphological rearrangement and production of infectious viral particles (Sundquist & Kräusslich 2012; Pokorná et al. 2009). Inhibition of the HIV protease would therefore prevent spreading of the viral particles to other cells (Pokorná et al. 2009) and is in fact one of many approaches used to fight HIV.

The development and successful use of protease inhibitors (PIs) is deemed one of the most remarkable achievements of molecular medicine (Pokorná et al. 2009). In fact, the introduction of PIs (combined with other anti-retrovirals) lead to a significant drop in the number of deaths and an increased life expectancy in the mid 1990's (Pokorná et al. 2009). It is a mixed success however, in that their use was accompanied by problems of antiviral resistance, due to HIV's error-prone reverse transcriptase, which gives rise to high mutation rates (Pokorná et al. 2009). Other problems have followed, namely those of tolerability and toxicity in HIV-positive patients, in addition to the high

price and lack of patient adherence to the treatment (Pokorná et al. 2009). Also, the inhibitors have unpredicted side-effects, interacting with other molecules in lipid metabolism and lipid trafficking pathways (Pokorná et al. 2009) resulting in the treatment being an even greater health risk than the HIV infection itself in certain cases (Wohl et al. 2006; Nolan et al. 2005; Shibuyama et al. 2006).

1.8. Structural features and molecular mechanism of HIV1-protease

HIV-1 protease forms part of the aspartic protease family where all of its members require two aspartic acid residues to catalyze the breakdown of peptide bonds via a water-based nucleophilic attack (Klebe 2013b). The protease has several preferred targets for peptide linkage cleavage, namely those between the following residue pairs: phenylalanine-proline, tyrosine-proline, phenylalanine-tyrosine, leucine-phenylalanine, phenylalanine-leucine, methionine-methionine and leucine-alanine (Klebe 2013b). In the case of HIV-1, the protease exists as a symmetric homodimer, composed of two identical polypeptide chains of 99 amino acids each (Klebe 2013b; Navia et al. 1989; Wlodawer et al. 1989). Each of its chains contributes an aspartic acid (residue 25), to the catalytic site of the protein (Klebe 2013b; Navia et al. 1989; Wlodawer et al. 1989). A typical structure of the protease is depicted in Figure 1.2, showing the highly-flexible β-hairpins which form the flaps that control access to the active site of the enzyme (Huang et al. 2014). The protease has been classified into 3 different conformations, namely the open, semi-open and the closed conformation (Mcgee 2010; Huang et al. 2014). These conformational changes are important for the catalytic activity of the enzyme (Huang et al. 2014). The flaps have to be in the opened state to enable substrate entry and product release (Hornak et al. 2006; Freedberg et al. 2002). After substrate entry, cleavage occurs in a general acid/base mechanism (Mcgee 2010). It is presumed that the flaps of the dimeric protein are held in a closed conformation by hydrogen bonding between a water molecule and the amide group of isoleucine 50 (Czodrowski et al. 2007). A study carried out by Torbeev and co-workers (2011), support the same catalytic mechanism for HIV-1 protease as for that of general aspartyl proteases, that is, one aspartate side chain carboxyl (COO^-) group acts as a general base, deprotonating a single water molecule, while the other aspartic acid side chain (COOH) acts as a general acid, donating a proton to the carbonyl oxygen atom of the scissile peptide bond (Suguna et al. 1987; Davies 1990; Torbeev et al. 2011).

Other features of HIV-1 proteases include the threonine 26, and glycine 27 that together with aspartic acid 25, form the catalytic triad (Meher & Wang 2015). The highly-mobile flaps of the protease (residues 33-62) (Scott & Schiffer 2000; Torbeev et al. 2011) are glycine rich and are

composed of anti-parallel beta sheets that cover the active site (Cai et al. 2012). The dimer is held in place by a network of hydrogen bonds at their interface found at the N and C termini, at the active site and at the flap tips (Cai et al. 2012). These flap tips are each arranged in a beta turn conformation that each contain a conserved glycine at position 51 (Torbeev et al. 2011). The active site is further delineated by additional landmarks termed subsites – labeled S1-S4 (on one chain) and S1'-S4' (on the mirror chain) which are residues that interact with the substrate moieties (otherwise proteins side chains) labeled P1-P4 and P1'-P4' for the matching positions (Graham 2013).

1.9. FDA-approved inhibitors of HIV-1 protease

Out of the 26 anti-HIV compounds that have been approved by the US Food and Drug Administration (FDA), ten of them are protease inhibitors, as mentioned by Lv and co-workers (2015). The FDA-approved protease inhibitors include saquinavir (SQV), indinavir (IDV), nelfinavir (NFV), ritonavir (RTV), amprenavir (APV), fosamprenavir (FPV), lopinavir (LPV), atazanavir (ATZ), tipranavir (TPR) and darunavir (DRV) (Lv et al. 2015). All of them are peptidomimetic except for TPR (Ali et al. 2010).These compounds share structural similarities and binding patterns (Lv et al. 2015) and are described in the following subsections. Two dimensional renderings (using Discovery Studio) of the protease inhibitors are shown in Figure 1.3, where oxygen and nitrogen atoms have been labeled according to the nodes present in the hydrogen bonding network graphs (as shown later in Chapter 4, Figures 4.7-4.16) and in and profile graphs for DRV (Chapter 4, Figures 4.17-4.18). It should be noted that the numbering of these atoms is based on the models used in the study, and is not according to systematic nomenclature.

1.9.1. Saquinavir

SQV (brand name: Invirase) was the first HIV protease inhibitor to be approved by the FDA (Colvin & Haas 1995). The rationale behind its use lies in the fact that the viral protease cleaves between phenylalanine (at the P1 site) and proline (at the P1' site) while mammalian proteases do not cleave peptide bonds that comprise proline on at the P1' site (Lv et al. 2015). The final structure of SQV replaced proline by a z(S,S,S)-decahydroisoquinoline-3-carbonyl (DIQ) group to enhance the potency of inhibition of the drug (Krohn et al. 1991). The carbonyl part of the DIQ group interferes with the water molecule (Krohn et al. 1991) that is required for catalysis. Saquinavir has a mean 50% effective concentration (EC_{50}) of 37.7nM (US Food and Drug Administration 2010b), however, the main problem with the drug is that of its low bioavailability (Cameron et al. 1999).

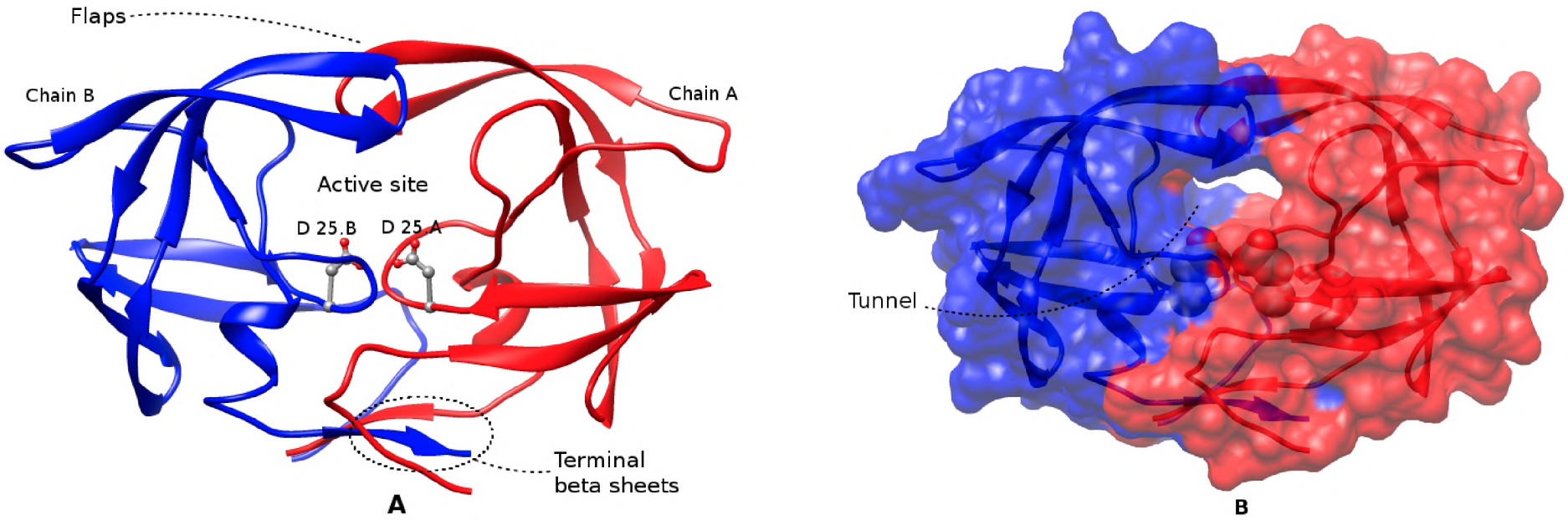
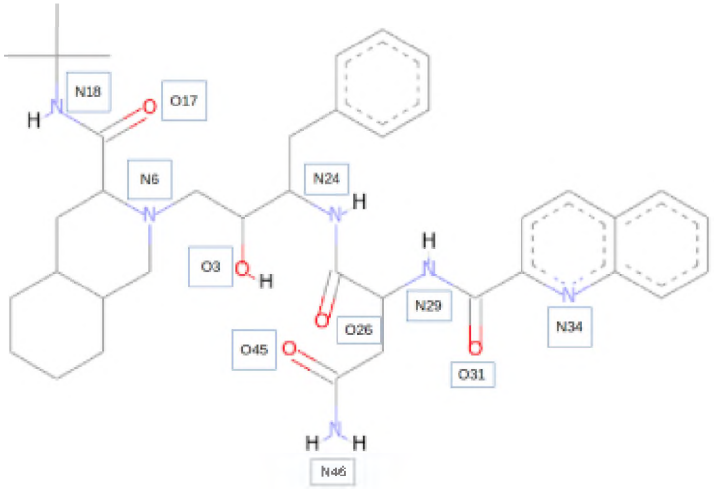


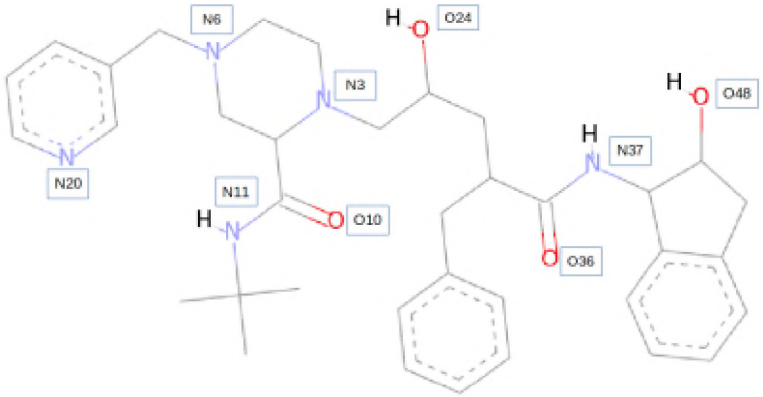
Figure 1.2: Structure of an HIV-1 protease (PDB accession 1HXB), (A) Showing the active site, flap regions and the terminal β sheets formed by chains A and B of the dimer. Catalytic aspartic acids (D 25.A and D 25.B) are shown as ball-and-stick representations. (B) Surface of the protease, showing the tunnel formed by the two chains. Aspartic acid residues 25 (from both chains) are represented as spheres.

(Adapted from Louis et al. 2011, Goodsell 2000 and Krohn et al. 1991).

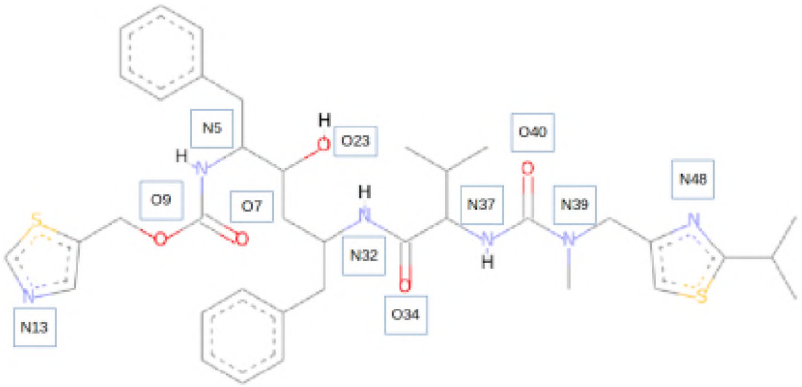
SQV



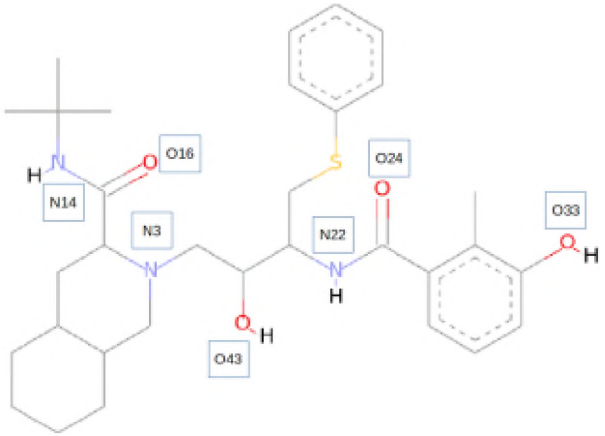
IDV



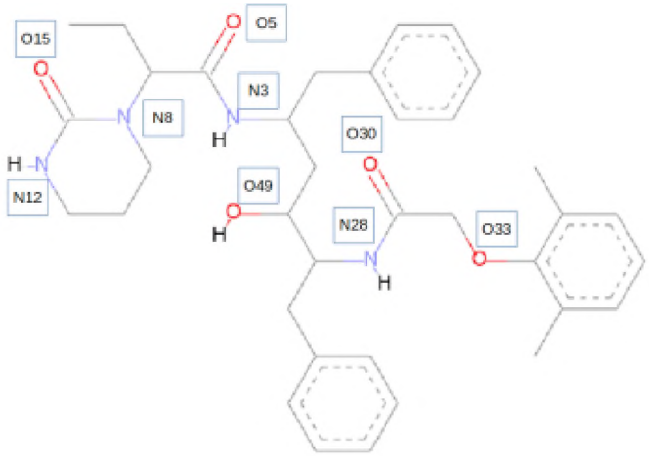
RTV



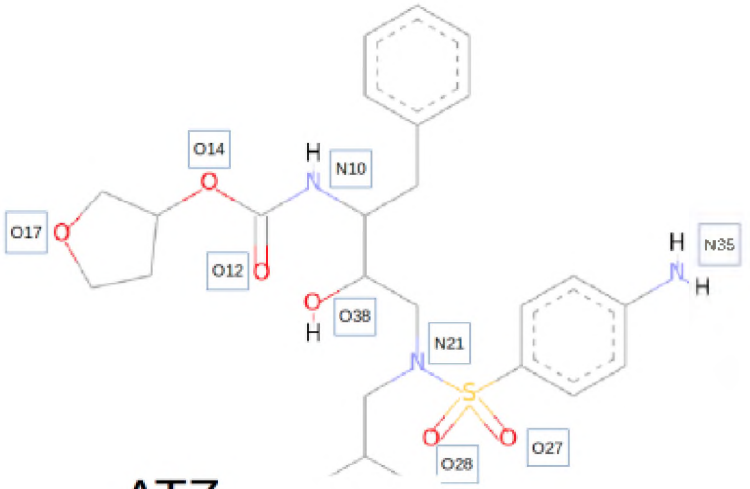
NFV



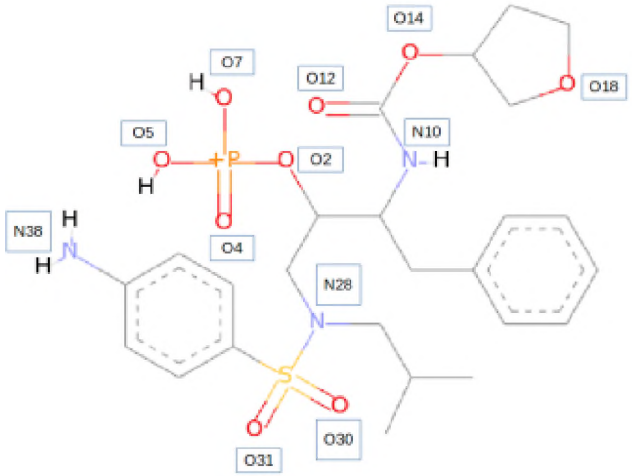
LPV



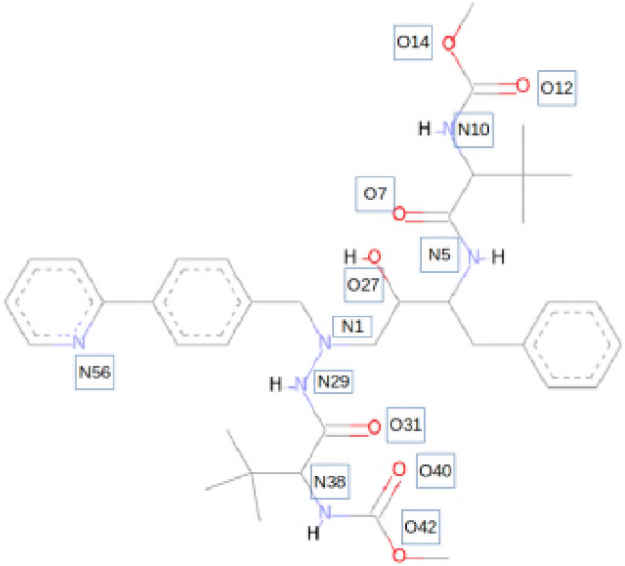
APV



FSV



ATZ



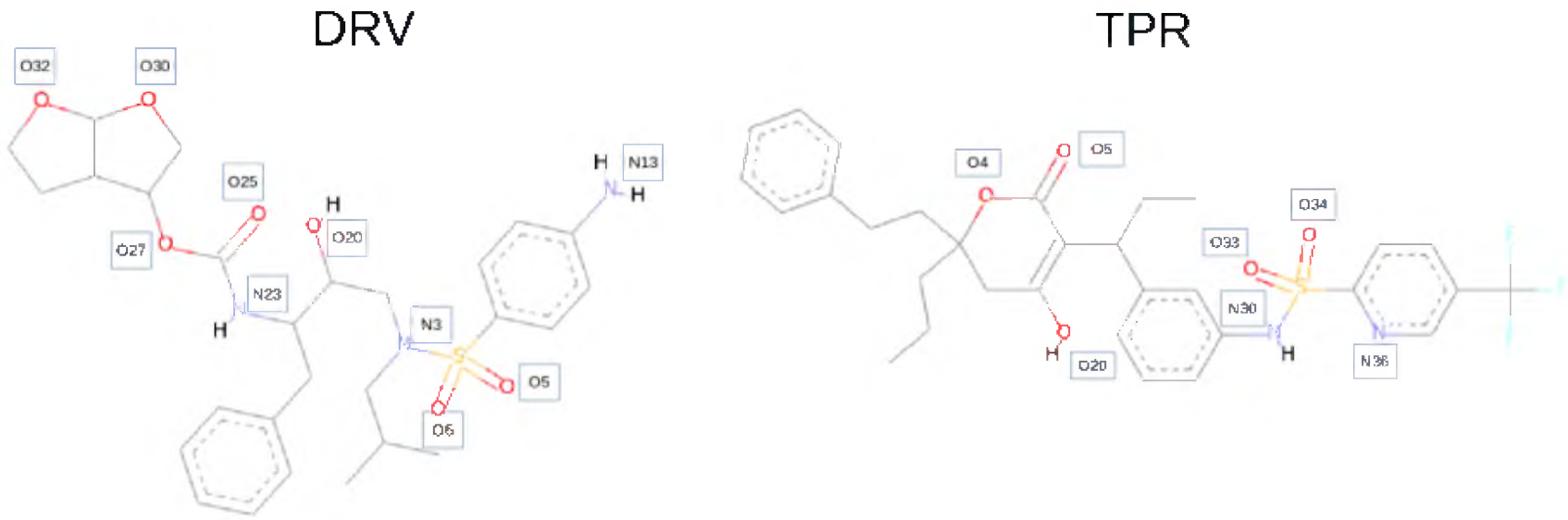


Figure 1.3: Current FDA-approved HIV-1 protease inhibitors (PDB accessions: 478, AB1, 1UN, DR7, RIT, TPV, MK1, 017, ROC; PubChem accession: 131536), with oxygen and nitrogen atoms labeled according to the labels used in Chapter 4 Figures 4.7-4.18

1.9.2. Indinavir

IDV (brand name: Crixivan) has an *in vitro* IC_{95} ranging between 25 and 100nM against HIV-1 (US Food and Drug Administration 2014). However, many problems are related to its use: plasma concentration decreases rapidly, which leads to treatment failures (González de Requena et al. 2003); low solubility, which can result in the development of kidney stones; potential competitive inhibition with the cytoplasmic glucose binding site of GLUT4 (Hresko & Hruz 2011); lipodystrophy syndrome association; high frequency of drug taking. In fact IDV has been replaced by second generation protease inhibitors (Lv et al. 2015).

1.9.3. Ritonavir

RTV (brand name: Norvir) has a mean EC_{50} of about 22nM and is active against both HIV-1 and HIV-2 (US Food and Drug Administration 2011a). It has an isopropyl thiazolyl P3 group that is longer than that of other FDA-approved PI's (Lv et al. 2015). Ritonavir boosting of other drugs normally require less frequent dosing of the combined drug regimen (Kempf et al. 1997). It should be noted that higher doses of RTV may lead to hyperlipidemia in both people suffering from AIDS and in healthy patients (Purnell et al. 2000).

1.9.4. Nelfinavir

NFV (brand name: Viracept) has an EC_{95} in the range of 7 to 196nM (US Food and Drug Administration 2011b). It shares the same DIQ group as saquinavir on one terminus and has a 2-methyl-3-hydroxybenzamide group at the other terminus (Prashar et al. 2015). An S-phenyl group located at the P1 site intensifies the effect of the inhibitor, however the inhibitor has to be taken twice daily and is accompanied with side-effects such as diarrhea and nausea (Lv et al. 2015; Max & Sherer 2000).

1.9.5. Lopinavir

LPV (brand name: Kaletra) was developed as an RTV-based drug, being used in conjunction with ritonavir (US Food and Drug Administration 2010a). It's main constituent consists of a hydroxyethylene dipeptide isostere, similar to that of ritonavir (Lv et al. 2015). It's P2 and P2' groups comprise a phenoxyacetal group and a cyclic urea group respectively (Lv et al. 2015). The small size of the two groups decrease contacts with highly variable residue at position 82 of the HIV-1 protease (Sham et al. 1998).

1.9.6. Amprenavir

APV (brand name: Agenerase) has an IC_{50} in the range of 0.012 to 0.08μM (US Food and Drug Administration 2005). It consists of a benzyl group and an isobutyl group at the P1 and P1' sites respectively (Lv et al. 2015). The presence of fewer chiral centers (compared to other PI's) increases its oral availability (St Clair et al. 1996). The only main side effect consists of benign rash (US Food and Drug Administration 2005).

1.9.7. Fosamprenavir

FPV (brand name: Lexiva) is built as a phosphate ester precursor of amprenavir (Yu et al. 1999). It is metabolism by the body to releases amprenavir, which is the active form of the drug (US Food and Drug Administration 2009b). The slow release means that the drug dosage is reduced (Eron et al. 2006) . In fact, clinical studies have showed that FPV is actually safer than amprenavir (Gathe et al. 2006; Judd et al. 2014).

1.9.8. Atazanavir

ATZ (brand name: Reyataz) is an aza-dipeptide analog, with an EC_{50} in the range of 2.6 to 5.3nM in cell culture (Lv et al. 2015). It has a large phenylpyridyl P1 group and a benzyl P1' group. The drug has good bioavailability (Becker 2003) and has no effect on insulin sensitivity and blood lipid concentrations but does correlate with a higher incidence of proximal tubulopathy in some patients (Dauchy et al. 2011; Calza et al. 2013).

1.9.9. Tipranavir

TPR (brand name: Aptivus) has a lactone oxygen that interacts directly with isoleucine 50 located at the flap region of the protease (Wang 2012). Its EC_{50} ranges between 30 and 70nM, however, it has an increased severity of side effects, which include intra-cranial hemorrhage and a form of hepatitis amongst others (US Food and Drug Administration 2009a).

1.9.10. Darunavir

DRV (brand name: Prezista) is the latest PI on the market as at April 2015 (Lv et al. 2015). It is structurally related to amprenavir with the exception of the P2 group, which allows for a higher number of hydrogen bonds, especially with the protease backbone atoms (Lefebvre & Schiffer 2008; King et al. 2004). The median EC_{50} of this ARV ranges between 1.2 and 2.8nM, decreasing by a factor of 5.4 in the presence of human blood (US Food and Drug Administration 2008).

1.10. Protease mutations associated with resistance to protease inhibitors

Resistance mutations are a major complication during antiretroviral therapy (ART) and protease inhibitors are the class of inhibitors that elicit the highest amount of mutations in HIV-1 (Flor-Parra et al. 2011). Table 1.3, displays the different positions of recorded resistance mutations (up to year 2014) for different protease inhibitors. These resistance mutations can be divided into major and minor mutations (Wensing et al. 2014). Major mutations are those that tend to be selected first in the presence of a drug - they tend to affect primary contact residues that are involved in drug binding, and thus have the potential to considerably reduce drug efficacy (Wensing et al. 2014). On the other hand, minor mutations arise later and do not impact substantially the viral phenotype, however, these may improve replication of viruses that already have major mutations against given drugs. (Wensing et al. 2014).

Despite being a protease inhibitor, RTV is not listed on it own in Table 1.3. The reason being that ritonavir is currently used at sub-therapeutic doses (100-200mg) as a pharmacological booster in combination with other protease inhibitors (Wensing et al. 2014; Rock et al. 2014). RTV is in fact a well-known inactivator of cytochrome P450 (CYP3A) (Koudriakova et al. 1998) – a major human drug-metabolizing enzyme that would otherwise eliminate PI's by oxidizing them by a mechanism which is not well understood (Rock et al. 2014).

The inactivation of liver and intestinal CYP3A by RTV increases its own bioavailability and half-life - a property that is utilized for the co-administration of other protease inhibitors in standard HIV therapy to increase the plasma concentration of the co-administered HIV drug (Koudriakova et al. 1998; Rock et al. 2014).

Table 1.3: Mutations associated to resistance against HIV-1 protease inhibitors

Position	Atazanavir +/- ritonavir	Darunavir /ritonavir	Fosamprenavir /ritonavir	Indinavir* /ritonavir	Lopinavir /ritonavir	Nelfinavir /ritonavir	Saquinavir/ ritonavir	Tipranavir** /ritonavir
10	L → (I,F,V,C)		L → (F,I,R,V)	L → (I,R,V)	L → (F,I,R,V)	L → (F,I)	L → (I,R,V)	L → V
11		V → I						
16	G → E							
20	K → (R,M,I,T,V)			K → (M,R)	K → (M,R)			
24	L → I			L → I	L → I		L → I	

Position	Atazanavir +/- ritonavir	Darunavir /ritonavir	Fosamprenavir /ritonavir	Indinavir* /ritonavir	Lopinavir /ritonavir	Nelfinavir /ritonavir	Saquinavir/ ritonavir	Tipranavir** /ritonavir
30						D → N		
32	V → I	V → I	V → I	V → I	V → I			
33	L → (I,F,V)	L → F			L → F			L → F
34	E → Q							
36	M → (I,L,V)			M → I		M → I		M → (I,L,V)
43								K → T
46	M → (I,L)		M → (I,L)	M → (I,L)	M → (I,L)	M → (I,L)		M → L
47		I → V	I → V		I → (V,A)			I → V
48	G → V						G → V	
50	I → L	I → V	I → V		I → V			
53	F → (L,Y)				F → L			
54	I → (L,V,M,T,A)	I → (M,L)	I → (L,V,M)	I → V	I → (L,V,M,T,A,S)		I → (L,V)	I → (V,M,A)
58								Q → E
60	D → E							
62	I → V						I → V	
63					L → P			
64	I → (L,V,M)							
69								H → (K,R)
71	A → (V,I,T,L)			A → (V,T)	A → (V,T)	A → (V,T)	A → (V,T)	
73	G → (C,S,T,A)		G → S	G → (S,A)	G → S		G → S	

Position	Atazanavir +/- ritonavir	Darunavir /ritonavir	Fosamprenavir /ritonavir	Indinavir* /ritonavir	Lopinavir /ritonavir	Nelfinavir /ritonavir	Saquinavir/ ritonavir	Tipranavir** /ritonavir
74		T → P						T → P
76		L → V	L → V	L → V	L → V			
77				V → I		V → I	V → I	
82	V → (A,T,F,I)		V → (A,F,S,T)	V → (A,T,F)	V → (A,F,S,T)	V → (A,F,S,T)	V → (A,F,S,T)	V → (L,T)
83								N → D
84	I → V	I → V	I → V	I → V	I → V	I → V	I → V	I → V
85	I → V							
88	N → S					N → (D,S)		
89		L → V						L → (I,M,V)
90	L → M		L → M	L → M	L → M	L → M	L → M	
93	I → (L,M)							
No. of major mutation positions	3	5	2	3	4	2	2	6
No. of minor mutation positions	21	5	9	11	13	8	9	8

Notes:

1. Major resistance mutations are highlighted in grey

2. * Information concerning indinavir is not comprehensive due to lack of recent research/ updates on the drug.

3. **Resistance mutations against tipranavir have not been validated on large datasets

(Adapted from Wensing et al. 2014)

Problem statement

Most of the protease inhibitors have been developed based on HIV-1 subtype B, which does not tally with their worldwide proportions (Velázquez-Campoy et al. 2003; Wensing et al. 2010). Rapid mutations (Ragland et al. 2014) and buildup of resistance by HIV-1, combined with patient tolerability, toxicity and adherence to treatment (Pokorná et al. 2009) make that current treatment regimens might not be effective in the future as resistance is built up against them. Wet lab methods such as NMR and X-Ray crystallography enable drug discovery, but are too slow (Floudas et al. 2006) when faced to the complexity of HIV subtypes and their rapid mutation rates. Comparative modeling combined with docking and molecular dynamics present themselves as viable tools for accelerating drug discovery against the HIV-1 critically-important protease, being a known successful target (Sundquist & Kräusslich 2012; Pokorná et al. 2009) to slow down the progression of HIV-1.

Aims, objectives and motivation

Aims:

1. To investigate the effect of Lopinavir (LPV) treatment on the structure of HIV-1 (subtype C) proteases from South African patients and its impact on resistance against other FDA-approved protease inhibitors

2. The discovery of new potential HIV-1 subtype C protease inhibitors

Objectives:

1. Retrieving HIV-1 subtype C protease protein sequences from drug-treated South African patients from the Stanford University HIV Drug Resistance Database.
2. Detecting sequence variations by carrying out per-patient pairwise sequence alignments.
3. Gathering mutation information (minor & major mutations) for every protease sequence.
4. Building the 3D structure of every HIV-1 protease sequence by comparative modeling.
5. Validating the homology models using a series of quality assessment tools.
6. Ligand-docking of the homology models to investigate the effect of the protease mutations on ligand binding.

7. Molecular dynamics of the modeled protein, in the presence of docked potential lead compounds.

Motivation:

Rapid build-up of resistance against the conventional protease inhibitors elicits the need to find alternative to them. The study of affinity-altering protease mutations in HIV-1 (subtype C) patients undergoing treatment using current PI's can help improve the understanding of the establishment of the underlying resistance mutations and can at the same time provide information suitable for the discovery or design of new drugs, as potential alternatives. The screening a database of natural products such as the South African natural compounds database (SANCDB) might reposition compounds of known use, shortening the time for testing and reviewing before approval (normally around 5-10 years (Mullard 2014)), such as is the case for compounds with no known previous use.

The flowchart below gives an global overview of the project research methodology, starting with sequence retrieval to modeling and receptor-ligand molecular dynamics (MD).

Figure 1.4: Overview of methodology to be used for molecular simulations using HIV-1 proteases of the C subtype

CHAPTER 2: Identification of protease mutations and phylogenetic analysis

In this chapter, HIV-1 protease mutations are revealed by different sequence alignment approaches and a phylogenetic tree is drawn to give a bird's eye view of the extent of mutations on the distance between shortlisted patient protease sequences before and after LPV treatment.

INTRODUCTION

Phylogenetic methods are often used to group HIV strains in different ways, such that the type of information can give hints about the strain type, provenance, virulence and timing of infections (Castro-Nallar et al. 2012). Many methods of building phylogenetic trees are available and have their advantages and disadvantages, particularly when it comes to accuracy, speed and the type of information that can be derived from the computations (Yang & Rannala 2012; O'Meara 2011). These methods can be generally classified into character-based and distance-based methods (Yang & Rannala 2012), depending on the information they take from a defined multiple sequence alignment (MSA). It is essential that the OTU's (Operational Taxonomic Units) are homologous, as it is assumed that every individual column is homologous across each of a set of aligned sequences (Hall 2013; Brocchieri 2001; Baldauf 2003). In other words the input sequences of the alignment have to be derived from a common ancestor (Patwardhan et al. 2014). The OTU, which can be represented by a stretch of nucleic acids or amino acids may in some cases reflect the ancestries of the whole organisms (species trees) or in others, reflect only those of the gene(s) or sequences being studied (gene trees) (Page & Charleston 1997).

HIV protease sequences are usually obtained by RNA extractions from blood plasma, followed by reverse transcription, amplification and sequencing (Ariffin et al. 2014; Jiao et al. 2014; de Felipe et al. 2011). Strain subtype information (as well resistance information) can then be obtained by submitting sequences to the HIVdb program, available from the Stanford HIV Database (Liu & Shafer 2006).

Phylogeny has many applications in the context of HIV-1 research. It has been used in the study transmission networks of the disease population-wide (Brenner & Wainberg 2013), in understanding the origins of the virus (Hemelaar 2012; Tongo et al. 2015), in the classification of the virus into different subtypes (Sharp, P. M. and Hahn 2011) and in many more cases. The data that is usually utilized to reconstruct HIV phylogenies is in the form of cDNA obtained from the extracted RNA,

as mentioned earlier. The Neighbor Joining algorithm together with the Kimura two-parameter (K2P) model of nucleotide substitution are commonly used for building the phylogenetic trees of HIV-1 (Monno et al. 2012; Pessoa et al. 2011; Fu-xiang et al. 2007; Galkin et al. 2006; Boeri et al. 2004). The approach used by neighbor-joining is to pair OTU's such that the total branch length at each clustering step is minimized (Saitou & Nei 1987), while the K2P model is a model of DNA evolution that attempts to account for unobserved mutation states by including a rate for transitions and another one for transversions (Patwardhan et al. 2014).

METHODOLOGY

2.1. Data retrieval and organization

The raw dataset for the HIV-1 proteases was obtained from Dr Soo-Yoon Rhee from the Stanford HIV Database and parsed with a Python script to extract and reorganize patient ids, treatment regimens, protein sequences, reference ids and year of treatment. The sequences were filtered so that full-length (99 residues long) protease sequences were retrieved for South African HIV-1 subtype C patients, with unambiguous amino acid residues. Redundancy between sequences was removed in a 2-step approach. The dataset was already arranged by year, however the patient entries were scattered amongst other entries – i.e. a patient could have several entries for drug treatment and several other without. In this context, unique sequences were gathered thus: for every patient in the list of entries in the raw dataset, only the sequence corresponding to the last year of treatment (or without treatment) was retained, so that each patient ended up with at most a sequence before and another one after treatment. Finally, pairwise sequence comparisons (only by string matching) were performed to flag (but not remove) the duplicates.

2.2. Data filtering

For every patient, the drug-naive and the LPV-treated (actually LPV/ RTV) aligned on top of the other and the mutations were represented as wild-card characters underneath every patient entry, as shown in the Figure 2.1. No sequence alignment tool was used due to the way by which the protein sequences were filtered from the raw dataset (sequences were of same lengths and no insertions/ deletions were allowed).

These sequences were concatenated into a single FASTA-formatted record without MSA. cDNA sequences were also concatenated as a separate FASTA file without MSA. Both records were

considered henceforth as MSA's (due to their equal lengths and homologous positions) and were visualized in Jalview (Waterhouse et al. 2009) (shown in Figures 2.2 and 2.3 respectively).

2.3. Retrieval of HIV-1 protease mutation information

Mutation information (after LPV treatment) was retrieved by sending the FASTA-formatted cDNA sequences to the Stanford HIVdb web tool.

2.4. Phylogenetic tree construction

A phylogenetic tree was also built for the aligned protein sequences, using the Maximum Likelihood method as implemented in the MEGA6 (Tamura et al. 2013) however the method produced a high proportion of very low bootstrap values (due to the high sequence identities across the sequences). Therefore, the coding nucleotide sequences were used instead to represent more of the variability between the sequences. The Neighbor Joining tree building method was used with 1000 bootstrap replicates using the Kimura 2-parameter (K2P) nucleotide substitution model, as used in Yang et al. (2000) and Jiao et al (2014). An outgroup (NCBI accession: JN835461.1) was retrieved by a local BLASTN search against the SIV taxon using a protease-coding nucleotide sequence from the alignment – the aligned hit was selected (and concatenated to the alignment) on the basis of low E-value, 100% coverage, with moderate sequence identity (83% in this case). The concatenated alignment was visually examined in Jalview and no anomalies were found, as the hit was a 100% coverage match, so no tool was used for alignment.

RESULTS AND DISCUSSION

Out of the 83,654 HIV-1 proteases sequences found in the received raw dataset containing sequences from multiple subtypes from patients following different treatment regimens (in different countries worldwide), only 11 patients were found to have sequences that had an entry corresponding to before and after treatment (labeled as “None” and “LPV”, respectively in the raw dataset). Only one of the protein sequences was found to be redundant across all patient sequences before and after LPV treatment (Figure 2.1).

For each patient ID, the mutations observed and their resistance status (as determined from HIVdb) after LPV treatment are summarized in Table 2.1.

Table 2.1: Summary of the mutations after LPV treatment and their resistance status

Patient ID	Protease mutations after LPV treatment	Resistance status of HIV protease after LPV treatment (HIVdb interpretation from consensus B HIV protease)
115754	S12T, V15I	T74S: minor resistance mutation; reduced susceptibility to NFV
116574	I14T, T19K	T74S: minor resistance mutation; reduced susceptibility to NFV
116329	P63L	No major or minor mutations
116499	R20K, K45R	No major or minor mutations
116136	N37S	No major or minor mutations
115568	L63P	No major or minor mutations
115755	R45K	T74S: minor resistance mutation; reduced susceptibility to NFV
83052	T12S	No major or minor mutations
115316	L89M	No major or minor mutations
115633	T19I	No major or minor mutations
116153	S12T, I19L, K20R, E35D, N37D , K41R, R56K, P62L, L89M	No major or minor mutations

As can be observed from the above table, most of the mutations did not lead to any resistance mutation against the common PI's, except for patient 115754, 116574, 115755 where minor resistance occurred, surprisingly not against LPV, but against NFV. As could be estimated from the years available from the records (refer to Figure 2.1), the amount of time between the viral cDNA sampling varied between a minimum of 1 year and 5 years (at most) and in spite of the mutations that arose after treatment, none were observed against LPV, meaning that it was most likely that the viral particles could still be successfully-controlled if used, as long as the regimen was respected. In addition, other drugs could also have been successfully used to treat the patients, with the only decreased efficiency of NFV in some of the cases.

A summary of the pairwise sequence alignments is shown in Figure 2.1 for the shortlisted HIV-1 protease sequences. The entries are arranged with the sequences tagged before LPV treatment being on top of the sequences obtained after LPV treatment. No indels were observed due to the fact that the sequences had been filtered to have identical lengths. The wild card characters show that the number of mutations occurring after treatment are few, except in the case of patient 116153, where a higher number of mutations (nine) is observed. Possible explanations for the higher number of mutations are linked to a higher rate of mutation of the virus, reasons for the latter probably being linked to variants of the host HIV-1 mRNA editing A3 enzymes and variants of the HIV-1 Vif protein.

```
>115754 refID: 2122
 None =2010  PQITLWQRPLVSIKVGGQTKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIVIEICGKKAIGSVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2011  PQITLWQRPLVTIKIGGQTKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIVIEICGKKAIGSVLVGPTPVNIIGRNMLTQLGCTLNF
                         *  *

>116574 refID: 2122
 None =2008  PQITLWQRPLVTIIIGGQTREALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDEVLIEICGKRAIGSVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2011  PQITLWQRPLVTITIGGQKREALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDEVLIEICGKRAIGSVLVGPTPVNIIGRNMLTQLGCTLNF
                           *    *

>116329 refID: 2122
 None =2007  PQITLWQRPLVTIKVGGQLKEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVKQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2009  PQITLWQRPLVTIKVGGQLKEALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVKQYDQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
                                                                  *

>116499 refID: 2122
 None =2009  PQITLWQRPLVSIKVGGQIREALLDTGADDTVLEDINLPGKWKPKMIGGIGGFIKVRQYEEILIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2011  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEDINLPGKWKPRMIGGIGGFIKVRQYEEILIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
                                 *                           *

>116136 refID: 2122
 None =2009  PQITLWQRPLVTIRIGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIAIEICGKKAIGTVLVGPTPINIIGRNMLTQLGCTLNF
 LPV  =2011  PQITLWQRPLVTIRIGGQIKEALLDTGADDTVLEEISLPGKWKPKMIGGIGGFIKVRQYDQIAIEICGKKAIGTVLVGPTPINIIGRNMLTQLGCTLNF
                                            *

>115568 refID: 2122
 None =2008  PQITLWQRPLVSVKIGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYEQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2011  PQITLWQRPLVSVKIGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYEQIPIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
                                                                  *
```

```
>115755 refID: 2122
 None =2007  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEINLPGKWKPRMIGGIGGFIKVRQYDQIPIEICGKKAIGSVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2011  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGSVLVGPTPVNIIGRNMLTQLGCTLNF
                                                          *

>83052 refID: 1922
 None <2006  PQITLWQRPLVTIKVGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNLLTQLGCTLNF
 LPV  <2010  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNLLTQLGCTLNF
                      *

>115316 refID: 2122
 None =2010  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEISLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNLLTQLGCTLNF
 LPV  =2010  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEISLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
                                                                                                      *

>115633 refID: 2122
 None =2007  PQITLWQRPLVSIKVGGQTKEALLDTGADDTVLEEINLPGKWKPRMIGGIGGFIKVRQYDQITIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
 LPV  =2009  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEINLPGKWKPRMIGGIGGFIKVRQYDQITIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
                           *

>116153 refID: 2122
 None =2009  PQITLWQRPLVSIKVGGQIKEALLDTGADDTVLEEINLPGKWKPKMIGGIGGFIKVRQYDQIPIEICGKKAIGTVLVGPTPVNIIGRNLLTQLGCTLNF
 LPV  =2010  PQITLWQRPLVTIKVGGQLREALLDTGADDTVLEDIDLPGRWKPKMIGGIGGFIKVKQYDQILIEICGKKAIGTVLVGPTPVNIIGRNMLTQLGCTLNF
                      *     **              * *   *              *     *                        *
```

Figure 2.1: Figure: Per-patient sequence alignment of full-length HIV-1 protease protein sequences before (None) and after treatment by lopinavir (LPV). The first line of each entry contains the patient ID, followed by the reference ID. Sequences for patient ID 83052 (for the LPV treatment) and patient ID 116153 (before treatment, labeled as "None") are identical.

Shown in Figure 2.2 is the MSA of all of the selected proteases protein sequences. The bar plots at the bottom of the figure clearly show a high degree of conservation with only a few variable residue positions. It is this lack of sequence divergence that lead to poor bootstrap values when used for phylogenetic analysis by the Maximum Likelihood tree building approach. For this reason, the cDNA sequences were used instead – the MSA is shown in Figure 2.3, where more variation is observed (disregarding the ambiguous bases), indicating that most mutations were synonymous.

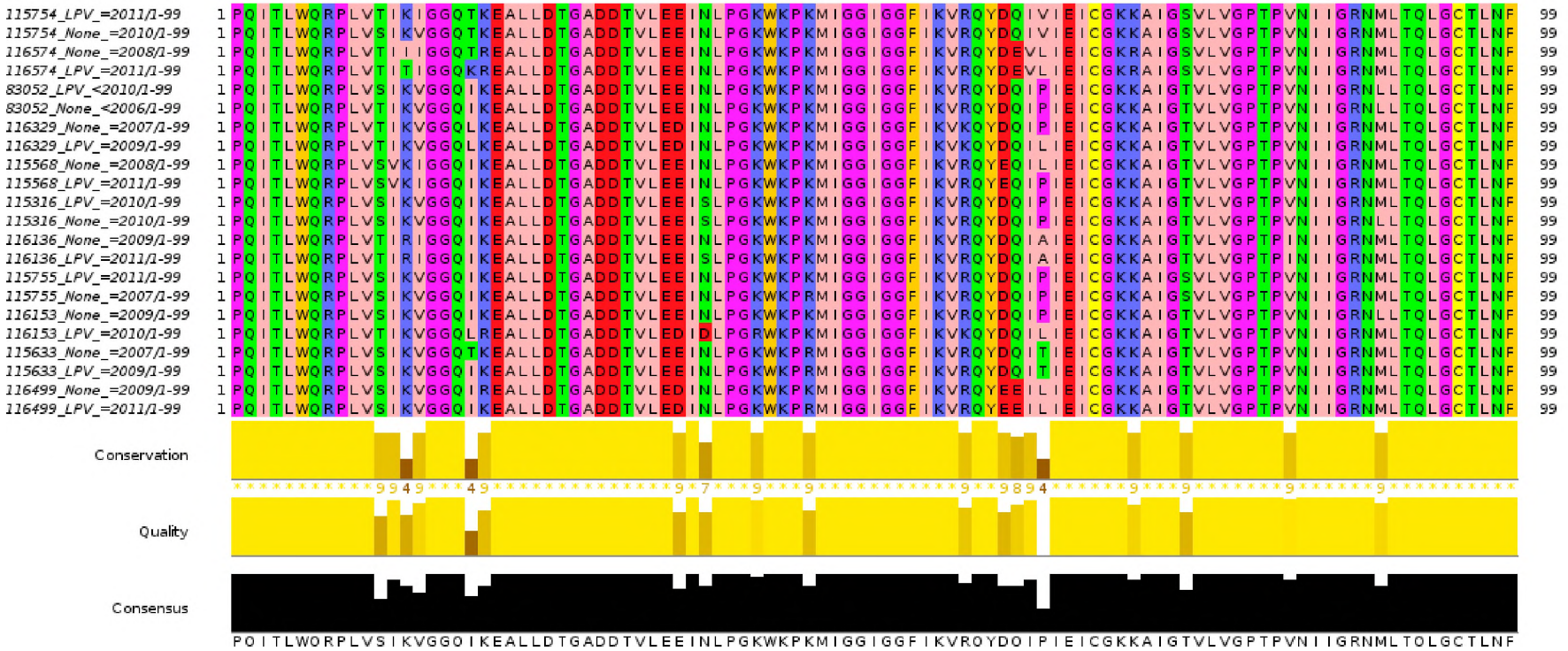


Figure 2.2: Multiple sequence alignment of protease protein sequences from patients before and after LPV treatment, colored by physicochemical property.

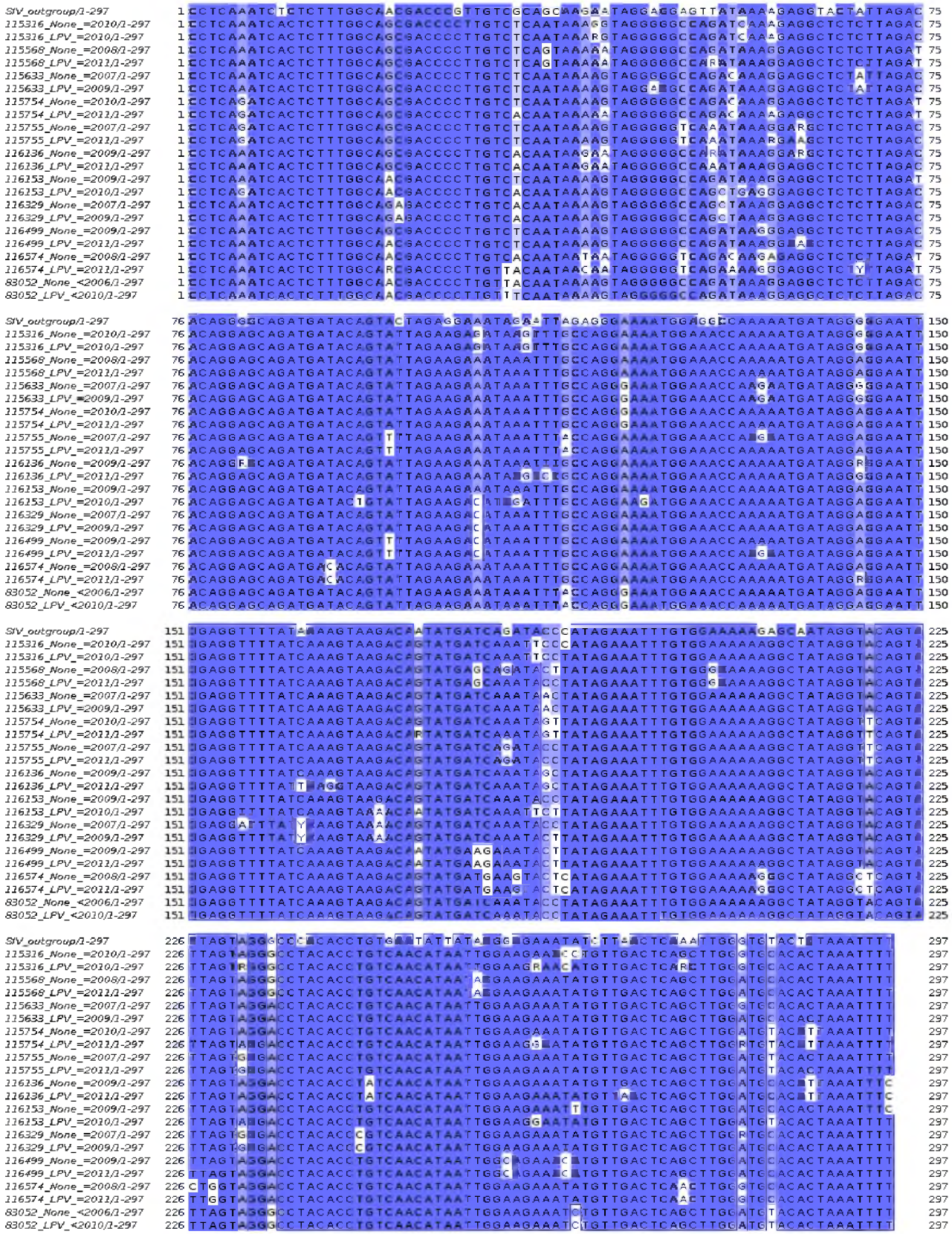


Figure 2.3: Multiple sequence alignment of the protease protein-coding nucleotide sequences, color-coded by sequence identity, using JalView. An outgroup from a Simian Virus (SIV) was retrieved from a BLAST search and added to the alignment to root the phylogenetic tree.

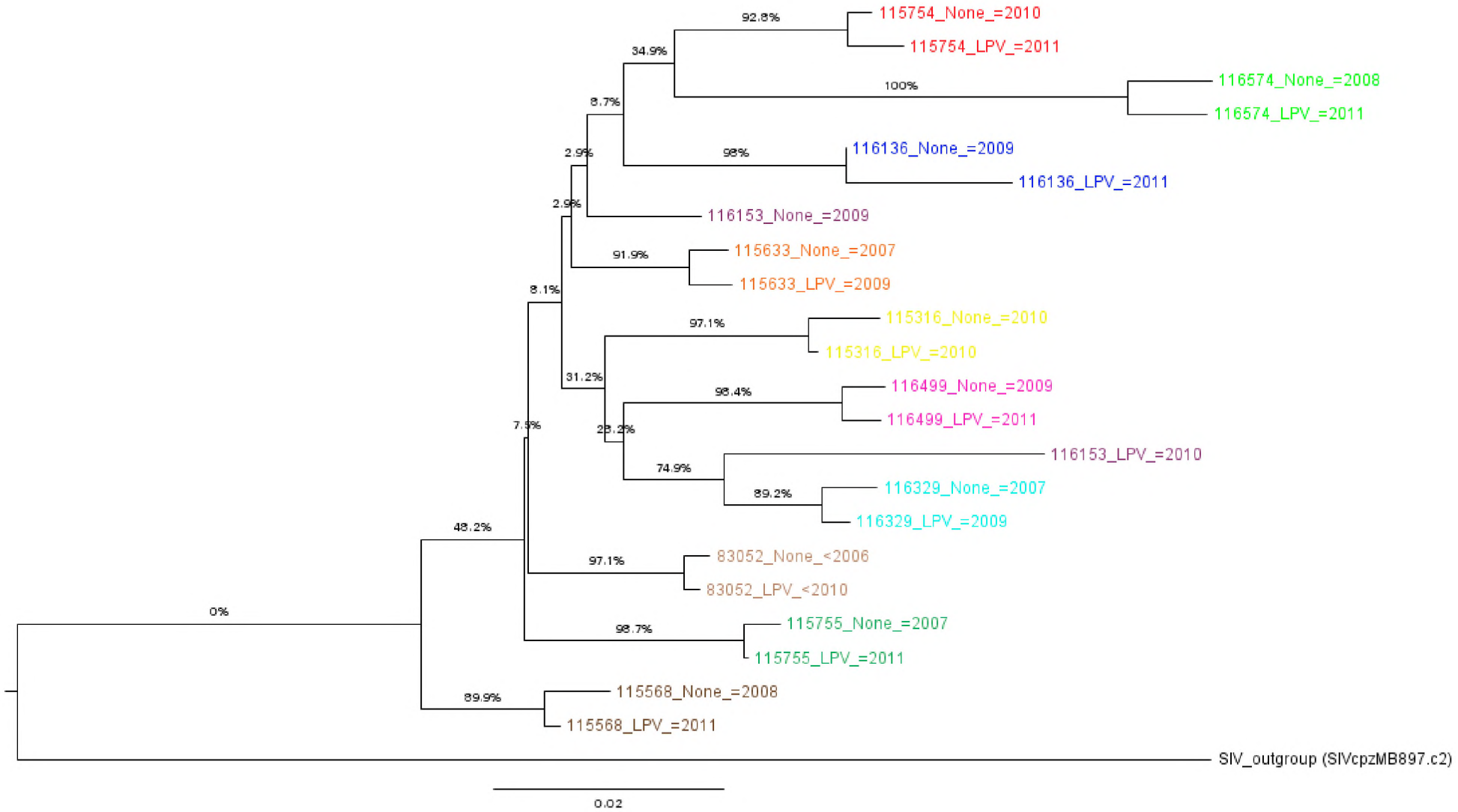


Figure 2.4: Neighbor-joining tree of the generated, using 1000 bootstraps and the K2P nucleotide substitution model, using MEGA6. The nodes were color-coded patient-wise, using the FigTree tool.

The neighbor-joining tree shown in the previous figure (Figure 2.4) shows clusterings among the different cDNA sequences analyzed, with the nodes representing the patient proteases, the branches representing the connections between them, the lengths of which represent the genetic (additive) distances between the sequences. As expected, most of the sequences before and after LPV treatment cluster together with high bootstrap support, however there is one sequence that does not follow the same trend and clusters across different clades (sub-trees) in the tree – the patient sequence with ID 116153. This observation should be handled with caution as the bootstrap support values between the clades connecting the clusters are very low (<70%), decreasing the confidence in the architecture of the clustering. In this case it would be more reasonable to remain on the observation that the sequences are moderately distant, with the sequence after LPV treatment being closer to the cluster for patient ID 116329. The sequence before LPV treatment was not reliably clustered, probably due to its high degree of similarity to many other sequences present in the tree. This hypothesis was verified by calculating the identity matrix (all against all, using BioEdit) (Hall 1999) of the same sequence set and computing the row means to identify their average identity. The sequence labeled "116329_None" was found to have the highest mean identity (shaded in gray in Table 2.2). Also observable from the same table is that most sequences are also individually very close to all others, explaining the occurrences of more low bootstrap values despite using the nucleotide sequences.

Table 2.2: Row-wise (one versus all) average identity

Sequences	Average identity
SIV_outgroup (SIVcpzMB897.c2)	0.785
116136_None_=2009	0.891
115316_None_=2010	0.888
115316_LPV_=2010	0.885
115568_None_=2008	0.896
115568_LPV_=2011	0.897
115633_None_=2007	0.900
115633_LPV_=2009	0.901
115754_None_=2010	0.899

115754_LPV_=2011	0.888
115755_None_=2007	0.891
115755_LPV_=2011	0.891
116136_LPV_=2011	0.883
116153_None_=2009	0.906
116153_LPV_=2010	0.880
116329_None_=2007	0.891
116329_LPV_=2009	0.895
116499_None_=2009	0.895
116499_LPV_=2011	0.891
116574_None_=2008	0.872
116574_LPV_=2011	0.868
83052_None_<2006	0.901
83052_LPV_<2010	0.901

CONCLUSION

Out of the 83,654 HIV-1 proteases sequences obtained, 22 sequences matching selection criteria were retrieved with differences occurring after treatment. The cDNA-based phylogenetic tree and the table of one-versus-all protein sequence identity support that the majority of the sequences have enough variability (with the exception of identical protein sequences from patients 83052 (LPV-treated) and 116153 (before treatment), which can potentially introduce changes in the 3D structure of the proteases, the effects of which are later assessed by firstly modeling the proteases (Chapter 3), followed by docking against various FDA-approved protease inhibitors (Chapter 4).

CHAPTER 3: Homology modeling of the HIV-1 proteases

This chapter describes the concept of homology modeling and explains the steps involved, starting from a target protein sequence to obtaining a 3D model of the sequence. There are many steps during protein structural modeling where things can go wrong. Fortunately, various independent tests are available to support these models, and these are also described.

INTRODUCTION

3.1. Homology (Comparative) modeling

Structure determination is very important for protein characterization and drug discovery. However, the *de facto* methods (NMR and X-ray spectroscopy) can fail with some proteins such as membrane proteins, and are still too slow (Floudas et al. 2006) to keep up with the massive surge in the number of incoming sequenced genes and genomes (Ginalski 2006), limiting their functional characterization (Fiser & Šali 2003). For this reason, comparative modeling (also referred to as homology modeling) still remains a method of choice (Ginalski 2006) for structure determination as it presents itself as a fast tool for drug discovery (Vyas et al. 2012). The accuracy and reliability of the method are deemed adequate to bridge this sequence/ structure gap (Ginalski 2006).

In general, comparative modeling methods comprise the following steps:

3.1.1. Template identification

The target sequence is used to search against a database of sequences with known structures, which usually is a Protein Data Bank (PDB) – related database (Fiser & Šali 2003). A variety template retrieval tools (such as NCBI, PDB, CATH, SCOP, MODBASE, BLAST and HHPred) and model-building tools (such as SWISS-MODEL, MODELLER and HHPred) are available, the addresses of which can be found in the list of useful web links. In case of low similarity of the template sequence to known structures, it can be useful to use different methods of finding related structures (Fiser & Šali 2003). However, higher sensitivities may be achieved when the number of sequences matching the target is large (Fiser & Šali 2003). Such a condition would enable the building of profiles and/ or the use of Hidden Markov Models, which are more sensitive (Fiser & Šali 2003).

3.1.2. Selection of templates

The search for templates may return several potential templates. Depending on the modeling problem, one or more templates regarded as appropriate may be retained (Fiser & Šali 2003). Several parameters need to be considered for choosing suitable templates, namely, the overall sequence similarity of the target and template(s) should be high; the target and the template should be relatively close inside their protein family; the environments of the target and the template(s) should be similar; the quality of the experimentally-determined template structure(s) should be high; the context of the comparative modeling simulation should coincide with the selected template - for example, the template might require a bound ligand, if docking needs to be done against the target at a similar site (Fiser & Šali 2003).

3.1.3. Alignment of target sequence with (one or more) template(s)

Homology modeling requires the alignment of the target sequence against one or more template sequences for which structures are available. The alignment relies on the structural equivalence between the target and template residues (Fiser & Šali 2003) at each aligned position. The alignment is more straightforward when the percentage identity between template(s) and the target sequence is above 40%, however lower percentage identities between them usually increase the probability of misaligned residues (Fiser & Šali 2003). A single misaligned residue can result in a discrepancy of about 4Å (Fiser & Šali 2003). Such alignments generally also require the introduction of gaps, which may cause problems if placed in regions corresponding to secondary structure elements, buried regions or between physically distant residues (Fiser & Šali 2003). Possible solutions to the alignment problems include 1) manual alignment editing, 2) prior structure-based alignment of the templates amongst themselves before aligning against the target, and 3) the use of independent target and template alignment profiles with the help of non-redundant sequence databases, which result in a profile-profile alignment between the template and target (Fiser & Šali 2003).

Overall, alignment accuracies using several modeling approaches are nearly equivalent when used optimally and it is in fact the template choice and the alignment accuracy which can substantially impact on the accuracy of the resulting model(s), especially when there is less than 40% sequence identity between the template and target sequences (Fiser & Šali 2003).

3.1.4. Model-building for the target

Many approaches are available for model-building, including rigid-body assembly, modeling by segment-matching and modeling by statistical potential-guided conformational searches (Fiser & Šali 2003; Blundell et al. 1987; Kolinski et al. 2001; Jones & Thirup 1986; Unger et al. 1989; Claessens et al. 1989; Levitt 1992). However, it is the approach used by MODELLER that will be used for the modeling, and is termed modeling by satisfaction of spatial restraints. This method is based on distance geometry and optimization techniques that help satisfy spatial restraints obtained from the sequence/ structure alignment (Havel & Snow 1991; Fiser & Šali 2003; Brocklehurst & Perham 1993; Aszódi & Taylor 1996). More details of the approach are given in the section entitled "Using MODELLER for comparative modeling".

Loops often define functional specificity, but are normally found as as gaps in MSA's as they are not evolutionarily conserved, allowing room for structural variability to often result in models that are more difficult to predict (Vyas et al. 2012). Nevertheless, loops can be modeled in two ways, namely by using the *ab initio* or by the use of database approaches (Fiser & Šali 2003). The a*b initio* approach is guided by an energy-scoring function combined with conformational searches (Fiser & Šali 2003; Samudrala et al. 1999). On the other hand, the database approach uses main chain segments corresponding to the stem regions of the loop to search for known structures (even non-homologs) that are then superposed, annealed and finally optimized by an energy function (Fiser & Šali 2003).

3.1.5. Model evaluation

Model evaluation is an important step that checks for possible errors introduced during modeling (Fiser & Šali 2003). In general, the accuracy of a model can be represented as a function dependent on the percentage sequence identity of the target to the templates - it is especially critical when the percentage sequence identity falls below 30% as model reliability drops significantly at these levels (Eswar et al. 2008). Other factors such as differences in the environment can result in significant deviations from the expected model due to conformational changes that arise in a functionally-irrelevant environment (Fiser & Šali 2003).

The evaluations themselves can be categorized as internal or external, based on whether the evaluations utilize factors that are intrinsic or extrinsic to the model, respectively (Fiser & Šali 2003). Internal evaluation is based on self-consistency checks and/ or the satisfaction of restraints (Fiser & Šali 2003), while external evaluation is based on additional external information not used during model construction (Lüthy et al. 1992; Sippl 1993). Internal evaluation of model stereochemistry includes the verification of bonds, bond angles, phi and psi angle anomalies, non-

bonded atomic distances, close contacts, amongst others, and can be carried out using the tools WHAT_CHECK and PROCHECK (Fiser & Šali 2003; Laskowski et al. 1993; Hooft et al. 1996; Xiong 2006). On the other hand, external evaluation is primarily used to check for the correctness of the templates used, and is specially important when dealing with sequence similarities less than 30% or when the templates used have different alternative folds (Fiser & Šali 2003). An example of such a tool is PROSAII (Fiser & Šali 2003; Sippl 1993). Various other model validation approaches are available including z-DOPE (Šali 2013), ANOLEA (Atomic Non-Local Environment Assessment) and Verify3D (Xiong 2006), which assign a score (local) at every residue position, while others such as DFire (Zhou & Zhou 2002) and QMEAN6 (Benkert et al. 2011), which give a single score (global) representative of the whole structure.

The z-DOPE is a standardized metric used by Modeller, derived from the knowledge-based DOPE (Discrete Energy Optimized Protein Energy) potential - a positive z-DOPE value would indicate a poor model while a model with a value of -1 or lower would indicate closeness to the native conformation of the protein (Šali 2013). The DOPE score itself is a statistical potential calculated from a sample of existing PDB structures having native conformations (Shen & Sali 2006).

ANOLEA is a web service that performs energy calculations for a given model. Local regions of the model that have high energy are usually correlated with errors or potential interacting zones of proteins, and are flagged as unfavorable (Melo et al. 1997).

The DFire score is derived from an all-atom, distance-dependent potential energy calculation used to predict protein stability and assist in their selection (Zhou & Zhou 2002; Zhang et al. 2004).The lower the score, the closer the protein conformation is to being native (Zhang et al. 2004).

QMEAN6 is one of the QMEAN derivatives, where the number 6 corresponding to the number of structural descriptors (including potentials) used in a linear combination to evaluate the score (Benkert et al. 2011). As DFire, the QMEAN score also provides the degree of nativeness of the protein, but is normalized by protein length, standardized, and sign adjusted to range between zero and one such that the more native-like the protein, the closer the score is to one (Benkert et al. 2011).

It should however be noted that no single method is the best and that it is advisable to use multiple validation tools and find the consensus between them (Xiong 2006).

3.1.6. Iteration

Based on the output from the model evaluations, the model can be improved by repeating the process of template selection, alignment, model construction and quality assessments until no further improvement is observed (Fiser & Šali 2003; Sánchez & Sali 1997; Guenther et al. 1997; John & Šali 2003).

3.2. Useful quality metrics from PDB to assist in template selection

PDB includes various types of information to support the validity of the protein crystal structures it hosts, including resolution and several other metrics such as the R-free values, clash scores, side chain outliers and RSRZ outliers (Read et al. 2011) that are summarized as multi-percentile reports.

The R-free value indicates how well the protein model fits a non-refined fraction of the protein while the clash score indicates the number of clashing atom pairs (wwPDB 2014). The side chain outliers indicate the percentage of residues with unusual side chain conformation, and the RSRZ is a Z-score based on a statistic called real space R-value that can only judge the relative quality of proteins (wwPDB 2014; Read et al. 2011).

3.3. Using MODELLER for comparative modeling

MODELLER is an automated tool that finds the most probable structure for a target sequence based on a sequence alignment against sequences having similar structures (Gromiha 2010). To do so, spatial restraints are extracted and optimized from the coordinates of the atoms corresponding to the aligned template sequences (Feyfant et al. 2007). The model is then expressed as probability density functions (pdfs) based on the restrained features, which include the C^{α}-C^{α} distances, main chain N-O distances and dihedral angles for the side chain and main chain (Šali & Blundell 1993). The process also includes smoothing procedures to minimize problems related to the use of sparse databases (Gromiha 2010). The final model is then optimized so that it does not violate the input restraints by much (Šali & Blundell 1993).

METHODOLOGY

3.1. Homology modeling

3.1.1. Template selection

Possible templates were initially searched using HHpred (locally and online), however the hits were few and the template were generally of low resolutions. Thus, local BLAST (Protein BLAST 2.2.28+) searches against PDB were resorted to, using the BLOSUM62 substitution matrix, with an E-value cut-off of 1e-4:

Seven criteria were used for template selection: two chains were required to be present for every structure; crystal structures were to be of high high resolution (≤ 2 Angstroms); residues should not be missing; a ligand should be found at or close to the active site; the template should have a high coverage with respect to the target sequence; there should be a high percentage identity between the target and the template sequence; template structures should have a good overall quality, based on PDB quality reports, which involved low R-free values, low clash scores, minimal side chain outliers and RSRZ (real space R-value Z-Score) outliers values less or equal to 2. Python and bash scripts were written to carry out the preliminary filtering tasks (2 chains; high resolution; no missing residues; ligand at active site; high coverage and percentage identity). Another Python script was then used to retrieve the multi-percentile validation charts from PDB for further filtering, according to the different available metrics, comprising any of the R-free scores, clash scores, side chain outliers and RSRZ outliers – resolution was also included. An R script was then used to generate a bar-plot (Figure 3.2). The presence of the ligand was assessed visually using the Discovery Studio software (version 4.1) (Dassault Systèmes BIOVIA 2015). The method only returned candidates for the closed conformation template. Therefore, the open conformation template (PDB accession: 3BC4) was independently searched and retrieved on the basis of high resolution (1.82 Angstroms), the presence of a ligand at the active site and the use the structure in a publications. Structure 3BC4 was thus retrieved, available as a biological assembly and thus the other half of the protease had to be mirrored by using a transformation matrix provided with the crystal structure.

Discovery Studio was used to remove the solvent (water) and the co-crystallized ligand(s) from the PDB structures. The ligands were separated from the structures and saved in PDB format. The use of the ligands is reviewed in Chapter 4.

3.1.2. Sequence alignment

The MUSCLE software (version 3.8.31) was used for the template/ target sequence alignments for every model to be built and the alignments were visually assessed for appropriate residue placement. The FASTA files were then converted to ".ali" format, which is very similar to the PIR alignment format, using a Python script.

3.1.3. Model-building

Modeling and z-DOPE scoring scripts were adapted from the Modeller (version 9.14) automodel example and assessment files respectively. One hundred models were built for each patient's treatment status with both the open and closed conformations. A schematic is shown below:

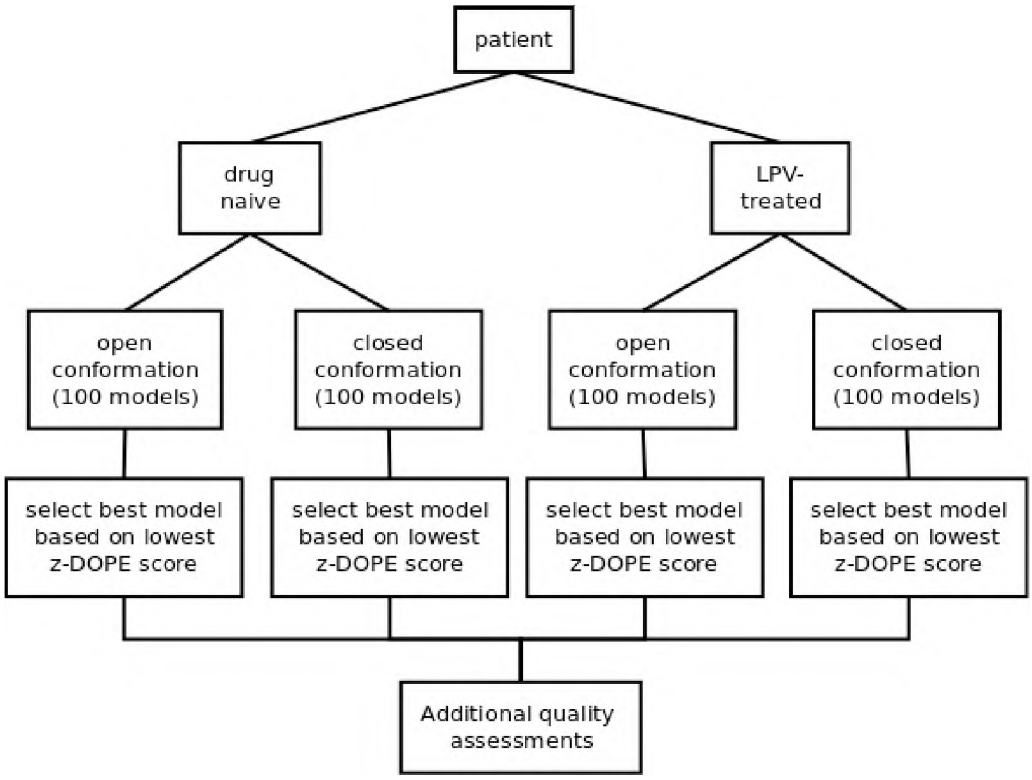


Figure 3.1: Schematic of the organization of model building, selection and quality assessments.

Each of the 100 models were summarized by their z-DOPE scores as box plots in Figure 3.1. One model was selected for every modeled conformation according to the lowest z-DOPE score obtained from MODELLER. Scripts used for modeling, z-DOPE scoring and plotting are shown in Supplementary Materials Section 4.2, 4.3 and 4.4, respectively.

3.2. Independent evaluation of model quality

Each of the best models (from z-DOPE score evaluations) were uploaded in PDB format to the SWISS-MODEL server to evaluate both local (ANOLEA) and global quality metrics (DFire and

QMEAN6). ANOLEA plots are shown in the Supplementary Materials Section 2, while DFire and QMEAN6 were plotted against each other in a scatter plot (Figure 3.4).

RESULTS AND DISCUSSION

The following bar plot (Figure 3.2) shows standardized metrics obtained from PDB for the selection of the closed conformation template. Each of these metrics have been standardized individually, using their individual means and standard deviations. The rationale was to obtain a structure for which all the metrics were available and which had the lowest values among the gathered crystal structures. Even though many structures displayed metrics that had much lower standardized values, structure 3PWM was selected as closed template as it had all the chosen metrics.

Figure 3.3. shows the distribution z-DOPE scores that were evaluated for all of the models built using MODELLER. It can be easily seen that the scores for the closed conformation models are much lower overall compared to those modeled from the open conformation template, indicating that the quality of the modeled structure is very much dependent on the initial quality of the template. The gray line (at y = -1) marks the region where the models are native-like. Most of the minimum points from the box plots are below this line if not very close (refer to Table 3.1 for actual values), meaning that the models are of good quality. Global scores from additional quality assessments from DFire and QMEAN6 also show that the models are of good quality, and once more a clustering is observed whereby the structures open and closed conformation models inherited characteristics similar to their respective templates. In general the open conformation models were of lesser quality than the closed conformation models. This discrepancy might be due to the fact that the protease was maintained in an unusual opened conformation by the ligand interfering with flap closure for the protease (the model is shown in chapter 4, Figure 4.1 C-D).

Figure 3.2: Summary of standardized multi-percentile quality metrics for choosing a template from PDB

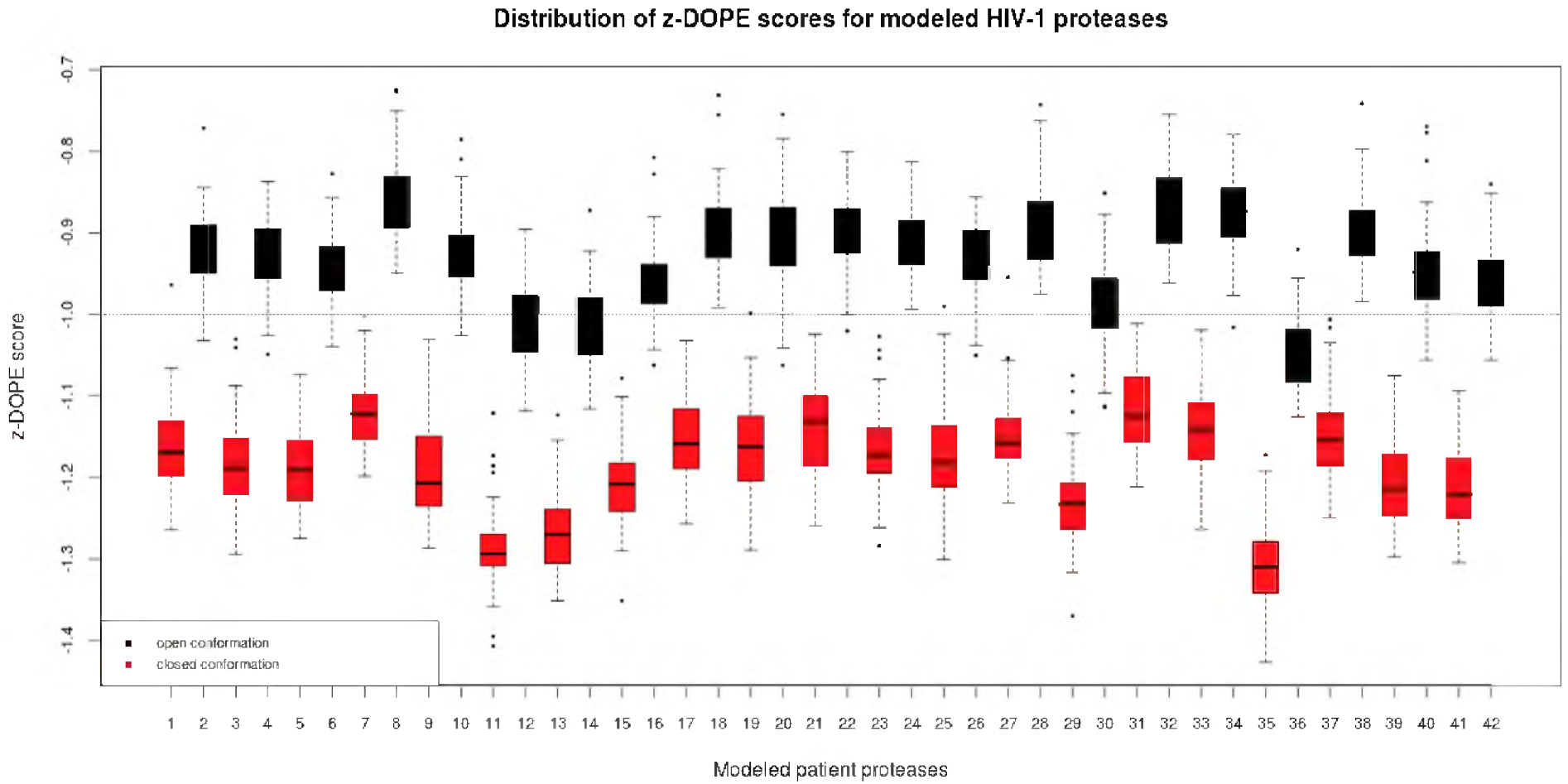


Figure 3.3: Showing the distribution of z-DOPE scores for all the models.

Summary of the z-DOPE scores for the modeled HIV-1 proteases, for the open and closed conformations, before and after LPV treatment.

Table 3.1: Models of lowest z-DOPE scores (Labels correspond to the bar plot x-axis labels in Fig. 10.)

Label	Conformation	Model name (Patient id/ treatment/ year/ model no.)	Z-DOPE score (3dp)
1	closed	115316_LPV_=2010.B99990064.pdb	-1.264
2	open	115316_LPV_=2010.B99990011.pdb	-1.032
3	closed	115316_None_=2010.B99990092.pdb	-1.295
4	open	115316_None_=2010.B99990096.pdb	-1.049
5	closed	115568_LPV_=2011.B99990066.pdb	-1.275
6	open	115568_LPV_=2011.B99990043.pdb	-1.040
7	closed	115568_None_=2008.B99990025.pdb	-1.199
8	open	115568_None_=2008.B99990062.pdb	-0.950
9	closed	115633_LPV_=2009.B99990097.pdb	-1.287
10	open	115633_LPV_=2009.B99990075.pdb	-1.026
11	closed	115633_None_=2007.B99990098.pdb	-1.407
12	open	115633_None_=2007.B99990099.pdb	-1.119
13	closed	115754_LPV_=2011.B99990058.pdb	-1.351
14	open	115754_LPV_=2011.B99990020.pdb	-1.116
15	closed	115754_None_=2010.B99990091.pdb	-1.351
16	open	115754_None_=2010.B99990032.pdb	-1.062
17	closed	115755_LPV_=2011.B99990003.pdb	-1.257
18	open	115755_LPV_=2011.B99990016.pdb	-0.992
19	closed	115755_None_=2007.B99990070.pdb	-1.289
20	open	115755_None_=2007.B99990063.pdb	-1.063
21	closed	116136_LPV_=2011.B99990026.pdb	-1.259
22	open	116136_LPV_=2011.B99990019.pdb	-1.021

23	closed	116136_None_=2009.B99990011.pdb	-1.284
24	open	116136_None_=2009.B99990001.pdb	-0.993
25	closed	116153_LPV_=2010.B99990051.pdb	-1.300
26	open	116153_LPV_=2010.B99990023.pdb	-1.050
27	closed	116329_LPV_=2009.B99990051.pdb	-1.232
28	open	116329_LPV_=2009.B99990055.pdb	-0.975
29	closed	116329_None_=2007.B99990095.pdb	-1.370
30	open	116329_None_=2007.B99990086.pdb	-1.114
31	closed	116499_LPV_=2011.B99990063.pdb	-1.212
32	open	116499_LPV_=2011.B99990027.pdb	-0.961
33	closed	116499_None_=2009.B99990081.pdb	-1.264
34	open	116499_None_=2009.B99990053.pdb	-1.016
35	closed	116574_LPV_=2011.B99990027.pdb	-1.427
36	open	116574_LPV_=2011.B99990090.pdb	-1.126
37	closed	116574_None_=2008.B99990058.pdb	-1.250
38	open	116574_None_=2008.B99990084.pdb	-0.984
39	closed	83052_LPV_<2010.B99990024.pdb*	-1.297
40	open	83052_LPV_<2010.B99990081.pdb*	-1.056
41	closed	83052_None_<2006.B99990044.pdb	-1.305
42	open	83052_None_<2006.B99990050.pdb	-1.056
A	open	3PWM (apo template)	-1.339
B	closed	3BC4 (apo template)	-1.424

Other quality metrics (DFire and QMEAN6), obtained from the SWISSMODEL are summarized in the figure below for the modeled receptors.

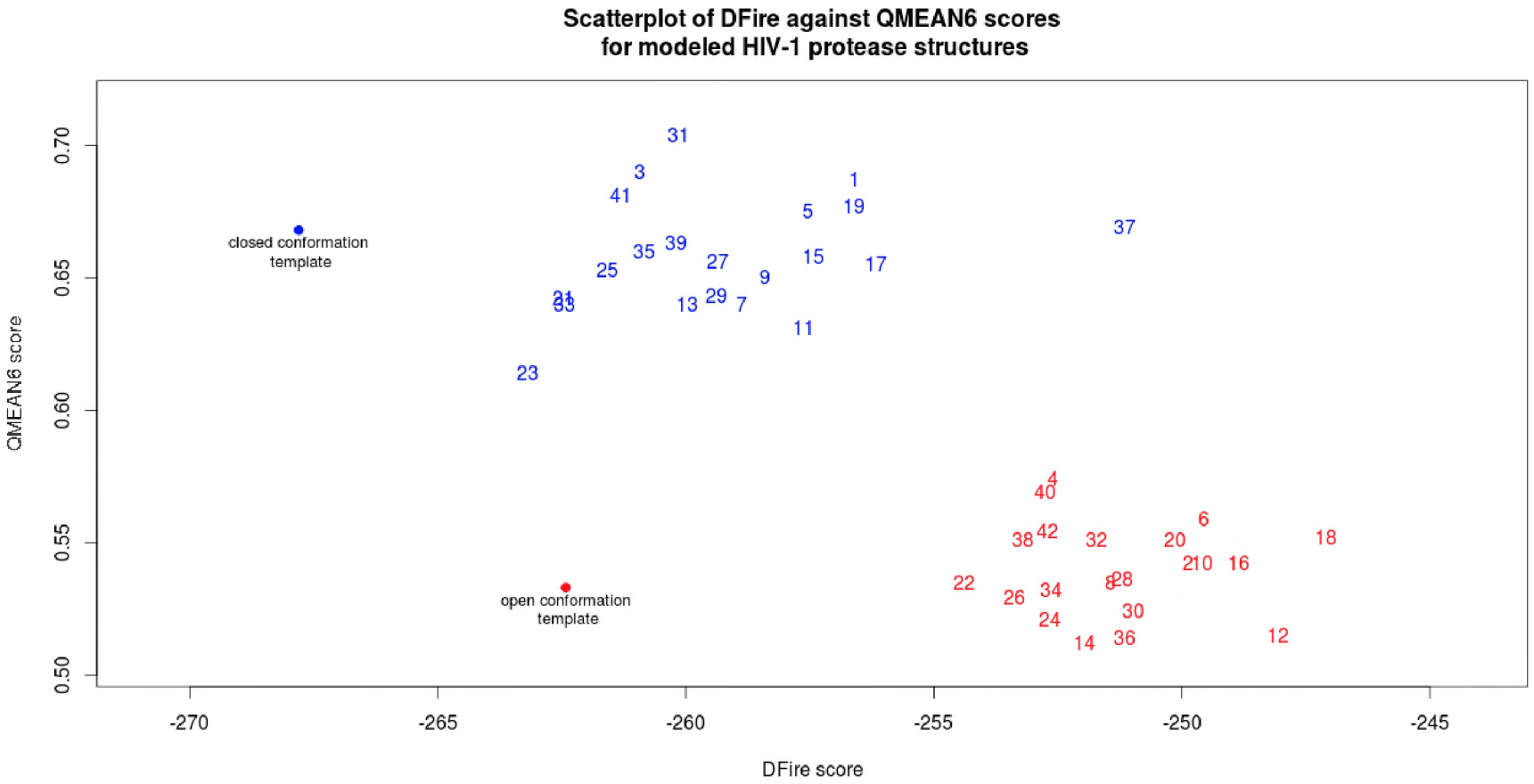


Figure 3.4: Model evaluation: DFire scores plotted against QMEAN6 scores, obtained from the SWISSMODEL server. Models are colored according to the template conformation initially used for modeling

Shown in the supplementary materials section 2 are the ANOLEA local scores for each of the modeled proteases in open and closed conformations. All models/ templates of closed conformations have odd number figures and model names while those of open conformations are even numbered. The models are good (green colored) on average but seem to have comparatively fewer unfavorable regions, according to both ANOLEA and QMEAN. Not immediately noticeable though are the consistent unfavorable regions for each of the even numbered models, around amino acid positions 51 (GLY 51 chain A) and 151 (GLY 51 chain B) , supported by both ANOLEA and QMEAN. It simply indicates that the open conformation template has an anomaly (probably because initial ligand present at the flap, which initially provided support) and all the models built from it are inheriting the same problem at the flap regions.

CONCLUSION

Good quality models were built for each of the selected protein sequences. However the general observed trend was that the closed conformation models were of better quality, having lower z-DOPE scores, better QMEAN6 and DFire global scores. ANOLEA and QMEAN seem to indicate an anomaly with the open conformation template and models that is most likely related to the absence of the flap-stabilizing ligand in the initial crystal structure,

CHAPTER 4: Molecular (receptor-ligand) docking

The current chapter introduces the concept of receptor-ligand docking and elaborates on one of the various available free energy scoring approaches – the one used by AutoDock 4 in this case. An independent docking validation tool (X-Score) is also introduced. A networking approach of visualizing and analyzing the docking results is also explained.

INTRODUCTION

Molecular docking is a widely used technique for predicting the binding mode between complexes composed of two or more molecules of known structures (Huang & Zou 2010). Ligand docking to a protein is one form of molecular docking and finds its importance in the discovery of candidate therapeutic drugs (Huang & Zou 2010).

The technique relies on the ability to predict a ligand binding mode and its affinity of binding to a receptor, based on computer-simulated conformations and the use of scoring schemes to rank and filter a defined space of conformations (also called poses) of lowest energies (Kuntz et al. 1982; Goodsell et al. 1996). The procedure, if done exhaustively, would involve the generation of thousands of possible binding conformations and the scoring (Kuntz et al. 1982) of each individual state for a single ligand and a single protein. It is easily conceivable that such the procedure easily escalates in computational complexity as the number and size of the ligand increases. For this reason, the docking procedure uses sampling to decrease the size of the search space to provide computational tractability, based on improvements to the work of Kuntz and his co-workers (1982) on the docking problem.

Sampling can be grouped into two activities, namely ligand sampling and protein flexibility (Huang & Zou 2010). Sampling entails the generation of possible ligand binding conformations at a binding site, while scoring gives a measure of binding tightness (suitable for ranking) based on physical or empirical energy functions (Huang & Zou 2010).

4.1. AutoDock and AutoDock Vina

AutoDock is an automated tool for predicting protein-ligand interactions (Morris et al. 2012). It aims at balancing robustness and accuracy of the interaction predictions, while keeping the computational demands reasonable (Morris et al. 2012). Two main features of AutoDock are its fast

grid-based energy scoring and its efficient search of torsion degrees of freedom (Morris et al. 2012). As at version 4.2 of AutoDock, the Lamarckian Genetic Algorithm (LGA) and an semi-empirical free energy-scoring algorithm are used – the approach being reproducible for up to 10 rotatable bonds (Morris et al. 2012). Alongside AutoDock, AutoDock Vina is developed in parallel, offering faster searches and reproducibility for ligands with over 20 rotatable bonds (Morris et al. 2012). AutoDock Vina uses an empirical scoring function like that of AutoDock 4, but differs in the local search function and in the way the scoring function is parameterized and optimized (Chang et al. 2010). AutoDock Vina's potential has only three terms, namely those for hydrophobic interactions, hydrogen bonding and torsion penalty (Chang et al. 2010).

The free energy scoring function of AutoDock 4 is evaluated over 3 interaction modes between a receptor and the provided ligand (protein-protein, ligand-ligand and protein-ligand) for both their bound and unbound conformations, yielding a total of 6 evaluations (Morris et al. 2012). The forcefield (energy scoring function) consists of 4 terms, as seen in the equation below,

$$V = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec}\sum_{i,j}\frac{q_i q_j}{e(r_{ij})r_{ij}} + W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{(-r_{ij}^2/2\sigma^2)}$$

Equation 4.1: Free energy-scoring function used by AutoDock 4.2 (Morris et al. 2012).

which comprises the potentials from Van der Waals interactions, hydrogen bonding, the Coulomb potential for atomic electrostatic interactions and the desolvation potential, which caters for the effect of solvent shielding of atoms (Morris et al. 2012).

The free energy scoring from AutoDock Vina is evaluated likewise (Trott & Olson 2009):

$$c = c_{inter} + c_{intra} \quad \text{, Equation 4.2.}$$

$$c = \sum_{i<j} \mathrm{f}_{t_i t_j}(r_{ij}) \quad \text{, Equation 4.3.}$$

From Equation 4.2. and Equation 4.3., it can be seen that the score "c" is computed as the summation of the inter- and intra-molecular interactions. Subscripts "i" and "j" relate to the ligand and receptor atom indices. "r_{ij}" is the internuclear distance. The symbols t_i and t_j refer to the atom types that are separated by 3 covalent bonds and can move relative to each other. The function "f" is the actual interaction calculation function, taking into account the different AutoDock Vina terms mentioned earlier. An optimization algorithm is also applied, whereby the global minimum of "c" and low-scoring conformations are determined and ranked (Trott & Olson 2009).

In AutoDock 4.2 and AutoDock Vina, both the target protein and the ligand need to have their polar hydrogens explicitly defined before proceeding with docking (Morris et al. 2008; Chang et al. 2010) as they both use united atoms types in their force field. United atoms are built by deleting non-polar hydrogens and merging their charges to the heavy atoms to which they were bound (Morris et al. 2012).

4.2. Receptor and ligand requirements for docking

Receptor preparation involves the addition of polar hydrogen atoms, the setting of partial charges, removal of atoms that do not form part of the receptor, such as those belonging to solvent molecules and ligands. All hydrogens have to be added for tools using all-atom forcefields, whereas only the polar hydrogens need to be specified for the those based on the united-atom forcefields, as is the cases with AutoDock and its AutoDock Vina. Atomic partial charges are calculated by the Gasteiger algorithm and the atom types are defined.

Ligand preparation entails the addition of Gasteiger charges, the assignment of rotatable bonds (the AutoDockTool's "prepare_ligand4.py" script sets the backbone as rotatable, while the amide bonds and guanidinium are set to non-rotatable by default. In addition to these, the ligand geometries should be optimized prior to docking.

4.3. Visualization

Several graphical and command-line interface tools are available for analyzing receptor-ligand interactions, following docking. These include, but are not limited to, tools such Discovery Studio (Dassault Systèmes BIOVIA 2015), which displays 3D images of the interactions and LigPlot+ (Wang et al. 2003), which renders 2D plots of the interactions. Yet another tool, the Protein-Ligand Interaction Profiler (PLIP) is available both as a web service and as a command-line executable, which not only produces text files containing the interaction information but can generate 3D images as well (Salentin et al. 2015) in a high throughput manner. The PLIP files are based on 7 atomic interaction types, including hydrogen bonding, hydrophobic interactions and salt bridges (Salentin et al. 2015) that can be found between the receptor residues and ligand atoms.

Other methods of visualizing the free energies of binding are possible due to the fact that a single value is produced, representing how favorable interactions are between a receptor and a ligand, which can easily be summarized by any of the common graphical data explorations tools available

from common statistical packages (such as R in this case) (R Core Team 2015) as bar charts, box plots in addition to heat maps.

The fact that PLIP gives the exact receptor residues and ligands involved in defined interaction types, means that the information can be interpreted as a relationship (an edge or a branch, in networking terminology). To build such a network, the output from PLIP can be parsed and converted to nodes (the residues and atoms) and edges. In this study, an edge represents a hydrogen bond shared between a protease residue atom and a ligand atom on a network graph. Several different network analysis methods are available, however an interesting one in this case would include the number of edges that a ligand atom has to a residue or vice versa – this is easily computed as the degree of a node. In network graphs, the edges may have directions (represented as arrows) or not (represented as lines) – in this case direction is not important and such a graph would be termed as an undirected graph.

4.4. Cross-validation of docking results

An independent validation tool such as X-Score (Wang et al. 2003) may be used to give additional confidence in the AutoDock-calculated energies. It uses three empirical energy-scoring functions that are each built using a variation of each of the different potentials, namely those of the Van der Waals interactions, hydrogen bonding, hydrophobic interactions and a weighted count of the number of rotatable single bonds (Wang et al. 2003; Wang 2003). The end result is simply the average of the three energies, however any combination of the three terms may also be used (Wang 2003). X-Score has been specially developed to evaluate and re-rank already docked complexes, and is as such not a docking tool (Wang et al. 2003).

METHODOLOGY

4.1. Receptor preparation

The selected models (termed receptors henceforth) were renamed and prepared using "prepare_receptor4.py" from AutoDock Tools to produce ".pdbqt" files. All of the receptors (including the open and closed conformation templates) were aligned in PyMOL (Schrödinger 2010) to an arbitrary receptor among the group of modeled receptors (using a Python script), so that a single center could be used for high throughput (HT) ligand docking. The center (Table 4.1) was

defined from the reference receptor using Discovery Studio, by taking the midpoint from the alpha carbons of the two catalytic aspartic acid residues located at position 25 of each chain.

Table 4.1: Determining the center of the ligand-binding site

Feature	Position
chain_A_Asp25 coordinate	(14.465, 16.297, 20.525)
chain_B_Asp25 coordinate	(12.801, 18.248, 26.446)
center	(13.633, 17.2725, 23.4855)

4.2. Ligand preparation

Optimized ligands (599 compounds) were retrieved from the SANCDB database. Only one compound could not be retrieved from SANCDB (Spongiostatin 5, SANCDB accession: SANC00213) as it was unavailable for download. These ligands were prepared by using the AutoDock Tools' (ADT) "prepare_ligand4.py" command, which produced "pdbqt" files.

On the other hand, all FDA-approved ligands were obtained from the Protein Data Bank with the exception of FPV that was retrieved from PubChem. A total of 10 ligands were retrieved in Structure Data File (sdf) format, each with an ideal coordinate instance from PDB except for FPV where same information was not available, but where ligand optimization had been already carried out based on the MMFF94 force field. The accessions for the ligands are given Table 4.2. Each structure file was manually edited (by trimming out atom alias annotations) before converting to PDB format by using OpenBabel (O'Boyle et al. 2011) and preparing them using ADT's "prepare_ligand4.py" script.

Table 4.2: Accessions of the FDA-approved PI's

FDA-approved PI	Accession
APV	478 (from PDB)
ATZ	DR7 (from PDB)
DRV	017 (from PDB)
FPV	131536 (from PubChem)

IDV	MK1 (from PDB)
LPV	AB1 (from PDB)
NFV	1UN (from PDB)
RTV	RIT (from PDB)
SQV	ROC (from PDB)
TPR	TPV (from PDB)

4.3. Docking

AutoDock Vina was used for the protein/ ligand docking. The docking simulations were submitted as jobs via a computer cluster at the Rhodes University Chemistry Department, on the priority queue with a limit of 1000 jobs at any given time. Control docking (positive control) experiments were carried out on the open and closed conformation templates prior to the HT docking with their original ligands. For the closed conformation template, the ligands originally found at the active site was removed using Discovery Studio and both the ligands and the apo proteins were prepared in the same fashion as previously described for the HT docking.

As mentioned in chapter 3 (section 3.1.1.), the open conformation protease structure was obtained as a biological assembly in which there was only half the receptor and ligand. The same ligand was found both at the flap and near the catalytic aspartic acid residues. Therefore, the other half of the ligands was built in Discovery Studio using one of the template's transformation matrices. The image thus obtained, required manual corrections – the ligands pair found at each of the flap and the catalytic regions were independently combined to produce two ligands. A C-C single bond had to be inserted between the then C18 and C2 atoms and redundant nitrogen atoms were deleted and the bonds reconnected. The original ligand (2-aminoethyl naphthalen-1-ylacetate) has the PDB accession LLG.

The exhaustiveness of search (in AutoDock Vina) was set to 4 for the purpose of screening the 599 SANCDB compounds and was supported by the control docking with the same docking parameters. The exhaustiveness was increased to 24 for higher docking accuracy against the 10 FDA-approved PI's due to their high number of flexible bonds.

4.4. Selecting the most energetically-favorable protein-ligand complexes

The optimal energy values (first conformation) produced from AutoDock Vina were first extracted from every (AutoDock Vina) log file for every docking experiment performed against both the FDA-approved and SANCDB compounds. Receptor/ ligand names were recorded, using a Python script. The information obtained from SANCDB ligand poses was then filtered to give a list containing the most energetically-favorable receptor-ligand complexes (refer to Table 4.1) by building a Python dictionary object that would overwrite a previous object's (the complex) key if the docking energy of the object is lower (better) than the previous one. Docking results obtained for the FDA-approved compounds were not filtered, but are all represented in Figures 4.5 and 4.6 as heat maps for the open and closed conformation receptors, respectively.

4.5. Docking cross-validation

All docking output from the previously shortlisted complexes from both SANCDB and FDA-approved compounds were re-evaluated by X-Score. A Python script was used as a wrapper for the X-Score tool for HT evaluations, taking as input a file containing the previously generated list of receptors and ligands, with the lowest binding energies for each patient and protease conformation. The same script also encapsulated the Open Babel tool, which was used to convert the ligands from "pdb" to "mol2" format as required by X-Score.

4.6. Analysis of receptor-ligand interaction using PLIP

PLIP (version 1.2.) was installed and run locally so that HT runs could be performed on all the energetically-favorable receptor-ligand complexes. The text output was parsed and converted to lists of network edges (edge lists) – one for the SANCDB compounds and another one for all of the FDA-approved compounds. The edge lists were de-identified from patient ID, patient treatment and receptor conformation information, to treat all of the receptor models as a single protease dimer entity representative of all the whole batches analyzed. The main focus of the network graph was targeted towards the SANCDB, however others were also built for the FDA-approved drugs to give an overview of the larger number of processed complexes. In all graphs, edges that were represented more than once were combined and shown as weighted edges (thicker edge widths representing more connections), decreasing the density of information displayed on the graph. Another dimension was also added to the network, namely the number of different edges emanating

from any given node would alter the node size – bigger nodes would represent those that have more (different) edges.

4.7. Adaptation of the AutoDock 4.2 potential for energy profiling

A Python script was written to encapsulate part of the free energy scoring functions used by AutoDock 4.2. Out of the six energy calculations that are normally evaluated by AutoDock 4.2 for inter- and intra-molecular interactions, only the one for bound intermolecular interactions between ligand and receptor was chosen. The potential was further simplified by including only the terms for Van der Waals, hydrogen bonding and electrostatics. The equations for the calculations are given in section 4.2. Cut-offs were implemented for the Van der Waals and hydrogen bonding (9 and 2.8 Angstroms respectively) to speed up calculations. Coefficients for the calculations were obtained from the AutoDock web page on AD4 parameters (http://autodock.scripps.edu/resources/parameters/AD4_parameters.dat/view) and the AutoDock 4.2 manual (Morris et al. 2012). Other coefficients for the calculations of A_{ij}, B_{ij}, C_{ij} and D_{ij} for the Lennard Jones (12,6 and 12,10) potentials were obtained from a paper published by Morris and his co-workers (1996). Energy evaluations were calculated for each ligand atom against every residue along the two protease chains. They were only computed on the DRV/ protease complexes. Multiple plots were generated (using an R script) onto which the energy values between each protease residue and each ligand atom were plotted.

RESULTS AND DISCUSSION

Docking controls are depicted in Figure 4.1. The open and closed conformation crystal structures are shown with the original ligand placement on the left and their re-docked conformations are shown on the right.

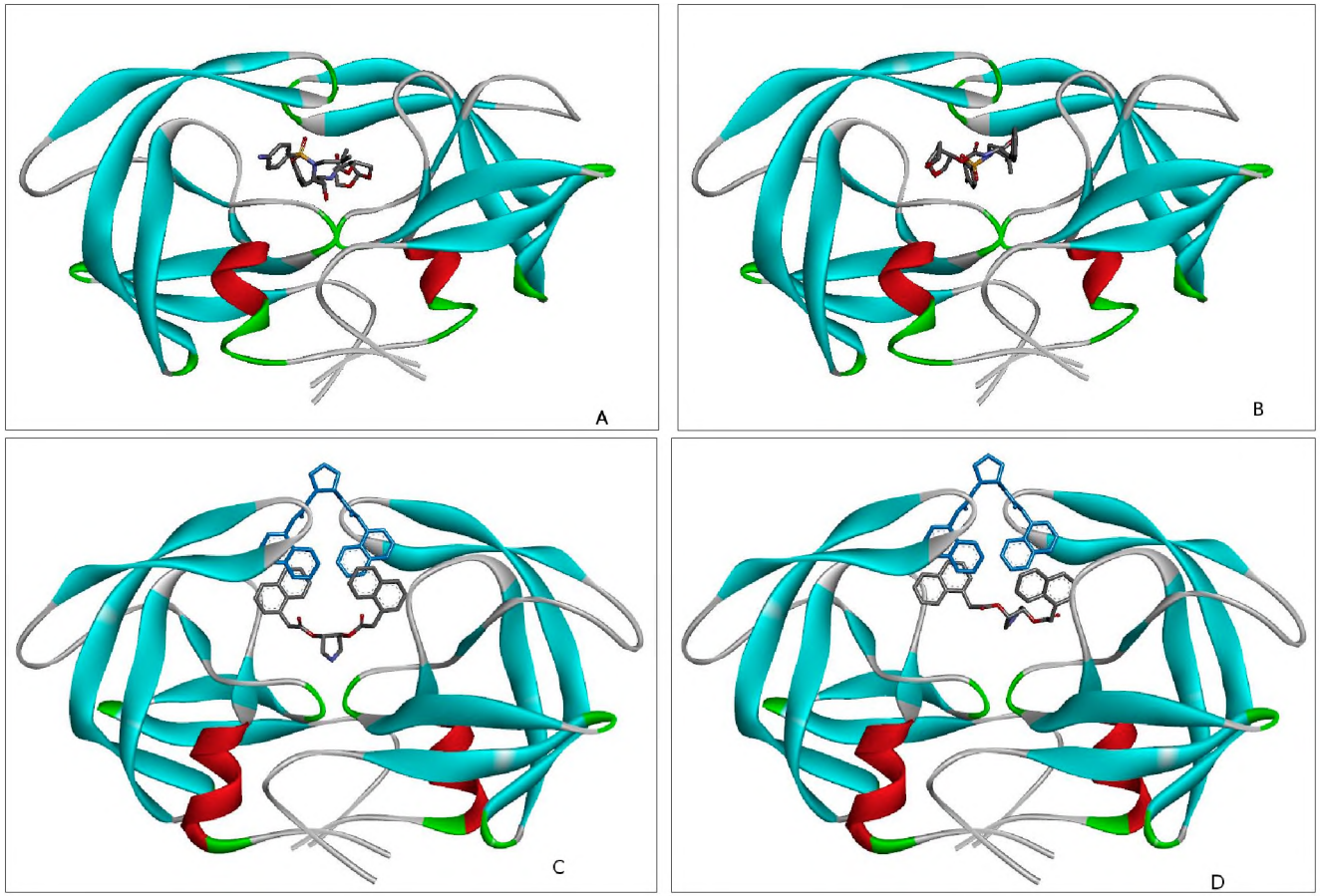


Figure 4.1: Control docking: Ligand poses before and after re-docking (A) Closed receptor conformation with original ligand, (B) Closed receptor conformation with re-docked ligand, (C) Open receptor conformation with original ligand and (D) Open receptor conformation with re-docked ligand

The re-docking for the closed conformation template resulted in a pose that is doubly rotated across one plane, but still binds at the active site with a low free energy value. On the other hand, the open conformation template re-docking was problematic at first when the all the ligands were removed from the re-built structure. Due to the high flexibility of the built ligand, the two "arms", each consisting of the fused aromatic rings were free to move and gave higher free energies (AutoDock Vina) of binding initially (about -9kcal/mol). Subsequently, the approach was altered only to put back the initial constraint imparted by the ligand originally found blocking the flap region, before reattempting docking. Indeed, the presence of the ligand at the flap did reduce the mobility of the two arms, most likely due to pi-stacking interactions between the ligand aromatic rings. The resulting energy lowered to -10.8kcal/mol (AutoDock Vina).

The fact that the ligands had favorable binding free energies to the receptors meant that the docking parameters could be used for HT ligand screening. The energies for best scoring receptor-ligand complexes from the screen against SANCDB compounds are shown in Table 4.1, including those for the re-docked ligands.

Table 4.3: Lowest binding energies for the receptor/ligand complexes

Receptor/ligand complexes	Binding energies determined by AutoDock Vina (kcal/mol)	Cross-validation of binding energies, using X-Score (kcal/mol)
A (open) / re-docked original ligand	-10.8	-8.87
B (closed) / re-docked original ligand	-10.6	-10.14
1_apo/SANC00174	-11.1	-11.11
2_apo/SANC00488	-10.1	-11.16
3_apo/SANC00290	-11.3	-9.85
4_apo/SANC00400	-10.1	-10.54
5_apo/SANC00594	-11.0	-10.86
6_apo/SANC00381	-10.1	-9.09
7_apo/SANC00347	-11.9	-10.16
8_apo/SANC00585	-10.0	-9.92

Receptor/ligand complexes	Binding energies determined by AutoDock Vina (kcal/mol)	Cross-validation of binding energies, using X-Score (kcal/mol)
9_apo/SANC00347	-11.2	-10.06
10_apo/SANC00518	-10.4	-11.21
11_apo/SANC00290	-11.5	-9.91
12_apo/SANC00595	-10.1	-10.19
13_apo/SANC00290	-10.9	-9.61
14_apo/SANC00381	-10.5	-9.28
15_apo/SANC00347	-11.5	-9.9
16_apo/SANC00381	-9.9	-9.05
17_apo/SANC00290	-11.0	-9.78
18_apo/SANC00381	-10.4	-9.3
19_apo/SANC00290	-11.2	-9.69
20_apo/SANC00381	-10.4	-9.16
21_apo/SANC00347	-11.3	-9.83
22_apo/SANC00518	-10.0	-11.2
23_apo/SANC00175	-10.8	-11.24
24_apo/SANC00383	-10.7	-9.38
25_apo/SANC00347	-11.3	-10.42
26_apo/SANC00518	-10.1	-11.15
27_apo/SANC00594	-11.9	-11.01
28_apo/SANC00380	-10.3	-9.52
29_apo/SANC00421	-11.1	-10.84
30_apo/SANC00381	-10.3	-9.24
31_apo/SANC00290	-11.2	-9.77

Receptor/ligand complexes	Binding energies determined by AutoDock Vina (kcal/mol)	Cross-validation of binding energies, using X-Score (kcal/mol)
32_apo/SANC00381	-10.8	-9.46
33_apo/SANC00347	-11.4	-10.46
34_apo/SANC00264	-9.8	-9.36
35_apo/SANC00422	-11.2	-11.04
36_apo/SANC00585	-9.9	-9.91
37_apo/SANC00290	-11.1	-9.89
38_apo/SANC00386	-10.1	-9.06
39_apo/SANC00585	-12.8	-11.57
40_apo/SANC00386	-10.1	-9.26
41_apo/SANC00342	-11.4	-10.36
42_apo/SANC00685	-9.8	-10.27

After computing the X-Scores, it was of interest to compare the results for AutoDock Vina and X-Score. To do so, the AutoDock Vina binding energies were plotted against those of X-Score for SANCDB compounds (Figure 4.2). In the figure, the most favorable binding AutoDock Vina energies are found at the left along the x axis, while those from X-Score are found towards the bottom of the y axis. Being generated from three different free energy calculations, there should normally be a higher confidence in its value – this is actually the case as there is no occurrence of energy values on the top left corner of the scatter plot, as this would mean that X-Score is giving lower (better) scores to high-scoring (worse) complexes from AutoDock Vina. X-score values were used in further analyses.

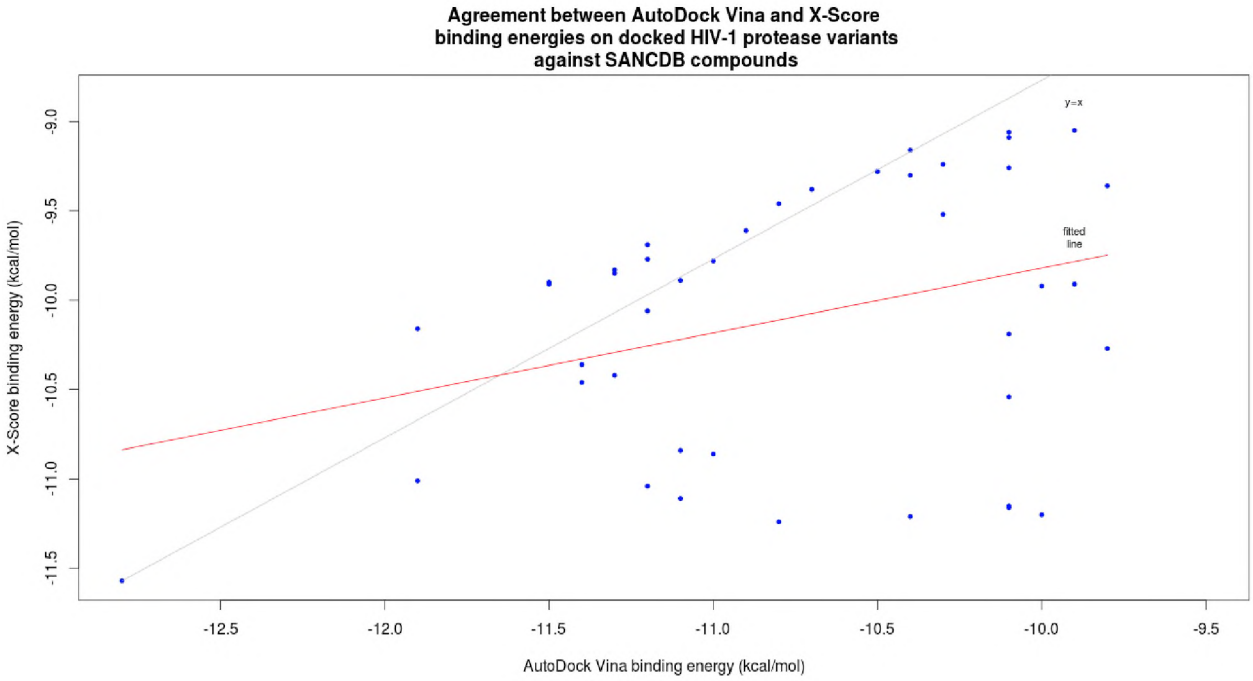


Figure 4.2: Scatter plot of lowest binding energies obtained from AutoDock Vina and X-Score for 42 docked SANCDB compounds. The regression line is shown in red. The line (y = x) is shown in gray, for comparison with the regression line.

The data from the previous table (Table 4.1) was mined further to display more of the available information. Bar plots were used for that purpose (Figure 4.3) to summarize the binding free energies across all the patient receptor-ligand complexes, separately for the open and closed conformations. Circled in red are the most frequently-occurring ligands that have been found to dock to the different variants of the HIV proteases, namely ligands 381 (SANCDB accession: SANC00381) and ligand 290 (SANCDB accession: SANC00290). Not highlighted in the figure, is another promising ligand 347 (SANCDB accession: SANC00347) that was the second most frequent ligand to bind closed conformation protease variants. The frequencies of occurrence of these three ligands are tabulated in Table 4.2, where the values have been computed context-wise, in other words, the total frequencies were based on the category taken into consideration (open and closed conformations).

Further interpretations can be made from the same figure, however with caution, as not all of the lowest binging energy ligands that were shortlisted are the same before and after treatment. It could simply mean that the ligand torsional space was not searched with enough exhaustiveness, or on the other hand, that the energy of binding became less favorable such that other ligands were picked up

with lowest energies. For those complexes, where the same ligands were retrieved before and after treatment, similar, or slightly higher binding energies were observed after treatment, namely in patients 115754 and 115755 for the open conformation models, while same was observed in patient 115755 among the closed conformation models, which may potentially indicate a marginal degree of resistance to LPV, however higher confidence would be obtained by increasing the conformational search space (exhaustiveness).

Figure 4.3: Summary of binding (X-Score) energies for open and closed conformation receptor models for each patient, before and after LPV treatment. The docking controls are also shown for both the closed and open conformation templates.

Table 4.4: Criteria and specifics from the SANCDB screening for compounds with potential protease inhibitory activity

SANCDB ID	Identity	Source	Additional information from SANCDB
Highest frequencies across closed conformation HIV protease variants			
SANC00290 (frequency: 7/22)	Clionamine D	*Cliona celata*	Aminosteroid (Cholane-21,24-Dioic Acid, 3-Amino-16,20-Dihydroxy-, 16,21:20,24-Dilactone, (3[Beta],16[Beta],20S), used as an autophagy modulator.
SANC00347 (frequency: 6/22)	Kraussianone 4	*Eriosema kraussianum*	Pyrano-isoflavone, with no recorded use.
Highest frequency across open conformation HIV protease variants			
SANC00381 (frequency: 7/22)	Cissacapine	*Cissampelos capensis*	Alkaloid (Bisbenzyltetrahydro-isoquinoline), with no recorded use.
Lowest energy of binding			
SANC00585	Scutiaquinone B	*Scutia myrtina*	Anthelmintic activity

Different selection criteria lead to the shortlisting of 4 compounds from the SANCDB screen with potential PI activity, as shown in Table 4.2. Details of ligand binding to their respective receptors are shown in Figure 4.4.

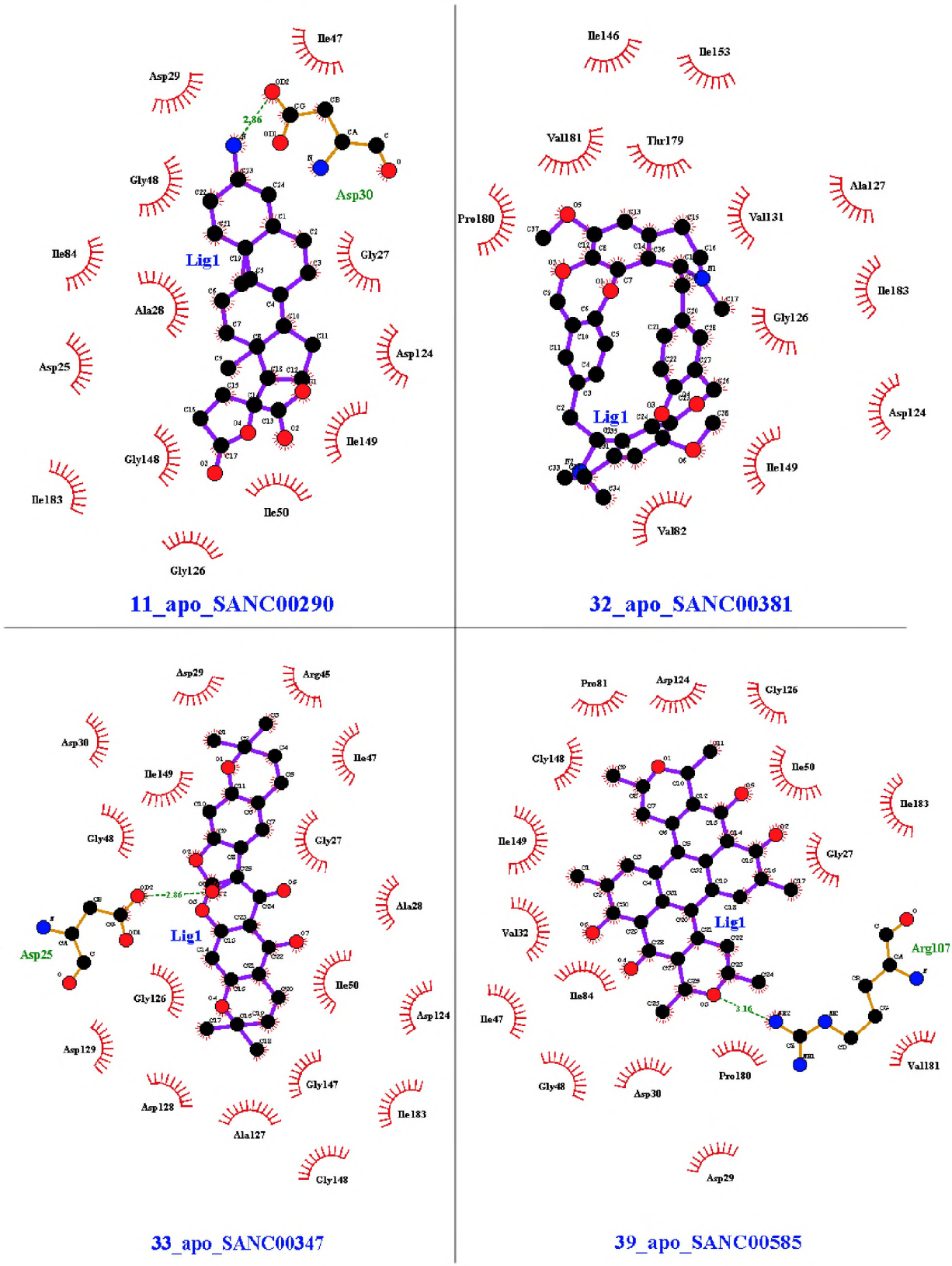


Figure 4.4: Residue-ligand interactions for selected SANCDB compounds determined by LigPlot+

From Figure 4.4, it can be observed that hydrophobic interactions (described as alkyl, pi-alkyl and pi-pi in Discovery Studio) play a major role in stabilizing the ligands inside the binding pocket of each of the proteases. Few hydrogen bonds are observed, namely between ASP30 (11_apo) and SANC00290, ASP25 (33_apo) and SANC00347 and another one between ARG107 (39_apo) and SANC00585.

Heat maps were chosen to represent the binding energies for the FDA-approved drugs due to the high number of docking experiments that were performed – (10 drugs x 2 conformations x 11 patients x 2 drug statuses – 2 redundant conformations for a patient). Figure 4.5 displays the docking energies obtained for open conformation protease models while those for closed conformation models are shown in Figure 4.6. In the case of the open conformations, it can be seen that all the PI's are performing well due to their lower binding energies, with the exception of Amprenavir, Darunavir and Fosamprenavir, with the latter performing the worst, most likely due to the fact that it is an Amprenavir precursor, requiring chemical modifications for properly fitting and blocking the protease active site. Saquinavir on the other hand is found to perform better than all on average in the open conformations.

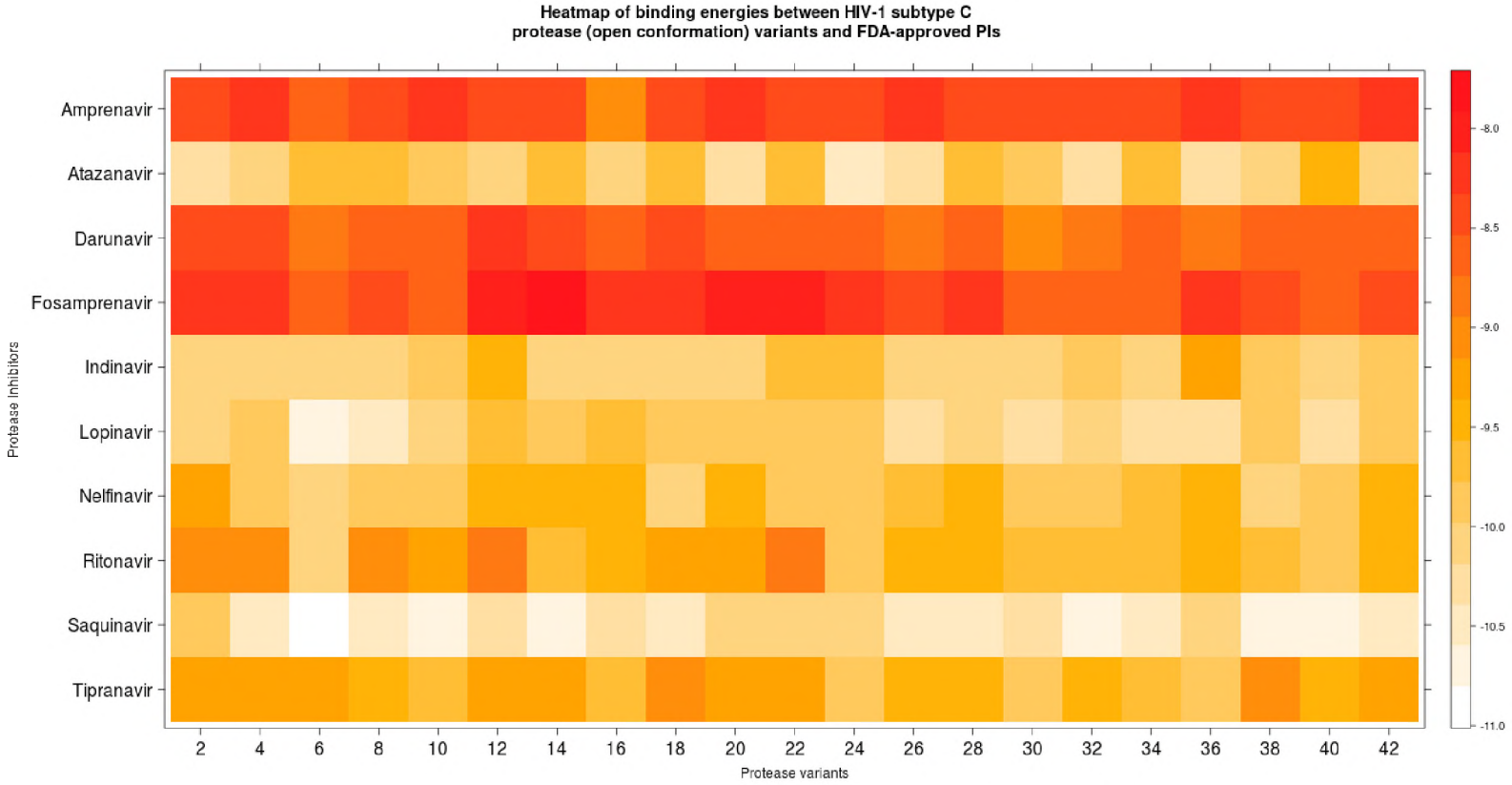


Figure 4.5: Heat map for docking energies obtained for open conformation protease models

The same trend is observed with the closed conformations (Figure. 4.6) as previously seen with the open conformation models, i.e. amprenavir, darunavir and fosamprenavir are performing worse than the other PI's, with fosamprenavir having the least favorable free energies of binding. Again, saquinavir is observed to perform better than the other PI's on average.

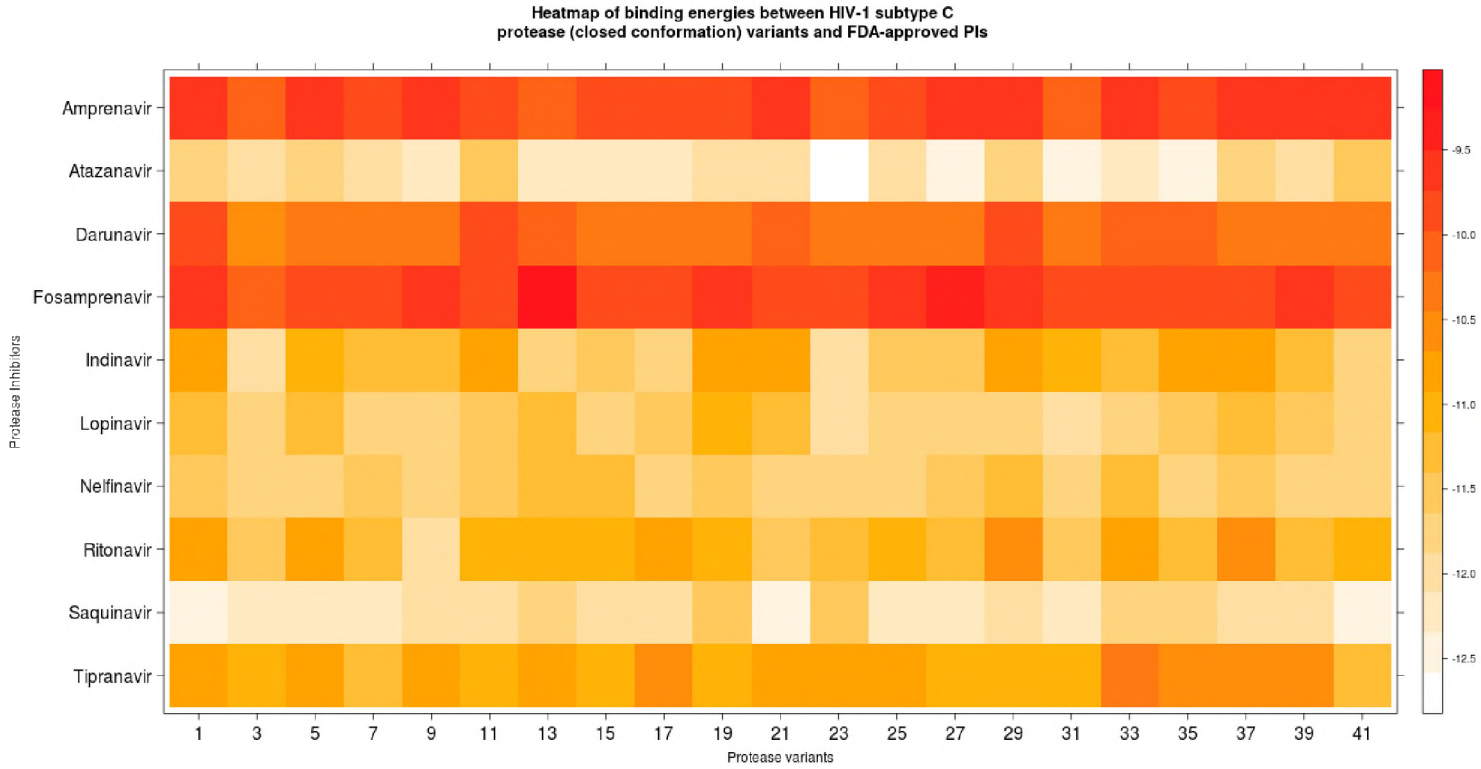


Figure 4.6: Heat map for docking energies obtained for closed conformation protease models.

Subsequently, network graphs were plotted in an attempt to represent the interactions occurring between the receptor-ligand complexes in the SANCDB compounds and to evaluate the performance of the same compounds in the context of these HIV-1 subtype C protease complexes. From Figure 4.7, it can be observed that the residues ASP124 (ASP25, chain B), ASP128 (ASP29, chain B), ARG107 (ARG8, chain B), ASP129 (ASP30, chain B), ASP29, ASP30, THR179, ILE50 and ILE149 (ILE50, chain B) were shared amongst different SANCDB ligands, meaning that several SANCDB compounds were consistently interacting with same protease residues at the active site despite their structural differences. In other words, ligands having hydrogen donors or acceptors at any of these positions in space would likely form a hydrogen bond with the mentioned protease residues. Viewed from a different angle, ligand atoms from compounds SANC00347 (oxygen 30), SANC00518 (oxygen 38), SANC00585 (oxygen 26), SANC00290 (oxygen 19; nitrogen 28) and SANC00380 (oxygen 40) are found to be connected to multiple different protease residues located at different positions. These ligands potentially have the advantage of having multiple ways of interacting with and blocking the protease active site, for more than one conformation (or variant) of the protease and would confer resilience to these mutations should they occur. In addition, oxygen 12 from ligand SANC00290 was found to bind to ILE50 multiple times, as seen by the thicker edge.

The network visualization approach was initially designed to summarize the hydrogen bonding between the SANCDB/ HIV protease complexes only. However, due to its ability to summarize and represent moderately dense interaction information, the same approach was applied to the complexes involving the FDA-approved ligands in figures 4.8-4.16. In the same figures, another level of information was more apparent, given the higher density of the information used for the FDA-approved drugs. Several network edges were clearly much bigger than the others, showing the number of times that the same hydrogen bonding interaction was found between the same ligand atom and protease residue. Note that atom names mentioned in the paragraphs below correspond to ligand atoms. Also, a stronger connection in this case only refers to the number of repeated edges found between two nodes, and does not directly refer to binding affinity. The degrees of the nodes are masked by combining repetitive edges in all of the built networks and it might not suitable to refer to the absolute degrees of the nodes to infer about node (atom/ residue) importance.

For APV (Figure 4.8), it can be seen that nitrogen 34, 37 and oxygen 16, 27 had many different residues to interact with. Several strong edges were observed between oxygen 37 and both catalytic ASP's. Other strong connections were observed between ASP30 and nitrogen 34, and between ARG107 and oxygen 27.

ATZ (Figure 4.9) did not have exceptionally large nodes, but had multiple residues connected to nitrogen 4, 9, 28, 37, 55 and oxygen 11, 13, 26, 30, 41. Several edges with strong connections were also observed within its complexes.

A different network was obtained with FPV (Figure 4.10) compared to AMP, with some similarities (e.g. a highly connected nitrogen atom), however the information should be treated with caution as the precursor requires chemical modification before it can function.

For IDV in Figure 4.11, it can be observed that nitrogen 10, 19, 36 and oxygen 9, 23, 35, 47 interact with multiple protease residues. Several strong connections were observed, for example between oxygen 25 and both catalytic ASP's..

LPV (Figure 4.12) comprises mostly of strongly-connected edges with the HIV proteases, some notable examples including those between ASP29 and nitrogen 11, and GLY27 and nitrogen 2.

NFV (Figure 4.13) has a very strong connection between ASP30 and oxygen 32. NFV oxygen atoms 15, 23, 32, 42 and nitrogen 13, 21 atoms bind to several different protease residues. Few strong connections are found, but a very strong one is observed between oxygen 32 and ASP30.

In the case of RTV (Figure 4.14), oxygen 6, 8, 22, 33, 39 and nitrogen 4, 12, 31, 36 all show multiple connection to different residues and at the same time several of them are strong connections towards different protease residues. Several strong connections are observed, examples of which are the ones between oxygen 22 and ILE149, and between nitrogen 36 and ARG8.

Nitrogen 45 of SQV (Figure 4.15) has a very central atom in the network having multiple connections to different protease residues, in addition to a very strong connection to ASP30. Also of importance are the nitrogen 17, 23, 28, 33, 45 and oxygen 2, 16 that have multiple connections to different protease residues. Several moderately strong connections were observed, with a very strong one between ASP30 and nitrogen 45. The high connectivity of the nitrogen 45 most probably explains the overall good performance of the drug in the previous heat maps (Figures 17 and 18).

In Figure 4.16 (TPR), multiple connections with protease different protease residues are observed between oxygen atoms 3, 4, 19, 33 and nitrogen atoms 29, 32. Some strong connections were observed, such as the one between ASP29 and nitrogen 29.

The information presented in the networks can potentially be used to extract information about important atomic and resilient architectures that may be of relevance in PI design.

Figure 4.9: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-atazanavir complexes

Figure 4.10: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-fosamprenavir complexes

Figure 4.11: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-indinavir complexes

Figure 4.12: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-lopinavir complexes

Figure 4.13: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-nelfinavir complexes

Figure 4.14: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-ritonavir complexes

Figure 4.15: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-saquinavir complexes

Figure 4.16: Network graph showing the protease/ ligand interactions determined by PLIP for 42 protease-tipranavir complexes

Hydrogen bonds could not be detected by PLIP in the case of DRV, despite increasing the distance threshold between the hydrogen acceptor (HA) and donor (HD) atoms and decreasing the threshold angle at the hydrogen bond bond donor. Same complexes were independently verified for non-covalent atomic interactions in both Discovery Studio and LigPlot+ (not shown) and they did show the existence of hydrogen bonds with suitable bond angles and distances between HA and HD atoms, indicating that PLIP might not be performing correctly on DRV. Instead of showing these plots, an energy profiling approach has been explored, to further mine the information available from the protease/ DRV complexes (Figures 4.17-4.18).

The energy semi-empirical scoring function used is only based on hydrogen bonding, Van der Waals and electrostatics interactions and is therefore only giving an estimate of the actual binding energy between each ligand atom and each protease residue. In the profile graphs, negative binding energies are colored from red (best) to yellow, positive binding energies are colored from green to blue (worst) and those with energies close to zero (zero was approximated by values ranging from -9e-4 to 9e-4 Kcal/mol) are colored white. Each of the profiles is showing variations of about eight zones of favorable binding energies all along the two protease chains (laid out from residue 1 to 198). Tick marks and individual axis labels have been removed from figures 29-31 for clarity and thus allow only qualitative assessment. In any given profile these zones represent the 3D binding pocket laid of in 2D and clearly shows the distance separation between each contacting residue for any given atom, or group of atoms on the y axis. Another information gathered from the DRV profiles is that of symmetry – the protease chains being laid out head to tail from N- to C-terminus can be found to be an imperfect translation of the profile image from the first chain to the second chain. In other words, similar positions along the protease chains are recruited for binding but the energy distribution is not the same, such as can be seen in profile 1_apo in Figure 29. In the same figures (Figures 29-31), unfavorable binding energies (of varying degree) can be seen for almost every profile calculated, with the most prominent one coming from the 3_apo/ DRV complex, close to the C terminus of second chain (chain B). Same is not observed on the first chain however, meaning that the energy profile of this complex is not symmetric.

The general trend observed in all the energy profiles is that the contacting residues from the first chain are repeating over to similar regions along the second chain, but are not symmetrical, such as can be seen in the 12_apo and 20_apo profiles amongst others. An important characteristic of the profiling here is that instead of displaying a global energy value, local values are shown, giving finer grain detail of what is happening for any particular snapshot of a docked ligand inside a receptor.

Magnified graphs (Figures 32 and 33) have been included for better visibility. Figure 32 is showing the profile for complex 3_apo/ DRV picked from the series of complexed DRV profiles. From this graph, it can be seen that a group of atoms around ligand atom 20 (boxed in red) are interacting with multiple residues inside the active site, some of which are yielding unfavorable energies such as is the case for residues 16 from (chain A) and 174 (residue 75 chain B). One of the complexes selected from the SANCDB screen (39_apo/ SANC00585) was also energy-profiled. From the figure, similar protease contacting residues are found, however asymmetry is observed (differences are circled in red) at the N-termini of the two protease chains. Several unfavorable binding energies are also revealed in the same plot against similar residues from both chains A and B of the protease. Modification of these bad-scoring atoms from the ligands may improve the binding at these local areas and lead to a better potential protease-inhibiting drug.

Figure 4.18: Details of complex 3_apo/ Darunavir docking free energy profiling

Figure 4.19: Details of complex 39_apo/ SANC00585 docking free energy profiling

CONCLUSION

Docking controls for both open and closed receptor conformations validated the subsequent docking experiments using AutoDock Vina. Placing the flap ligand was beneficial for the success of the docking against the open conformation template, due to the high flexibility of the re-constructed ligand. Good binding energies were obtained from AutoDock Vina and agreement was obtained from X-Score cross-validation.

Resulting from the SANCDB ligand screen, same ligand was not found to bind before and after LPV treatment, indicating possible reduction in binding affinity of the protease variants or a lack of exhaustiveness in searching ligand conformations due to the lower (but faster) exhaustiveness.

Two FDA-approved drugs, namely APV and DRV (disregarding the APV precursor FPV) appeared to perform consistently worse in all open and closed receptor conformations. SQV, on the other hand performed better overall against both receptor conformations.

A powerful network representation method was developed, representing moderately high density hydrogen bonding information between numerous (HIV protease) receptor residues and their docked ligand atoms. A dock profiling approach was also developed, showing the results of local binding free energy evaluations between HIV proteases and atoms from corresponding docked ligands.

CHAPTER 5: Energy minimization and molecular dynamics

This chapter introduces concepts that lead to the simulation of motion (molecular dynamics) of previously-obtained HIV-1 protease-ligand complexes in the presence of an environment controlled to mimic physiological conditions of the human body. A freely-available tool – GROMACS (van der Spoel et al. 2010) is used to perform the MD simulations and details of the steps used are given.

INTRODUCTION

Energy minimization is usually used in an attempt to correct for structural inconsistencies between atoms (such as steric clashes and strains) introduced during the modeling step (Xiong 2006). This is accomplished by displacing atoms in such a way that the overall potential energy of a given model is minimized, without significant modifications to the original structure (Xiong 2006). However, this step is to be used with caution as excessive minimization can move residues from their correct positions – therefore it is advisable to run only a few hundred iterations to correct for the most critical errors such as short bond lengths and close atomic clashes (Xiong 2006). For similar reasons, important residues such as those involved in cofactor binding may need restraining (Xiong 2006). Energy minimization is normally carried out prior to MD simulations.

MD is an improvement over minimization, the rationale of which is that the latter may land on a local optimum instead of searching through a wider space of conformations to find more optimal structures (Xiong 2006). MD widens this search space by modulating heating and cooling to simulate molecular motions uphill and downhill an energy landscape (Xiong 2006). MD typically solves Newton's equations of motion for a system of N interacting atoms, updating atomic coordinates of the system over a defined number of time steps, given a suitable temperature and pressure (Abraham et al. 2015; Cickovski et al. 2010; Frenkel & Smit 2001). From the simulation, it is expected that modeling and monitoring microscopic changes will predict changes that occur at the macroscopic level (Somer 2004). The method is very similar to real experiments in many respects, as it involves sample preparation and measurement of a given property over a set amount of time (Frenkel & Smit 2001). In the same manner, problems affecting real-life experiments also affect MD, such as the introduction of noise during measurement (that can be dampened by repeated measures) (Frenkel & Smit 2001). Sample preparation in this case involves the equilibration of the system of particles, whereby Newton's equations of motion are solved until the system stops changing (Frenkel & Smit 2001). Measurements are taken from observables that are generated by a function, which takes as input particle position and momentum for each of the

system's particles (Frenkel & Smit 2001). In GROMACS, the equilibration process involves the coupling of any, or both, of a thermostat and/ or a barostat for maintaining temperature and pressure, respectively (Abraham et al. 2015).

MD runs may include water and ions for more realistic simulations (Xiong 2006). It should nevertheless be noted that given the simulations are approximations using forcefields/ classical mechanics, they do not treat features of quantum mechanical nature and hence are inherently limited (Abraham et al. 2015).

5.1. Force fields and the potential energy functions

The potential energy functions are similar to those generally used in molecular mechanics, adding the local (bond length, angle and dihedrals) and non-local (electrostatic and Lenard Jones) atomic interactions (Rapaport 2004), the latter being similar to the one previously described for AutoDock (in chapter 3). These force field vary depending on their set of equations and parameters used for their calibration (Monticelli & Tieleman 2013), the latter of which are determined empirically or estimated by quantum mechanical methods (Ponder & Case 2003). Forces are determined simultaneously in small time increments (Δt) from the negative derivatives of the energy-scoring functions (Abraham et al. 2015). Atomic positions are updated and recorded at the set time intervals for the duration of the simulation to produce atomic trajectories (Abraham et al. 2015).

5.2. Kinetic energy and temperature

After minimization, the net force in the system (including solvent and ions) is as brought as small as possible (using algorithms such as the steepest descent or the conjugate gradient amongst others) (van der Spoel et al. 2010). One equation used to relate the average kinetic energy and temperature is given by: $\langle \frac{1}{2} m v_{alpha}^2 \rangle = \frac{1}{2}(K_B T)$, where K_B is the Boltzmann constant (Frenkel & Smit 2001).

To simulate atomic motion, these atoms are energized by randomly assigning starting velocities to the individual atoms from the Boltzmann distribution (Klebe 2013a) generated at the constant temperature defined for the system (NVT). The time for equilibration is used to then solve for the classical equations of motion (Klebe 2013a).

5.3. Pressure

NPT equilibration follows NVT equilibration, starting with energies and velocities generated from the NVT equilibration (van der Spoel et al. 2010). Pressure is kept constant during the NPT equilibration using a barostat, such as that of Berendsen, Parrinello-Rahman and Martyna-Tuckerman-Tobias-Klein, which are available from the GROMACS package (Berendsen et al. 1984; Martyna et al. 1996; Nosé & Klein 1983; van der Spoel et al. 2010).

In GROMACS, groups of atoms are independently defined for both NVT and NPT by coupling them for temperature and pressure control, respectively (van der Spoel et al. 2010).

5.4. Restraints and constraints

The purpose of a force restraint is to minimize movement of a given molecule during the steps of equilibration by taking into account position, distance and dihedral angles (Apostolov 2014). Deviations are possible but are heavily penalized such that the integrity of a desired component in the system is preserved (Abraham 2011; van der Spoel et al. 2010). Such restraining is warranted, as for example in the case of a receptor-ligand complex, both molecules should not be expected to move much during the initial set up of the reaction conditions before starting the experiment, which is the production dynamics.

Constraints on the other hand are used to fix bond lengths or bond angles after force integrations (for potential energy calculation) (Abraham 2011). Two constraint methods available for GROMACS are the LINCS and the SHAKE algorithms, the former being faster and more stable than the latter (van der Spoel et al. 2010).

5.5. GROMACS run Parameters

Parameters required for the MD simulation have to be specified inside ".mdp" files. Default parameters are available, but can be altered, depending on the ensemble to be simulated. Below is a list of some relevant parameters:

Table 5.1: MD parameters (Adapted from GROMACS manual and default ".mdp" files)

Command	Modifier	Purpose
define	-DPOSRES	

integrator	md	Specify leap-frog algorithm for calculating integrations.
	steep	Implement steepest descent algorithm for minimization.
emstep		Step size for energy output.
emtol		Force threshold (KJ/mol/nm) below which system is minimized.
dt		Time increment (ps) for integration calculation.
nsteps		Maximum number of steps to use for integration or minimization.
nstxout		Number of steps before writing coordinates to trajectory file.
nstvout		Number of steps before saving velocities to trajectory file.
nstlog		Number of steps before updating log file.
nstenergy		Number of steps before saving energies to energy file.
energygrps		Group(s) to save to energy file.
nstlist		Number of steps before updating neighbor list.
pbc	xyz	Specify periodic boundary conditions in all directions.
rlist		Threshold distance (nm) for the short-range neighbor list.
coulombtype	PME	Use Particle Mesh Ewald for long range electrostatics.
rcoulomb		Cut-off distance (nm) for short range electrostatics.
rvdw		Cut-off distance (nm) for short range Van der Waals interactions.
tcoupl	v-rescale	Temperature coupling using velocity-rescaling, with a random term.
tc-grps		Groups to be used for temperature coupling
tau-t		Time constant (ps) for temperature coupling (1 per group)
ref-t		Reference temperature to use for coupling (1 per group)
pcoupl	Parrinello-Rahman	Turn on pressure coupling
	no	Box size remains fixed – no pressure coupling.
tau-p		Time constant (ps) for pressure coupling (1 per group)

ref-p		Reference pressure to use for coupling (1 per group)
gen_vel	yes	Choose velocities from the Maxwell Boltzmann distribution
	no	
gen_temp		Target temperature

5.6. Steps for MD in GROMACS

(Adapted mainly from the GROMACS 4.5.x tutorial on protein-ligand complex, found at: http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/complex_old/index.html), GROMACS 4.5.6 man pages and the GROMACS user manual 4.5.6 (van der Spoel et al. 2010).

An overview of the steps generally followed in MD is given in the 5.1.

Figure 5.1: Overview of MD steps

5.6.1. Receptor topology preparation

The "system topology file" (*.top) is created by the g_pdb2gmx command and is required to specify links to the atom types of each of the molecules to be included in a system for an MD simulation for a specified force field. After being created for the receptor, the topology file is gradually updated as other molecules of possibly different atom types are added to the system. The system topology file works in conjunction with the "include topology" (*.itp) files, which are referenced inside of the topology file by include statements to ".itp" files, which contain the actual topological information, such as the protein chain topologies, water model topology, ligand position restraints, ligand topology, ion topologies, etc. These "itp" files contain the actual atom types, bonds, atom pairs, bond angles, dihedrals, constants and other parameters that are used by the force field.

5.6.2. Ligand topology preparation

Depending on the nature of the ligand, the "g_pdb2gmx" tool might not be able to recognize the input ligand and its atom types, due to the different residue names. For this reason, other tools (local and web-based) are used for that purpose. Some of them include ACPYPE (AnteChamber PYthon Parser interfacE) (Sousa da Silva & Vranken 2012), PRODRG (Schüttelkopf & Van Aalten 2004) and CGenFF (Vanommeslaeghe et al. 2010). ACPYPE, which is the tool used in this case, returns topology files (*.top and *.itp) suitable for use with the GAFF (General Amber Force Field) and OPLS force fields in GROMACS (Sousa da Silva & Vranken 2012). A variant of the AMBER force field is used here.

5.6.3. Building the complex and its topology

The complex is assembled by concatenating the ".gro" files of the prepared receptor and ligand. Subsequently, the number of atoms is updated at the header to reflect the total number of atoms, while the protein box vector is kept at the very end of the ".gro" file. The accompanying topology file (previously generated from the receptor) is edited to contain an include statement pointing to the ligand ".itp" filename, which allows the atom types to be recognized by GROMACS. In the same file, the ligand name (corresponding to that found in the ligand ".itp" file) is also added to the molecules section.

5.6.4. Defining the simulation box and adding water/ ions

In this step, the parameters of the simulation box are defined by various flags to the "g_editconf" command. The "-bt" flag sets the box type, which can be of triclinic, cubic, dodecahedron or octahedron shape. The cube type has the largest volume, and thus has implications on the number of

atoms that can be packed inside, which will affect the program number of calculation to be done on the system. The "-d" option specifies the clearance (in nanometers) between the complex and the edge of the box. While setting this parameter, it is especially important to give a reasonable amount of clearance away from the complex, as the box is implemented as a periodic boundary, which means that if insufficient space is used, the atoms on the far sides along any plane may come too close so that spurious forces are calculated. The topology file is updated and a new "gro" file is created, with adjusted box vector and adjusted coordinates.

Solvation is accomplished by the "g_genbox" command which updates the topology file to include the solvent (water) molecules, and creates a new "gro" file containing the water molecules in addition to the complex. The previous ".gro" file generated by the "g_editconf" command is specified as input by the "-cp" (protein configuration) and the water model parameters are specified by the "-cs" (solvent configuration) flag.

Ions are added by "g_genion" command, which randomly replaces solvent molecules by specified mono-atomic ions. These modifications are updated in the topology file and a ".gro" file is generated as output. The ions required are specified by "-nname" (negative ion) and "-pname" (positive ion) while the number of positive and negative ions are specified by the "-np" and "-nn" flags, respectively. The salt concentration (in mol/liter) can be specified by the "-conc" flag, overriding the "-np" and "-nn" flags. An "ions.tpr" file is also required as input, which is obtained from the "g_grompp" (GROMACS preprocessor) command by providing any ".mdp" file (for example, the "em_real.mdp" file, which is usually used for energy minimization). The topology file is updated at the end of this step, including the newly inserted ions (e.g. chloride and sodium ions).

5.6.5. Energy minimization

Energy minimization is carried out on the whole system, by running the "g_grompp" command, followed by the "g_mdrun" command. The preprocessor's parameters are specified in an "mdp" file to update the topology and produce a "tpr" file. It is the "tpr" file which is used as input to the "g_mdrun" command to run the minimization. The output from the minimization is a system with reasonable atom geometries and solvent orientations.

5.6.6. Equilibration of system & restraining of the receptor-ligand complex

In the case of a receptor-ligand complex, it is desired that both the ligand and the receptor remain relatively fixed in the solvent during the equilibration phases - the protein restraints are generated automatically for the heavy (non-hydrogen) protein atoms by the "g_pdb2gmx" command, while

ligand restraints are generated separately by the "genrestr" command. Protein restraints are specified by the "define = -POSRES" statement in each of the equilibration "mdp" parameter files, namely the ones for NVT and NPT.

In the NVT equilibration, the temperature is raised and stabilized to a predetermined value (set in the "nvt.mdp" parameter file) by energizing the system for a preset amount of time by randomly assigning velocities (picked from the Boltzmann distribution) to the atoms in the system. The ligand and receptor remain restrained throughout the equilibration procedure.

On the other hand, NPT equilibration stabilizes the pressure (set in the "npt.mdp" file), certain amount of time (for example 100ps), until the system's pressure is stabilized.

Both of the equilibration steps require the output from the "g_gromp" command that generates the "tpr" files required by the "g_mdrun" command of GROMACS.

5.6.7. Production MD

The restraints are removed from both the ligand, marking the end of equilibration and the system is allowed move with the given atomic velocities previously attained from equilibration.

5.6.8. Analysis

GROMACS provides several tools for analyzing and monitoring of molecular dynamic runs. Some examples include:

- g_energy: Used to analyze components of potential energy, but can also be used to retrieve other properties of the system such as pressure and temperatue (van der Spoel et al. 2010).
- g_rms: Computes the Root Mean Squared Deviation (RMSD) between two structures, typically frames along a trajectory against a reference structure (van der Spoel et al. 2010). RMSD is calculated as the square root of the averaged sum of squared deviations between atom pairs for a defined set of atoms (van der Spoel et al. 2010).
- g_hbond: Computes and displays the number of hydrogen bonds between all possible donors and acceptors of hydrogen bonds (van der Spoel et al. 2010).
- g_gyrate: Computes the gyration radius of groups of atoms as a function of time, giving an idea of the general compactness a given structure. The smaller the gyration radius, the more compact the group of atoms.

Correction for periodicity, i.e. when the atoms cross the periodic boundary defined by the box size and type, can be accomplished by using the "trjconv" command.

METHODOLOGY

5.1. Preparations of receptor/ ligand complexes for MD with GROMACS

5.1.1. Ligand preparation

The lowest-energy-scoring docked ligands were prepared for molecular dynamics by putting back the non-polar hydrogen atoms that had been previously removed by AutoDockTool's "prepare_ligand4.py" tool) using OpenBabel from within a Python script prior to making the GROMACS topology files using the ACPYPE "AnteChamber PYthon Parser interfacE" (Wang et al. 2006) tool. The jobs were submitted to the "perkin" cluster.

5.1.2. Receptor preparation and protonation

The protease was protonated to pH7 using the PDB2PQR tool (Dolinsky et al. 2007) to produce a "pdb" file with atom types corresponding to the AMBER forcefield. However, the atom types were not entirely recognized by GROMACS' "pdb2gmx" command for the AMBER03 forcefield. Therefore, a Python script was written to convert the atom types, which mainly consisted of renumbering several hydrogen atoms that had their numbering in a different format (e.g. starting from HD2 to HD3, without having an initial HD1 atom type). The N-termini and the C-termini residues from both of the protease chains were also adjusted accordingly, prefixing the residue types with an "N" and a "C" respectively.

5.1.3. Building the complex and updating its topology

In order to build the receptor-ligand complex, the GROMACS file produced from ACPYPE was merged with the processed receptor resulting from the previous step. Only one box vector (the one from the protein) was retained. The ligand molecule and its atom types were then updated in the topology file.

5.1.4. Preparing the simulation environment

The simulation box was built with a distance of 1.5 Angstroms between the complex and the edge of the box, which was itself set to the triclinic type. Water (spc model) was added, followed by sodium and chloride ions at a concentration of 0.15M, with the "-neutral" flag of the "genion" command to neutralize the system. The system was subsequently minimized in the created environment and the potential energy was evaluated by the "g_energy" command.

5.2. Molecular dynamics runs: energy minimization to production MD

A bash script was written to automate the MD procedure. The jobs were separated, for shorter queue waiting times and were run both at the chemistry department on the "perkin" cluster and at the Center for High Performance Computing (CHPC).

5.2.1. Energy minimization

The energy minimization was performed with the default parameters specified in the "em_real.mdp" file available from the GROMACS tutorial website (http://www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/complex_old/Files/) and the output was processed using an R script to give line plots of the potential energies.

5.1.2. Receptor and ligand restraints

Restraints were initially generated for the protein receptor by the "pdb2gmx" command in the form of ".itp" files. Ligand restraints on the other hand were obtained from the "genrestr" command, using as input the GROMACS ".itp" file obtained from ACPYPE for each ligand and the information was updated to the relevant topology file "topol.top".

5.2.3. NVT equilibration

NVT equilibration was performed with default parameters, as available from "nvt.mdp" parameter file (leap frog integrator, 100ps, LINCS constraint algorithm, velocity generation, V-rescale temperature coupling, default short range and long range interaction parameters) available from the GROMACS tutorial website, except for the temperature that was set to 310K, which corresponds to the human body temperature.

5.2.4. NPT equilibration

Default parameters were also used (leap frog integrator, 100ps, LINCS constraint algorithm, Parrinello-Rahman pressure coupling, V-rescale temperature coupling, 1 bar pressure, no velocity

generation, default short range and long range interaction parameters), as available from the "npt.mdp" file from the GROMACS tutorial website. The temperature was kept at 310K, as for NVT.

5.2.5. Production MD

The time for the production MD run was increased to 20ns to have a better view of any possible late stabilizations or fluctuations in the RMSD plots. Again, the temperature was set to 310K. All other parameters were used at their default values (LINCS constraints, Parrinello-Rahman pressure coupling, V-rescale temperature coupling , 1 bar pressure, no velocity generation, default short range and long range interaction parameters)

5.2.6. MD monitoring and analysis

Several line plots were used to check for the suitability of the parameters utilized and for analyses at a later stage. These included:

- Energy minimization: Potential energy of the system.
- NVT monitoring: Potential energy and temperature of the system.
- NPT monitoring: Potential energy, pressure and temperature of the system.
- Production MD monitoring and analyses: Potential energy and temperature of the system, RMSD, hydrogen bonds, gyration, active site ligand distance, flap distance.

RESULTS AND DISCUSSION

Graphs generated from the GROMACS analysis tools are shown in the current section in the following order: 1) energy minimization, 2) NVT equilibration, 3) NPT equilibration, 4) Production MD.

5.1. Energy minimization plot

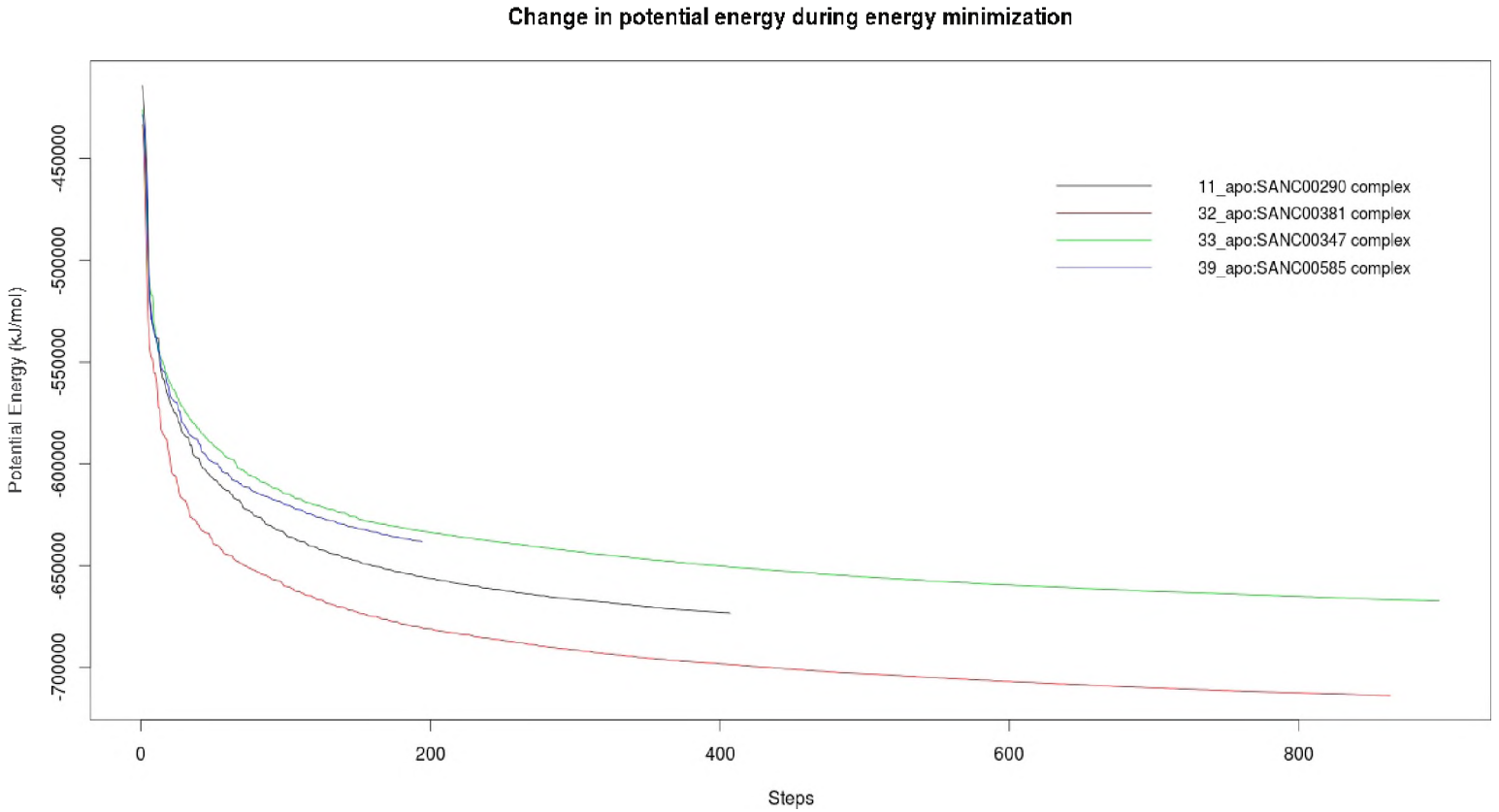


Figure 5.2: Change in potential energy during the energy minimization steps of four shortlisted HIV protease/ ligand complexes

The energy minimization ensured that each of the systems (each comprising the protein, ligand water, sodium ions and chloride ions) settled to their respective local potential energy minima after the addition of solvent (water) and ions, from their previous *in vacuo* states. During the minimization step (using steepest descent algorithm), the potential energies are seen to drop exponentially during the first few steps, to level off at different rates, depending on the complexes, with complex 39 apo/ SANC00585 finishing earlier than the others. Potential energy specifics of are reported in the table below:

Table 5.2: Energy minimization specifics of the receptor-ligand complexes

Receptor:ligand complex	Final potential energy gradient found (KJ/mol/ps)	Steps	Initial receptor conformation
11_apo:SANC00290	(-673284.9 + 673142.1)/2 = -71.4	407	closed
32_apo:SANC00381	(-713763.3 + 713707.5)/2 = -27.9	863	open
33_apo:SANC00347	(-667288.9 + 667237.3)/2 = -25.8	897	closed
39_apo:SANC00585	(-638189.8 + 637868.2)/2 = -160.8	194	closed

From Table 5.2, complex 39_apo:SANC00585 was found to have the steepest final gradient amongst the 4 systems. Further minimization could be carried out to obtain more favorable energies closer to their respective global minima for all of the complexes, however over-minimization may result in molecule distortions, worsening at every step leading to and including the final production MD. Following minimizations, the complexes were equilibrated at 310K in an NVT ensemble.

5.2. NVT equilibration plots

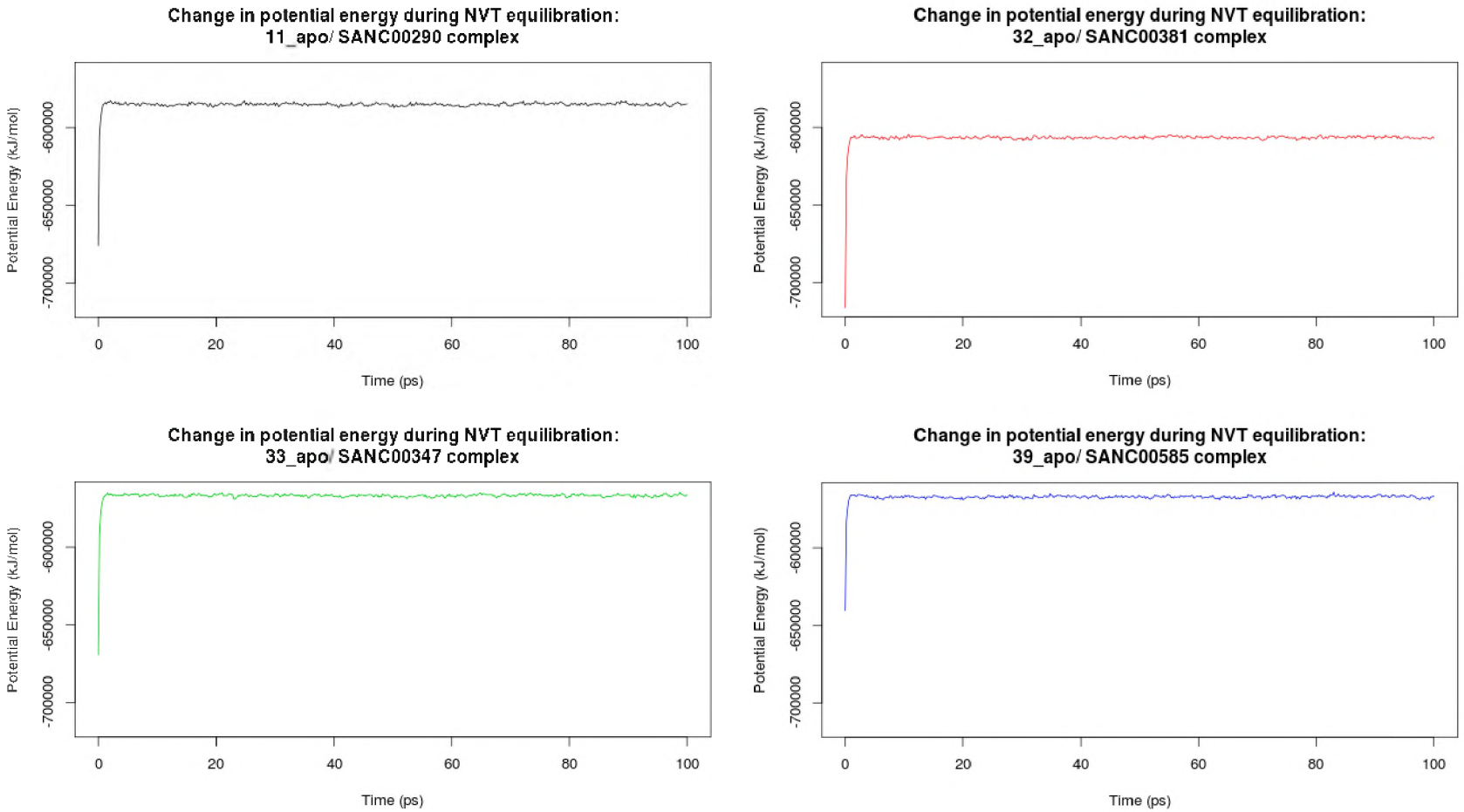


Figure 5.3: NVT equilibration: Potential energy

As can be seen from the potential energy plot of the NVT equilibration step (Figure 5.3), all the complexes start at initial energies that roughly equal those produced at the end of energy minimization that was previously done, exponentially increasing to quickly plateau off, revolving around the final energy values until the end of the 100ps. The potential energy values all stabilize at different levels most likely due to the fact that the complexes have different atomic compositions (number and identity of ligand/ protease atoms), which for example would involve different bonded (covalent) and non-bonded interactions (Lennard Jones potentials and electrostatic interactions). Potential energy values approximated in Table 5.3.

Table 5.3: NVT equilibration: potential energy specifics

Receptor:ligand complex	Initial potential energy (MJ/mol)	Final potential energy (MJ/mol)
11_apo:SANC00290	-675.7	-584.5
32_apo:SANC00381	-716.1	-606.6
33_apo:SANC00347	-669.4	-566.1
39_apo:SANC00585	-640.5	-566.9

The fact that the potential energies remained fairly constant for the most part of the NVT equilibration until the end of equilibration, hinted that the kinetic energy, and therefore the temperature of the system were also stabilized for the system. Temperature values during equilibration were also recorded and are shown in Figure 5.4. At time zero, it can be seen that the temperature is around the set value (approximately 310K) in every system, but decreases sharply within the first 2ps to immediately spike back and stabilize at around 310K until the end of the 100ps runs. Temperature stabilization is a good indication that the atoms of the system have the right distribution of velocities corresponding to 310K and that the system is suitable to proceed to pressure equilibration. The systems were subsequently brought to NPT equilibration.

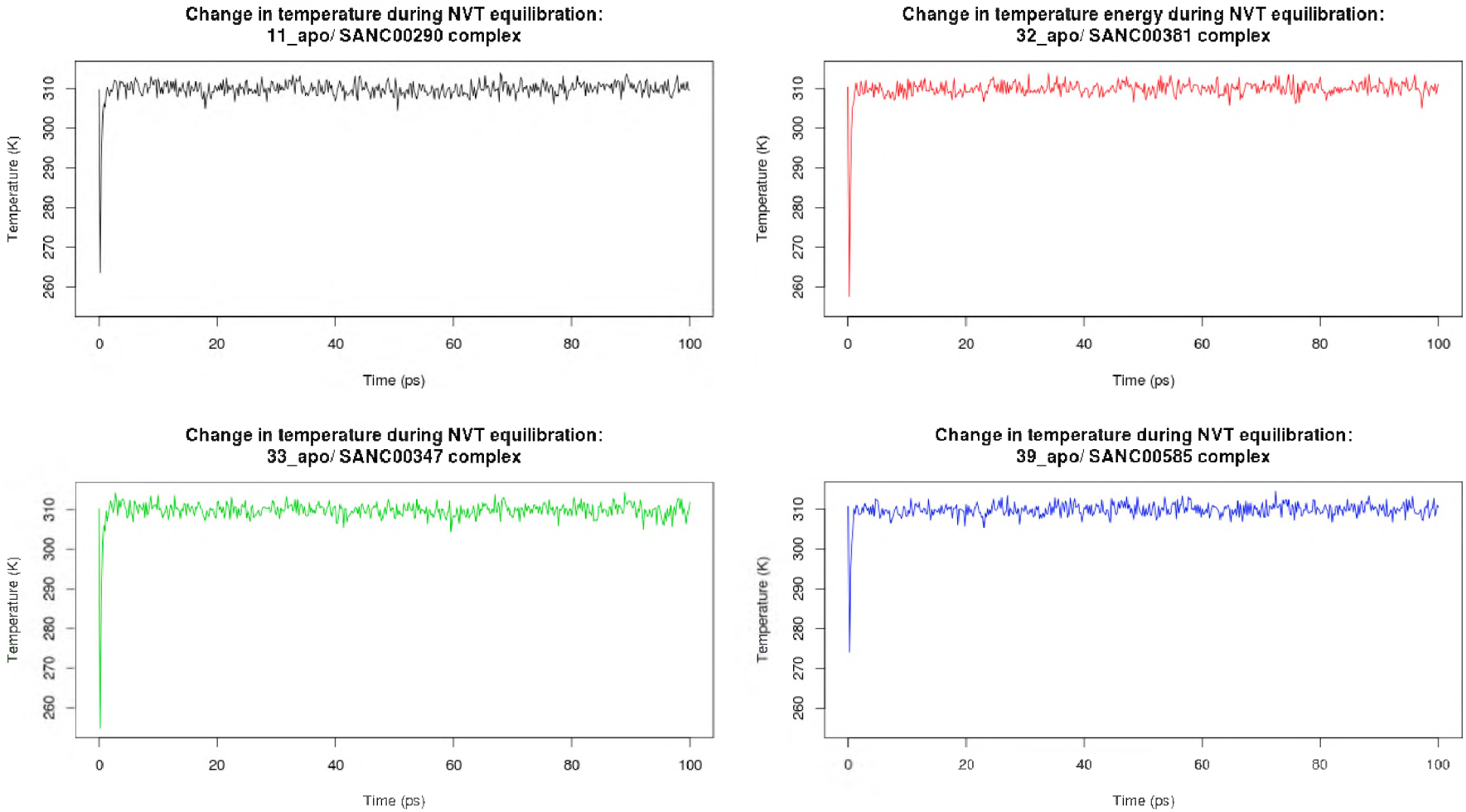


Figure 5.4: NVT equilibration: Temperature

Pressure, temperature and potential energy were inspected for the NPT equilibration. Running averages were computed (over a window size of 30 frames) for both the pressure and temperature equilibration plots and their respective values were found to be fluctuating around their respective means. The running averages (shown in thick red lines) eliminated some of the noise, showing more clearly that the temperature is revolving around 310K in Figure 5.7. On the other hand, pressure seems to stabilize around zero (Figure 5.6), however this is mainly due to a y-axis scaling artifact, as the pressure was initially set at 1 bar. Potential energies (Figure 5.5) revolve around their initial values until the end of the NPT equilibration. The fact that both the temperature and pressure were stabilized indicated that the system was ready for releasing the force restraints to proceed to production dynamics.

5.3. NPT equilibration plots

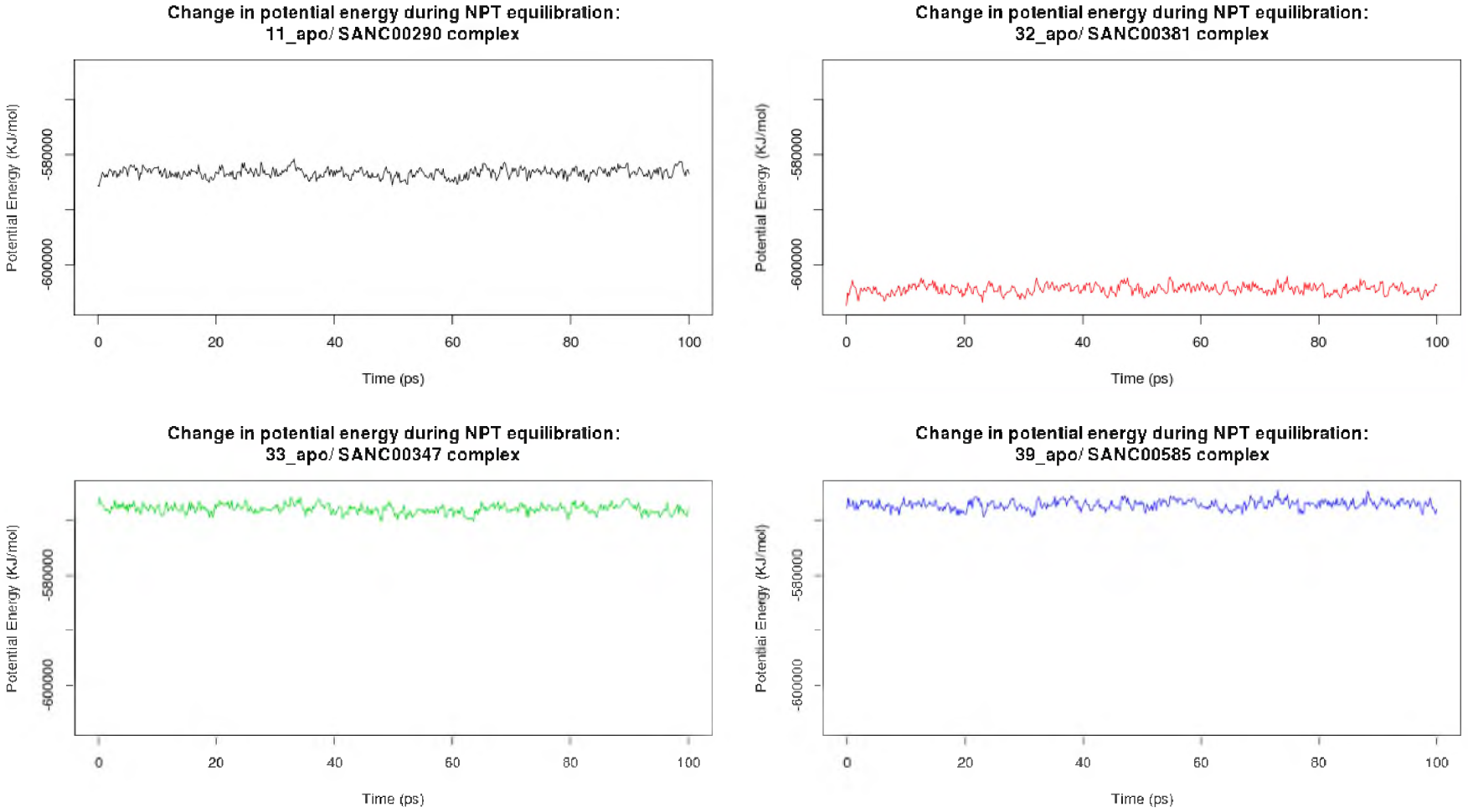


Figure 5.5: NPT equilibration: Potential energy

Figure 5.6: NPT equilibration: Pressure

Figure 5.7: NPT equilibration: Temperature

5.4. Production MD

After running the production MDs, the systems were corrected for the periodic boundary conditions as different sections of the complexes were found to appear at opposing faces of the triclinic simulation boxes, which severely affected the calculation of the RMSD values for the protein. Aafter adjusting for the periodic boundaries, one of the initially open conformation complexes was chosen for displaying flap movements. Translational movement inside the simulation box made visualization of flap movement difficult, therefore each frame for the resulting trajectory was aligned using the progressive fit method (from the "trjconv" tool) – the protein group was chosen for the least squares fit and the whole system was output. Visualization was accomplished using VMD.

5.4.1. RMSD analysis

Alpha carbon RMSD values were calculated for the production MD (Figure 5.8) by aligning protein backbones (by least square fitting) across every frame of the trajectory file using only the protein to obtain clearer signals of the motion based on the protein chains only, with energy feedbacks coming from the interactions with the ligand. It can be seen that the averaged deviations are slight, revolving around 0.15nm after an initial jump from zero in all cases except for complex 32_apo/SANC00381, where a gradual increase in RMSD is observed for the first 7-8ns before stabilizing after about 13ns. This change happened only for the initially open conformation model, in contrast to all the others that were built from closed conformation models. This difference was investigated further and found to be supported by several other calculated metrics, as explained later in the current section.

5.4.2. Choice of receptor-ligand complex

Overall, little movement was observed for the complexes built from the initially-closed conformation models, namely complex 11_apo/ SANC00290, 33_apo/ SANC00347 and 39_apo/ SANC00585. Therefore more emphasis was laid on explaining the observations for the initially-opened conformation complex (32_apo/ SANC00381)

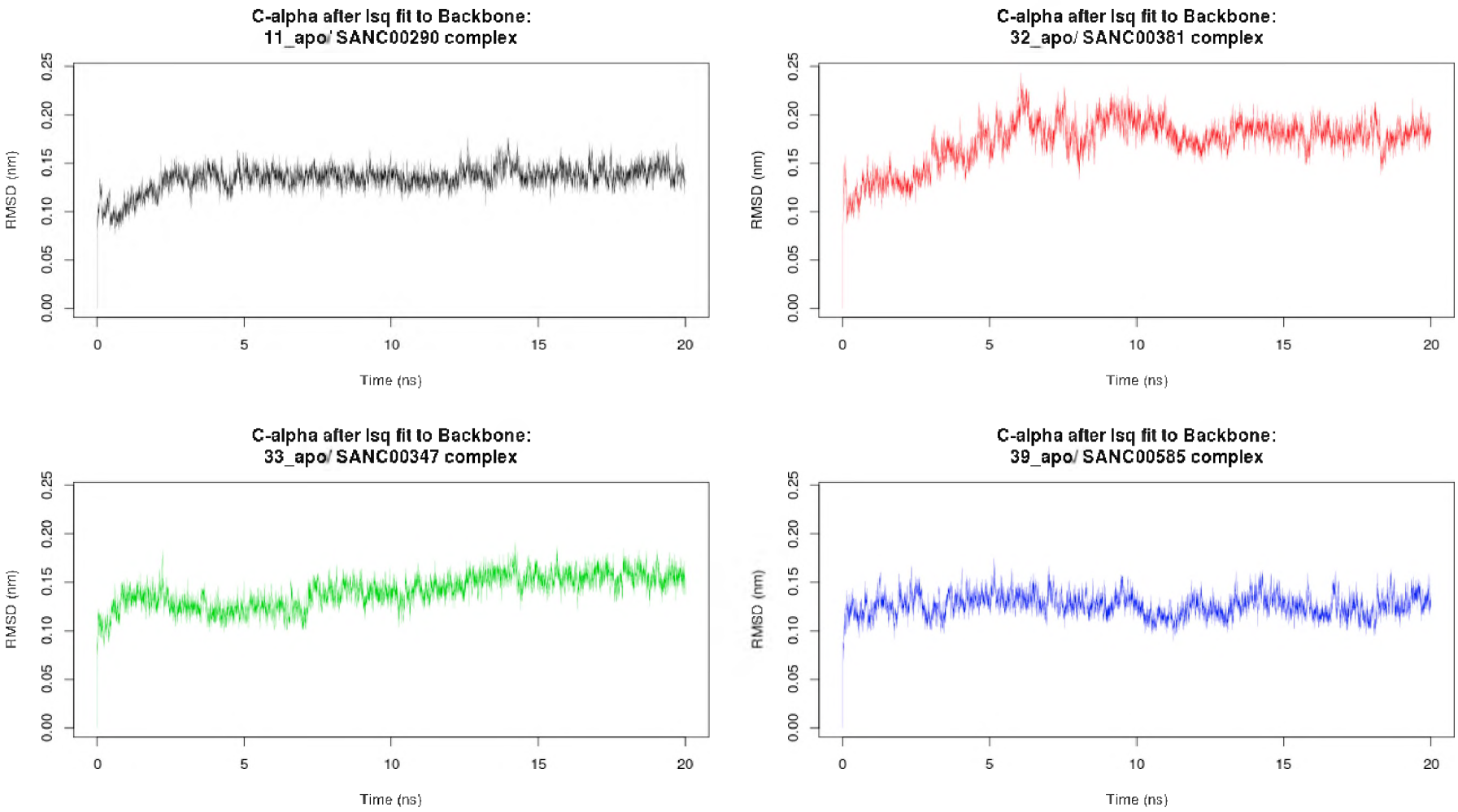


Figure 5.8: Production MD: RMSD of the protein

5.4.3. Analysis of hydrogen bonding

The changes in the number of hydrogen bonds within the complexes (Figure 5.9) show that with the exception of complex 32_apo/SANC00381, there is a high frequency of occurrence of hydrogen bonds all throughout the MD production run that might explain the differences in the evolution of the gyration radii of between the complex 32_apo/SANC00381 and the rest of the complexes. The high frequency of hydrogen bonding indicates that the distances between hydrogen bond donors and acceptors mostly remain within the acceptable margins so that they stabilize the complexes of the initially closed conformation models throughout the course of the production dynamics.

5.4.4. Analysis of radius of gyration

The radius of gyration during the MD run agree with the evolution of hydrogen bonding, revealing that the complexes do not change very much in their degree of compactness, except for complex 32_apo/ SANC00381 where the gyration radius drops from about 1.88nm to 1.78nm during the first 5ns and stabilize around 1.78nm until the end of the MD run. This corresponds with the initial absence of hydrogen bonding (Figure 5.9) and the larger initial distance between the flaps (Figure 5.12). The higher initial gyration radius can be explained by the initially unstable (energetically unfavorable) conformation of the flaps of the protease model and the low frequency of hydrogen bonding inside the complex, due to the removal of the co-crystallized ligands at the flaps from the protease model.

5.4.5. Tracking the distance between the ligand and the active site

According to Figure 5.11, ligand SANC00290 is stably bound to the active site of receptor 11_apo for the first 14ns, but then appears to fluctuate across a distance averaging 3nm, which is still small and probably is a sign of bond rotation/ atom translation along the ligand chain, moving the hydrogen atom that was designated to obtain ligand distance from the active site. Ligand SANC00585 appears to be the more strongly bound to the active site of receptor 39_apo after an initial fluctuation lasting about 4ns. In Figure 5.12. it can be seen that the flaps remain mainly closed for the three initially closed conformation receptors, except for the 32_apo/ SANC00381 complex which was initially open.

Figure 5.10: Production MD: Analysis of gyration radii

Figure 5.11: Receptor-ligand distances

5.4.6. Tracking the distance between the protease flaps of 32_apo/ SANC00381

After computing the distance between the flaps of the dimeric proteases, a high degree of fluctuation and large distances were found between the flaps of complex 32_apo/ SANC00381 during the first 8ns. The same explanation given before about the low protein compaction and low frequency of hydrogen bonding applies – the absence of the original ligand at the flaps (in the template crystal structure) destabilized the complex, giving the flaps a higher mobility, hence the higher fluctuations during the first 8ns, before finding a more stable conformation, as observed by the decrease in the amount of fluctuations for the given complex in Figure 5.12. Eight frames were extracted (in VMD) at different time steps from the trajectory file to give a better overview of what was happening within the complex (shown in Figure 5.13). From the sample of frames in Figure 5.13, the highest distance is initially observed and a high amount of variation is found to occur all throughout the selected frames – the smallest inter-flap distance is nevertheless observed at 8ns, indicating that the flaps were trying to close. However, a better picture of the ligand is also observed in the same figure, overcoming the initial bias of the choice of ligand atoms (in Figure 5.11) for computing the ligand-active site distance. It is easily seen that the ligand tries to exit (the unstable) complex at time 600ps and 1ns, but is found deeper inside the binding pocket after 16ns as the flaps try to close, with an inter-flap distance of 6.2nm.

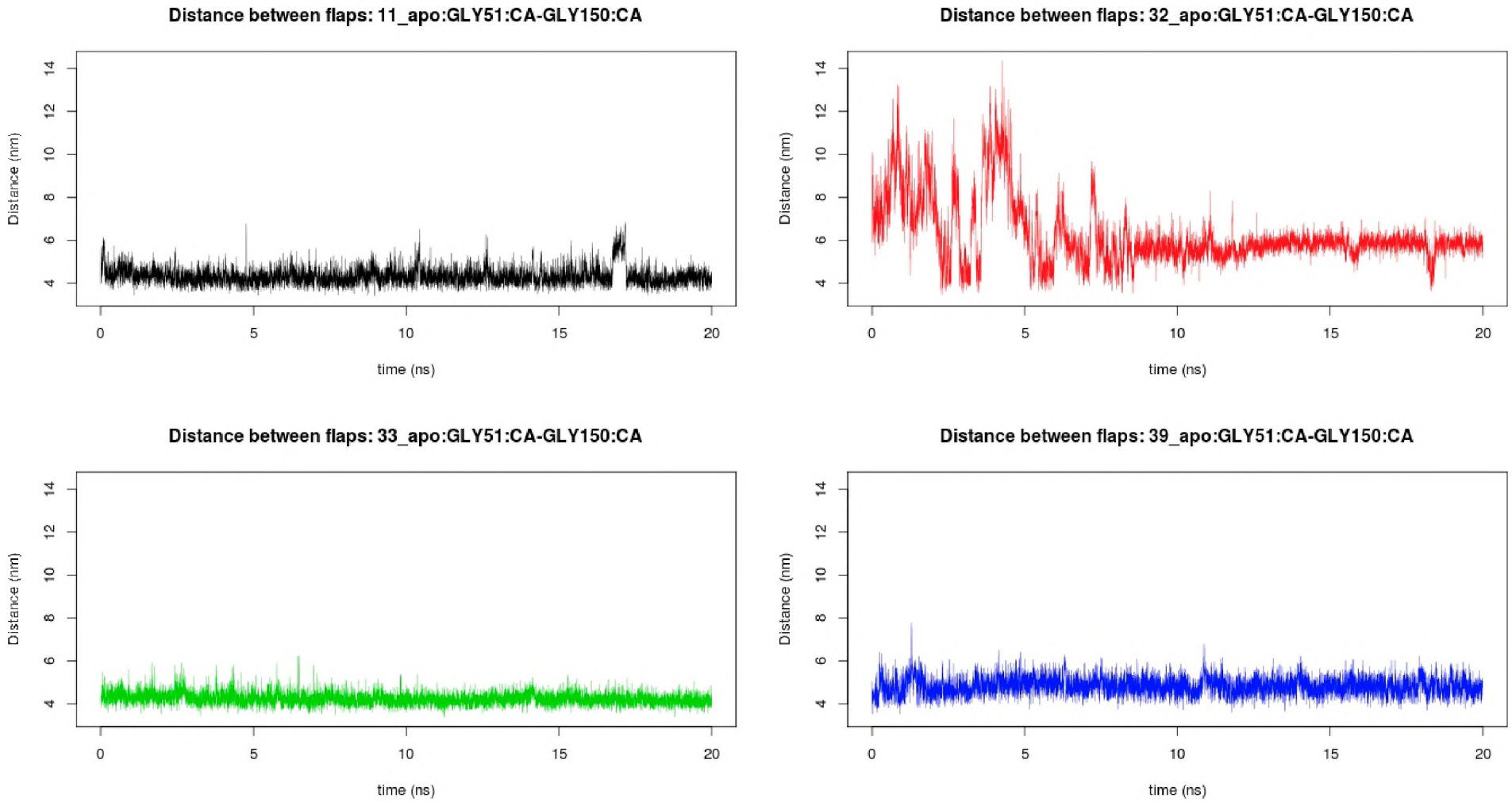


Figure 5.12: Distance between the protease flaps.

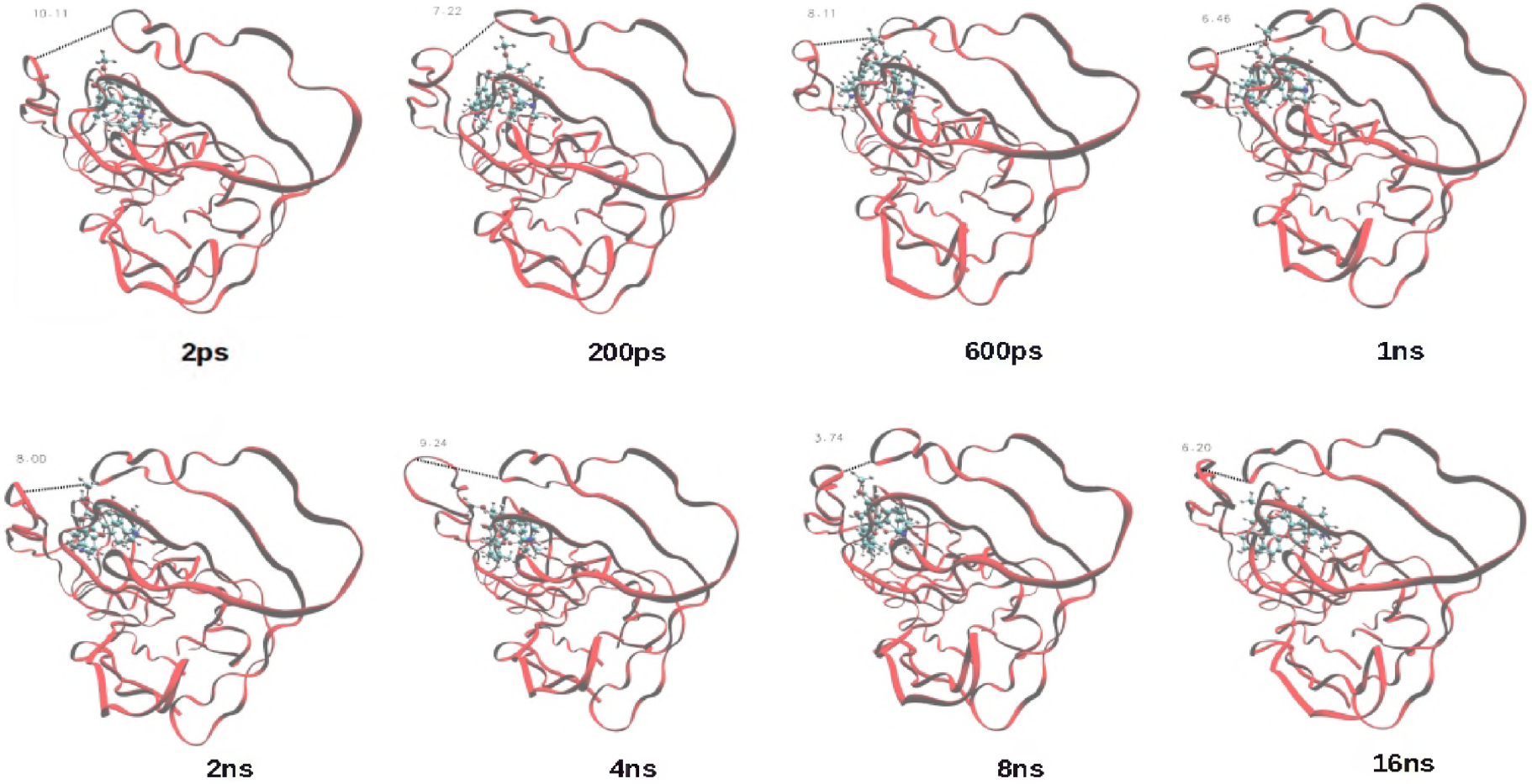


Figure 5.13: Overview of flap motions for the 32_apo/ SANC00381 complex during the first 16ns of production MD: Flap distances are shown as dotted lines between glycines 51 from chains A and B of HIV protease.

CONCLUSION

Results from MD analyses provide support as to the stability of four receptor-ligand complexes under human physiological conditions of ionic concentration, pH, solvent, pressure and temperature. Complexes with initially closed flap receptors were observed to be stable all along the MD runs. On the other hand, complex 32_apo/ SANC00381 (initially opened flap receptor conformation) was accompanied with initial flap movements and ligand motion, before stabilizing. The complex stabilities themselves, measured using receptor RMSD, gyration radii, ligand-active sire distance, hydrogen bonding and flap tip distances all agreed to initial flap movement only in the initially open conformation receptor complex, before stabilization.

Recommendations for further work

Further characterization methods such as the Molecular Mechanics Poisson Boltzmann Surface Area (MM-PBSA) could be carried out on the complexes selected for MD (over short time intervals) to compute differences between the bound and unbound states between the receptors and their ligands. Further laboratory tests (such as the Enzme-Linked Immunosorbent assay or the phenotypic resistance assay) could be used to test for receptor ligand binding and their effect on the ability of the virus particles to proliferate (AIDSMEDS 2010), accompanied by assays for drug toxicity (such as the tetrazolium test) (Feng et al. 2012), before approval of the potential PI's.

The scripts for energy calculations could also be improved, for example to take into consideration the allowable bond angle between the hydrogen acceptor and donor atoms. In addition, the script for plotting the energy profiles could be fine-tuned for application with other protein-ligand complexes.

References

Abraham, M.J., 2011. GROMACS - Terminology. Available at: http://www.gromacs.org/Documentation/Terminology [Accessed November 6, 2015].

Abraham, M.J. et al., 2015. GROMACS User Manual version 5.0.5. Available at: www.gromacs.org.

Adamson, C.S. & Freed, E.O., 2007. Human immunodeficiency virus type 1 assembly, release, and maturation. *Advances in pharmacology (San Diego, Calif.)*, 55(7), pp.347–87. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17586320.

Adamson, C.S. & Freed, E.O., 2010. Novel approaches to inhibiting HIV-1 replication. *Antiviral Research*, 85, pp.119–141.

AIDS.gov, 2015. Overview of HIV Treatments. Available at: https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/treatment-options/overview-of-hiv-treatments/ [Accessed December 2, 2015].

AIDS Foundation of South Africa, 2014. HIV/AIDS in South Africa - AIDS Foundation of South Africa. Available at: http://www.aids.org.za/hivaids-in-south-africa/ [Accessed November 9, 2015].

AIDSMEDS, 2010. Understanding Drug Resistance : What Is Phenotypic Resistance Testing? Available at: http://www.aidsmeds.com/articles/Resistance_7510.shtml [Accessed December 3, 2015].

Ali, A. et al., 2010. Molecular basis for drug resistance in HIV-1 protease. *Viruses*, 2, pp.2509–2535.

Apostolov, R., 2014. How-tos - Gromacs. Available at: http://www.gromacs.org/Documentation/How-tos [Accessed November 11, 2015].

Ariffin, T.A.A.T.M. et al., 2014. Antiretroviral drug resistance and HIV-1 subtypes among treatment-naive prisoners in Kelantan, Malaysia. *The Journal of Infection in Developing Countries*, 8(8). Available at: http://www.jidc.org/index.php/journal/article/view/4095.

Aszódi, A. & Taylor, W.R., 1996. Homology modelling by distance geometry. *Folding & design*, 1(5), pp.325–334.

Baldauf, S.L., 2003. Phylogeny for the faint of heart: A tutorial. *Trends in Genetics*, 19(6), pp.345–351.

Becker, S., 2003. Atazanavir: improving the HIV protease inhibitor class. *Expert Rev. Anti. Infect. Ther.*, 1, pp.403–413.

Benkert, P., Biasini, M. & Schwede, T., 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3), pp.343–350. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq662.

Berendsen, H.J.C. et al., 1984. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8), pp.3684–3690. Available at: http://link.aip.org/link/JCPSA6/v81/i8/p3684/s1&Agg=doi%5Cnpapers2://publication/doi/10.1063/1.448118.

Berger, E.A., Murphy, P.M. & Farber, J.M., 1999. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annual review of immunology*, 17, pp.657–700.

Blundell, T.L. et al., 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326(6111), pp.347–352. Available at: http://www.nature.com/nature/journal/v326/n6111/abs/326347a0.html.

Boeri, E. et al., 2004. Phylogenetic internal control for HIV-1 genotypic antiretroviral testing. *The new microbiologica*, 27(2 Suppl 1), pp.105–109.

Botes, L., van den Heever, W.M.J. & Pretorius, G.H.J., 2007. The Structural Biology of HIV. *Medical Technology SA*, 21(June), pp.13–18. Available at: http://www.smltsa.org.za/journal/archive/vol21no13.pdf.

Brenner, B.G. & Wainberg, M.A., 2013. Future of phylogeny in HIV prevention. *Journal of acquired immune deficiency syndromes (1999)*, 63 Suppl 2(SUPPL. 2), pp.S248-54. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84880210924&partnerID=tZOtx3y1.

Brocchieri, L., 2001. Phylogenetic inferences from molecular sequences: review and critique. *Theoretical population biology*, 59(1), pp.27–40.

Brocklehurst, S.M. & Perham, R.N., 1993. Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated H-protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure. *Protein Science*, 2(4), pp.626–639.

Cai, Y. et al., 2012. Differential flap dynamics in wild-type and a drug resistant variant of HIV-1 protease revealed by molecular dynamics and NMR relaxation. *Journal of Chemical Theory and Computation*, 8(10), pp.3452–3462.

Calza, L. et al., 2013. Incidence of renal toxicity in HIV-infected, antiretroviral-naïve patients starting tenofovir/emtricitabine associated with efavirenz, atazanavir/ritonavir, or lopinavir/ritonavir. *Scandinavian journal of infectious diseases*, 45(2), pp.147–54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22991923.

Cameron, D.W. et al., 1999. Ritonavir and saquinavir combination therapy for the treatment of HIV infection. *AIDS*, 13(2), pp.213–224.

Castro-Nallar, E. et al., 2012. The evolution of HIV: Inferences using phylogenetics. *Molecular Phylogenetics and Evolution*, 62(2), pp.777–792. Available at: http://dx.doi.org/10.1016/j.ympev.2011.11.019.

Chang, M.W. et al., 2010. Virtual screening for HIV protease inhibitors: A comparison of AutoDock 4 and Vina. *PLoS ONE*, 5.

Cickovski, T. et al., 2010. MDLab: A molecular dynamics simulation prototyping environment. *Journal of Computational Chemistry*, 31(7), pp.1345–1356. Available at: http://www3.nd.edu/~izaguirr/papers/Cick09.pdf [Accessed August 22, 2015].

Claessens, M. et al., 1989. Modelling the polypeptide backbone with "spare parts" from known protein structures. *Protein Engineering, Design and Selection*, 2(5), pp.335–345.

Colvin, R. & Haas, G., 1995. Protease inhibitor update. *Common Factor*, 10(15).

Cuevas, J.M. et al., 2015. Extremely High Mutation Rate of HIV-1 In Vivo S. L. Rowland-Jones, ed. *PLOS Biology*, 13(9), p.e1002251. Available at: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002251#sec007 [Accessed September 17, 2015].

Czodrowski, P., Sotriffer, C.A. & Klebe, G., 2007. Atypical protonation states in the active site of HIV-1 protease: A computational study. *Journal of Chemical Information and Modeling*, 47(4), pp.1590–1598.

Dassault Systèmes BIOVIA, 2015. Discovery Studio Modeling Environment, Release 4.5. Available at: http://accelrys.com/products/discovery-studio/.

Dauchy, F.-A. et al., 2011. Increased risk of abnormal proximal renal tubular function with HIV infection and antiretroviral therapy. *Kidney international*, 80(3), pp.302–309.

Davies, D.R., 1990. The structure and function of the aspartic proteinases. *Annual review of biophysics and biophysical chemistry*, 19(1), pp.189–215.

Debouck, C. et al., 1987. Human immunodeficiency virus protease expressed in Escherichia coli exhibits autoprocessing and specific maturation of the gag precursor. *Proceedings of the National Academy of Sciences of the United States of America*, 84(24), pp.8903–8906.

Delelis, O. et al., 2008. Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology*, 5, p.114.

Desimmie, B.A. et al., 2014. Multiple APOBEC3 restriction factors for HIV-1 and one vif to rule them all. *Journal of Molecular Biology*, 426(6), pp.1220–1245.

Dolinsky, T.J. et al., 2007. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(SUPPL.2).

Doms, R.W., 2000. Beyond receptor expression: the influence of receptor conformation, density, and affinity in HIV-1 infection. *Virology*, 276(2), pp.229–237.

Eron, J. et al., 2006. The KLEAN study of fosamprenavir-ritonavir versus lopinavir-ritonavir, each in combination with abacavir-lamivudine, for initial treatment of HIV infection over 48 weeks: a randomised non-inferiority trial. *Lancet*, 368(9534), pp.476–482.

Eswar, N. et al., 2008. Protein structure modeling with MODELLER. *Methods in molecular biology (Clifton, N.J.)*, 426, pp.145–159.

FDA, 2015. HIV/AIDS Treatment - Antiretroviral drugs used in the treatment of HIV infection. Available at: http://www.fda.gov/forpatients/illness/hivaids/treatment/ucm118915.htm [Accessed December 2, 2015].

de Felipe, B. et al., 2011. Prevalence and resistance mutations of non-B HIV-1 subtypes among immigrants in Southern Spain along the decade 2000-2010. *Virology journal*, 8, p.416.

Feng, L. et al., 2012. A potential in vitro and in vivo anti-HIV drug screening system for Chinese herbal medicines. *Phytotherapy research : PTR*, 26(6), pp.899–907. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22852142.

Feyfant, E., Sali, A. & Fiser, A., 2007. Modeling mutations in protein structures. *Protein science : a publication of the Protein Society*, 16(9), pp.2030–41. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2206969&tool=pmcentrez&rendertype=abstract.

Fiser, A. & Šali, A., 2003. MODELLER: Generation and Refinement of Homology-Based Protein Structure Models. *Methods in Enzymology*, 374, pp.461–491.

Flor-Parra, F. et al., 2011. The HIV type 1 protease L10I minor mutation decreases replication capacity and confers resistance to protease inhibitors. *AIDS research and human retroviruses*, 27(1), pp.65–70.

Floudas, C.A. et al., 2006. Advances in protein structure prediction and de novo protein design: A review. *Chemical Engineering Science*, 61(3), pp.966–988. Available at: http://www.sciencedirect.com/science/article/pii/S0009250905002988 [Accessed December 17, 2014].

Freedberg, D.I. et al., 2002. Rapid structural fluctuations of the free HIV protease flaps in solution: relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein science*, 11(2), pp.221–232.

Frenkel, D. & Smit, B., 2001. Understanding Molecular Simulation: From Algorithms to Applications. , 50(7), p.664. Available at: http://scitation.aip.org/content/aip/magazine/physicstoday/article/50/7/10.1063/1.881812.

Fu-xiang, W. et al., 2007. Subtype and sequence analysis of HIV-1 strains in Heilongjiang Province. *CHINESE MEDICAL JOURNAL*, 120(22), pp.2006–2010.

Galkin, A.N. et al., 2006. Full-length genomic sequencing and analysis of four HIV type 1 subtype B isolates circulating in the territory of Russia. *AIDS research and human retroviruses*, 22(11), pp.1192–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17147510.

Ganser-Pornillos, B.K., Yeager, M. & Sundquist, W.I., 2008. The structural biology of HIV assembly. *Current Opinion in Structural Biology*, 18(2), pp.203–217.

Gathe, J.C. et al., 2006. Long-term (120-Week) antiviral efficacy and tolerability of fosamprenavir/ritonavir once daily in therapy-naive patients with HIV-1 infection: an

uncontrolled, open-label, single-arm follow-on study. *Clinical therapeutics*, 28(5), pp.745–54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16861096.

Ginalski, K., 2006. Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, 16, pp.172–177.

González de Requena, D. et al., 2003. Indinavir plasma concentrations and resistance mutations in patients experiencing early virological failure. *AIDS research and human retroviruses*, 19(6), pp.457–459.

Goodsell, D.S. et al., 1996. Automated Docking of Flexible Ligands: Applications of AutoDock. *Journal of Molecular Recognition*, 9(November 1995), pp.1–5.

Goodsell, D.S., 2000. HIV-1 Protease. *RCSB Protein Data Bank*. Available at: http://dx.doi.org/10.2210/rcsb_pdb/mom_2000_6 [Accessed March 3, 2015].

Graham, L.P., 2013. *Introduction to medicinal chemistry* 5th ed., Oxford: Oxford University Press, United Kingdom.

Gromiha, M.M., 2010. *Protein Bioinformatics: From Sequence to Function*, Academic Press. Available at: https://books.google.com/books?id=sEQNm-9vRV0C&pgis=1 [Accessed December 11, 2015].

Guenther, B. et al., 1997. Crystal structure of the delta' subunit of the clamp-loader complex of E. coli DNA polymerase III. *Cell*, 91(3), pp.335–345.

Hall, B.G., 2013. Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution*, 30(5), pp.1229–1235.

Hall, T., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, pp.95–98. Available at: http://jwbrown.mbio.ncsu.edu/JWB/papers/1999Hall1.pdf.

Havel, T.F. & Snow, M.E., 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *Journal of molecular biology*, 217(1), pp.1–7.

Hemelaar, J., 2012. The origin and diversity of the HIV-1 pandemic. *Trends in Molecular Medicine*, 18(3), pp.182–192. Available at: http://dx.doi.org/10.1016/j.molmed.2011.12.001.

Hooft, R.W. et al., 1996. Errors in protein structures. *Nature*, 381(6580), p.272.

Hornak, V. et al., 2006. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4), pp.915–920.

Hresko, R.C. & Hruz, P.W., 2011. HIV protease inhibitors act as competitive inhibitors of the cytoplasmic glucose binding site of GLUTs with differing affinities for GLUT1 and GLUT4. *PLoS ONE*, 6(9).

Huang, S.-Y. & Zou, X., 2010. Advances and challenges in protein-ligand docking. *International journal of molecular sciences*, 11(8), pp.3016–34. Available at: http://www.mdpi.com/1422-0067/11/8/3016/htm [Accessed December 2, 2014].

Huang, X. et al., 2014. The role of select subtype polymorphisms on HIV-1 protease conformational sampling and dynamics. *The Journal of biological chemistry*, 289(24), pp.17203–14. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4059161&tool=pmcentrez&rendertype=abstract [Accessed June 2, 2015].

Jacks, T. et al., 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, 331(6153), pp.280–283.

Jiao, Y. et al., 2014. HIV-1 transmitted drug resistance-associated mutations and mutation co-variation in HIV-1 treatment-naïve MSM from 2011 to 2013 in Beijing, China. *BMC infectious diseases*, 14(1), p.689. Available at: http://www.biomedcentral.com/1471-2334/14/689 [Accessed September 2, 2015].

John, B. & Šali, A., 2003. Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Research*, 31(14), pp.3982–3992.

Jones, T.A. & Thirup, S., 1986. Using known substructures in protein model building and crystallography. *The EMBO journal*, 5(4), pp.819–822.

Judd, A. et al., 2014. Post-licensing safety of fosamprenavir in HIV-infected children in Europe. *Pharmacoepidemiology and Drug Safety*, 23(3), pp.321–325. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84895453894&partnerID=40&md5=592a92711f7bcd401b4bab59c5991fb5.

Kempf, D.J. et al., 1997. Pharmacokinetic enhancement of inhibitors of the human immunodeficiency virus protease by coadministration with ritonavir. *Antimicrobial Agents and Chemotherapy*, 41(3), pp.654–660.

King, N.M. et al., 2004. Structural and Thermodynamic Basis for the Binding of TMC114 , a Next-Generation Human Immunodeficiency Virus Type 1 Protease Inhibitor. *Journal of Virology*, 78(21), pp.12012–12021.

Klebe, G. (Editor), 2013a. *Drug Design* G. Klebe, ed., Springer Berlin Heidelberg. Available at: http://link.springer.com/10.1007/978-3-642-17907-5.

Klebe, G. (Editor), 2013b. *Drug Design: Methodology, Concepts and Mode-of-Action* G. Klebe, ed., Springer Berlin Heidelberg. Available at: http://link.springer.com/10.1007/978-3-642-17907-5.

Kolinski, A. et al., 2001. Generalized comparative modeling (GENECOMP): A combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins: Structure, Function and Genetics*, 44(2), pp.133–149.

Koudriakova, T. et al., 1998. Metabolism of the human immunodeficiency virus protease inhibitors indinavir and ritonavir by human intestinal microsomes and expressed cytochrome

P4503A4/3A5: Mechanism-based inactivation of cytochrome P4503A by ritonavir. *Drug Metabolism and Disposition*, 26(6), pp.552–561.

Krebs, F., Hogan, T. & Quiterio, S., 2001. Lentiviral LTR-directed expression, sequence variation, and disease pathogenesis. *HIV sequence ...*, (i), pp.29–70. Available at: http://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2001/partI/Wigdahl.pdf.

Krohn, A. et al., 1991. Novel binding mode of highly potent HIV-proteinase inhibitors incorporating the (R)-hydroxyethylamine isostere. *Journal of medicinal chemistry*, 34(11), pp.3340–3342. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1956054 [Accessed March 29, 2015].

Kuntz, I.D. et al., 1982. A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology*, 161(2), pp.269–288.

Laskowski, R.A. et al., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), pp.283–291.

Lefebvre, E. & Schiffer, C.A., 2008. Resilience to resistance of HIV-1 protease inhibitors: Profile of darunavir. *AIDS Reviews*, 10(3), pp.131–142.

Levitt, M., 1992. Accurate modeling of protein conformation by automatic segment matching. *Journal of molecular biology*, 226(2), pp.507–533.

Lihana, R.W. et al., 2012. Update on HIV-1 diversity in Africa: A decade in review. *AIDS Reviews*, 14, pp.83–100.

Liu, T.F. & Shafer, R.W., 2006. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 42(11), pp.1608–1618.

Los Alamos National Laboratory, 2012. HIV Sequence Database: HIV and SIV Nomenclature. Available at: http://www.hiv.lanl.gov/content/sequence/HelpDocs/subtypes-more.html [Accessed July 8, 2015].

Los Alamos National Laboratory, 2014. Landmarks of the HIV genome. Available at: http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html [Accessed April 12, 2015].

Louis, J.M. et al., 2011. 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*, 108(22), pp.9072–9077.

Lüthy, R., Bowie, J.U. & Eisenberg, D., 1992. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364), pp.83–85.

Lv, Z., Chu, Y. & Wang, Y., 2015. HIV protease inhibitors: a review of molecular selectivity and toxicity. *HIV/AIDS (Auckland, N.Z.)*, 7, pp.95–104. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4396582&tool=pmcentrez&rendertype=abstract [Accessed August 11, 2015].

Martyna, G.J. et al., 1996. Explicit reversible integrators for extended systems dynamics. *Molecular Physics*, 87(5), pp.1117–1157. Available at: citeulike-article-id:2219591%5Cnhttp://dx.doi.org/10.1080/00268979600100761%5Cnhttp://journalsonline.tandf.co.uk/Index/10.1080/00268979650027054.

Max, B. & Sherer, R., 2000. Management of the adverse effects of antiretroviral therapy and medication adherence. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 30 Suppl 2, pp.S96–S116.

Mcgee, D., 2010. *Molecular Dynamics Study of the Conformational Dynamics of HIV-1 Protease Subtypes A, B, C, and F*. University of Florida, Gainesville.

Meher, B.R. & Wang, Y., 2015. Exploring the drug resistance of V32I and M46L mutant HIV-1 protease to inhibitor TMC114: Flap dynamics and binding mechanism. *Journal of Molecular Graphics and Modelling*, 56, pp.60–73. Available at: http://dx.doi.org/10.1016/j.jmgm.2014.11.003.

Melikyan, G.B., 2008. Common principles and intermediates of viral protein-mediated fusion: the HIV-1 paradigm. *Retrovirology*, 5, p.111.

Melo, F. et al., 1997. ANOLEA: a www server to assess protein structures. *International Conference on Intelligent Systems for Molecular Biology*, 5, pp.187–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9322034 [Accessed October 31, 2015].

Messiaen, P. et al., 2013. Clinical Use of HIV Integrase Inhibitors: A Systematic Review and Meta-Analysis. *PLoS ONE*, 8(1).

Monno, L. et al., 2012. An outbreak of HIV-1 BC recombinants in Southern Italy. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology*, 55(4), pp.370–3. Available at: http://www.sciencedirect.com/science/article/pii/S1386653212003186.

Monticelli, L. & Tieleman, D.P., 2013. Force Fields for Classical Molecular Dynamics. In L. Monticelli & E. Salonen, eds. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 197–213. Available at: http://link.springer.com/10.1007/978-1-62703-017-5.

Morris, G. et al., 2012. User Guide AutoDock Version 4.2. , pp.1–66. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.363.3063&rep=rep1&type=pdf.

Morris, G.M. et al., 1996. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *Journal of computer-aided molecular design*, 10(4), pp.293–304.

Morris, G.M., Huey, R. & Olson, A.J., 2008. UNIT 8.14 using AutoDock for ligand-receptor docking. *Current Protocols in Bioinformatics*, (SUPPL. 24).

Mullard, A., 2014. 2013 FDA drug approvals. *Nature Reviews Drug Discovery*, 13(2), pp.85–89. Available at: http://www.nature.com/doifinder/10.1038/nrd4239.

Navia, M.A. et al., 1989. Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature*, 337(6208), pp.615–620.

NIAID, 2013. HIV/AIDS Antiretroviral Drugs Classes. Available at: http://www.niaid.nih.gov/topics/HIVAIDS/Understanding/Treatment/pages/arvdrugclasses.aspx [Accessed December 2, 2015].

Nolan, D., Reiss, P. & Mallal, S., 2005. Adverse effects of antiretroviral therapy for HIV infection: a review of selected topics. *Expert opinion on drug safety*. Available at: http://0-informahealthcare.com.wam.seals.ac.za/doi/abs/10.1517/14740338.4.2.201 [Accessed March 6, 2015].

Nosé, S. & Klein, M.L., 1983. Constant pressure molecular dynamics for molecular systems. *Molecular Physics*, 50(5), pp.1055–1076.

O'Boyle, N.M. et al., 2011. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1), p.33. Available at: http://www.jcheminf.com/content/3/1/33.

O'Meara, B.C., 2011. Evolutionary Inferences from Phylogenies: A Review of Methods. *Annual Review of Ecology, Evolution, and Systematics*, 43(1), p.120913143848009.

Özen, A., Haliloğlu, T. & Schiffer, C. a., 2011. Dynamics of preferential substrate recognition in HIV-1 protease: Redefining the substrate envelope. *Journal of Molecular Biology*, 410, pp.726–744.

Page, R.D. & Charleston, M. a, 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular phylogenetics and evolution*, 7(2), pp.231–240.

Patwardhan, A., Ray, S. & Roy, A., 2014. Molecular Markers in Phylogenetic Studies-A Review. *Journal of Phylogenetics & Evolutionary Biology*, 2(2), pp.1–9. Available at: http://esciencecentral.org/journals/molecular-markers-in-phylogenetic-studiesa-review-2329-9002-2-131.php?aid=30965.

Pessoa, L.S. et al., 2011. Genotypic analysis of the gp41 HR1 region from HIV-1 isolates from enfuvirtide-treated and untreated patients. *Journal of acquired immune deficiency syndromes (1999)*, 57 Suppl 3(3), pp.S197-201. Available at: http://online.liebertpub.com/doi/abs/10.1089/aid.2010.0057%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/2185731%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/21857318.

Pokorná, J. et al., 2009. Current and novel inhibitors of HIV protease. *Viruses*, 1, pp.1209–1239.

Ponder, J.W. & Case, D.A., 2003. Force Fields for Protein Simulations. *Advances in Protein Chemistry*, 66, pp.27–85. Available at: http://linkinghub.elsevier.com/retrieve/pii/S006532330366002X.

Prashar, V. et al., 2015. Structural basis of why nelfinavir-resistant D30N mutant of HIV-1 protease remains susceptible to saquinavir. *Chemical Biology and Drug Design*, 86(3), pp.302–308.

Purnell, J.Q. et al., 2000. *Effect of ritonavir on lipids and post-heparin lipase activities in normal subjects.*,

Rabson, A.B. & Graves, B.J., 1997. Synthesis and processing of viral RNA. In H. E. C. J. M. H. S. H. Varmus, ed. *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY). Available at: http://www.ncbi.nlm.nih.gov/books/NBK19367/.

R Core Team, 2015. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Available at: http://www.r-project.org/.

Ragland, D. a et al., 2014. Drug Resistance Conferred by Mutations Outside the Active Site through Alterations in the Dynamic and Structural Ensemble of HIV-1 Protease. *Journal of the American Chemical Society*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25091085.

Rapaport, D.C.C., 2004. *The Art of Molecular Dynamics Simulation*, Available at: http://books.google.com/books?hl=en&lr=&id=iqDJ2hjqBMEC&oi=fnd&pg=PR9&dq=The+art+of+molecular+dynamics+simulation&ots=krIRqBdl2O&sig=l4zaeQ2Snz4KDWO2yFaqpKuRD4c%5Cnhttp://www.amazon.fr/Art-Molecular-Dynamics-Simulation/dp/0521825.

Read, R.J. et al., 2011. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure*, 19(10), pp.1395–1412. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0969212611002851.

Rock, B.M. et al., 2014. Characterization of Ritonavir-Mediated Inactivation of Cytochrome P450 3A4. *Molecular Pharmacology*, 86(6), pp.665–674. Available at: http://molpharm.aspetjournals.org/cgi/doi/10.1124/mol.114.094862.

Saitou, N. & Nei, M., 1987. The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular biology and evolution*, 4(4), pp.406–425. Available at: http://mbe.oxfordjournals.org/content/4/4/406.short.

Salentin, S. et al., 2015. PLIP: fully automated protein-ligand interaction profiler. *Nucleic Acids Research*, 43(W1), pp.W443–W447. Available at: http://nar.oxfordjournals.org/content/43/W1/W443.

Šali, A., 2013. MODELLER: A Program for Protein Structure Modeling Release 9.12, r9480. *Rockefeller University*. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:MODELLER+A+Program+for+Protein+Structure+Modeling#6.

Šali, A. & Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3), pp.779–815. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022283683716268.

Samudrala, R. et al., 1999. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins: Structure, Function, & Bioinformatics*, Suppl. 3(May), pp.194–198.

Sánchez, R. & Sali, a, 1997. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins*, Suppl 1, pp.50–58.

Sarafianos, S.G. et al., 2009. Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. *Journal of Molecular Biology*, 385(3), pp.693–713.

Schrödinger, L., 2010. *The PyMOL Molecular Graphics System, Version~1.3r1*,

Schüttelkopf, A.W. & Van Aalten, D.M.F., 2004. PRODRG: A tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica Section D: Biological Crystallography*, 60(8), pp.1355–1363.

Scott, W.R.P. & Schiffer, C. a., 2000. Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. *Structure*, 8(12), pp.1259–1265.

Sham, H.L. et al., 1998. ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. *Antimicrobial agents and chemotherapy*, 42(12), pp.3218–24. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=106025&tool=pmcentrez&rendertype=abstract [Accessed August 12, 2015].

Sharp, P. M. and Hahn, B.H., 2011. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor Perspectives in Medicine*, 1(1).

Sharp, P.M. & Hahn, B.H., 2010. The evolution of HIV-1 and the origin of AIDS. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365, pp.2487–2494.

Shen, M.-Y. & Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. *Protein Science: A Publication of the Protein Society*, 15(11), pp.2507–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17075131%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/17075131?ordinalpos=1&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum.

Shibuyama, S. et al., 2006. Understanding and avoiding antiretroviral adverse events. *Current pharmaceutical design*, 12(9), pp.1075–1090. Available at: http://0-www.ingentaconnect.com.wam.seals.ac.za/content/ben/cpd/2006/00000012/00000009/art00006 [Accessed March 6, 2015].

Sippl, M.J., 1993. Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function and Genetics*, 17(4), pp.355–362.

Somer, F.L., 2004. Molecular Modelling for Beginners (Alan Hinchliffe). *Journal of Chemical Education*, 81(11), p.1573. Available at: http://dx.doi.org/10.1021/ed081p1573.

Sousa da Silva, A.W. & Vranken, W.F., 2012. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC research notes*, 5(1), p.367. Available at: http://www.biomedcentral.com/1756-0500/5/367 [Accessed January 11, 2015].

van der Spoel, D. et al., 2010. Gromacs User Manual version 4.5.6. In *SpringerReference*. Berlin/Heidelberg: Springer-Verlag, pp. 77–78. Available at: www.gromacs.org.

St Clair, M.H. et al., 1996. In vitro antiviral activity of 141W94 (VX-478) in combination with other antiretroviral agents. *Antiviral research*, 29(1), pp.53–56.

Suguna, K. et al., 1987. Binding of a reduced peptide inhibitor to the aspartic proteinase from Rhizopus chinensis: implications for a mechanism of action. *Proceedings of the National Academy of Sciences of the United States of America*, 84(20), pp.7009–7013.

Sundquist, W.I. & Kräusslich, H.G., 2012. HIV-1 assembly, budding, and maturation. *Cold Spring Harbor Perspectives in Medicine*, 2(7).

Suzuki, Y. & Craigie, R., 2007. The road to chromatin - nuclear entry of retroviruses. *Nature reviews. Microbiology*, 5(3), pp.187–196.

Swanstrom, R. & Wills, J.W., 1997. Synthesis, Assembly, and Processing of Viral Proteins. In *Retroviruses*. pp. 263–334. Available at: http://www.ncbi.nlm.nih.gov/books/NBK19456/.

Tamura, K. et al., 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12), pp.2725–2729.

Tongo, M. et al., 2015. Phylogenetics of HIV-1 subtype G env: Greater complexity and older origins than previously reported. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*.

Torbeev, V.Y. et al., 2011. Protein conformational dynamics in the mechanism of HIV-1 protease catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(52), pp.20982–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3248522&tool=pmcentrez&rendertype=abstract [Accessed February 22, 2015].

Trott, O. & Olson, A.J., 2009. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), p.NA-NA. Available at: http://doi.wiley.com/10.1002/jcc.21334.

Turner, B.G. & Summers, M.F., 1999. Structural biology of HIV. *Journal of molecular biology*, 285, pp.1–32.

UNAIDS, 2014. South Africa | UNAIDS. Available at: http://www.unaids.org/en/regionscountries/countries/southafrica [Accessed November 9, 2015].

Unger, R. et al., 1989. A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5(4), pp.355–373.

US Food and Drug Administration, 2005. AGENERASE® (amprenavir) Capsules. Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2005/021007s017lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2009a. APTIVUS® (tipranavir) capsules, APTIVUS (tipranavir) oral solution. Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2009/021814s006,022292s001lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2014. CRIXIVAN ® (indinavir sulfate) capsules. Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2014/020685s076lbl.pdf [Accessed September 22, 2016].

US Food and Drug Administration, 2010a. KALETRA® (lopinavir/ritonavir) capsules, (lopinavir/ritonavir) oral solution. Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/021226s030lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2009b. LEXIVA® (fosamprenavir calcium) Tablets and Oral Suspension. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/021548s021,022116s005lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2011a. NORVIR® (ritonavir) Capsules Soft Gelatin, (ritonavir) Oral Solution. Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/020945s032lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2008. PREZISTA™ (Tibotec, Inc.) (darunavir). Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2008/021976s003s004lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2010b. ROCHE, INVIRASE® (saquinavir) Capsules and tablets. Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2010/020628s032,021785s009lbl.pdf [Accessed August 26, 2016].

US Food and Drug Administration, 2011b. VIRACEPT ® (nelfinavir mesylate). Available at: http://www.accessdata.fda.gov/drugsatfda_docs/label/2011/020778s035,020779s056,021503s017lbl.pdf [Accessed September 22, 2016].

Vandegraaff, N. & Engelman, A., 2007. Molecular mechanisms of HIV integration and therapeutic intervention. *Expert reviews in molecular medicine*, 9(6), pp.1–19.

Vanommeslaeghe, K. et al., 2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4), pp.671–690.

Velázquez-Campoy, A. et al., 2003. Protease inhibition in African subtypes of HIV-1. *AIDS Reviews*, 5(3), pp.165–171.

Vyas, V.K. et al., 2012. Homology modeling a fast tool for drug discovery: current perspectives. *Indian journal of pharmaceutical sciences*, 74(1), pp.1–17. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3507339&tool=pmcentrez&rendertype=abstract [Accessed February 23, 2015].

Wang, J. et al., 2006. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2), pp.247–260.

Wang, R., 2003. User manual for X-Score. Available at: http://sw16.im.med.umich.edu/software/xtool/manual/index.html [Accessed October 29, 2015].

Wang, R., Lu, Y. & Wang, S., 2003. Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry*, 46(12), pp.2287–2303.

Wang, Y., 2012. *Drug resistance mechanisms and drug design strategies for human immunodeficiency virus and hepatitis c virus proteases*. Wayne State University, Michigan. Available at: http://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1483&context=oa_dissertations.

Warrilow, D. & Harrich, D., 2007. HIV-1 replication from after cell entry to the nuclear periphery. *Curr HIV Res*, 5(1873–4251 (Electronic)), pp.293–299.

Waterhouse, A.M. et al., 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9), pp.1189–1191.

Wensing, A.M. et al., 2014. 2014 update of the drug resistance mutations in HIV-1. *Topics in Antiviral Medicine*, 22(3), pp.642–650.

Wensing, a. M.J., van Maarseveen, N.M. & Nijhuis, M., 2010. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Research*, 85, pp.59–74.

Wlodawer, A. et al., 1989. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science (New York, N.Y.)*, 245(4918), pp.616–621.

Wohl, D.A. et al., 2006. Current concepts in the diagnosis and management of metabolic complications of HIV infection and its therapy. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 43(5), pp.645–653. Available at: http://0-cid.oxfordjournals.org.wam.seals.ac.za/content/43/5/645.short [Accessed March 6, 2015].

Wu, T.D. et al., 2003. Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different Protease Inhibitor Treatments Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different P. *Journal of virology*, 77(8), pp.4836–4847.

wwPDB, 2014. User guide to the wwPDB X-ray validation reports. Available at: http://www.wwpdb.org/validation/ValidationPDFNotes.html [Accessed October 31, 2015].

Xiong, J., 2006. *Essential Bioinformatics*, Available at: http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

Yang, C. et al., 2000. Phylogenetic Analysis of Protease and Transmembrane Region of HIV Type 1 Group O. *AIDS Research and Human Retroviruses*, 16(11), pp.1075–1081.

Yang, Z. & Rannala, B., 2012. Molecular phylogenetics: principles and practice. *Nat Rev*, 13(5), pp.303–314. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22456349.

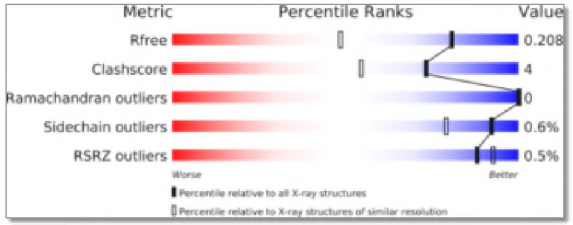
Yu, L. et al., 1999. Vitamin E-TPGS increases absorption flux of an HIV protease inhibitor by enhancing its solubility and permeability. *Pharm Res*, 16(12), pp.1812–1817.

Zhang, C., Liu, S. & Zhou, Y., 2004. Accurate and efficient loop selections by the DFIRE based all atom statistical potential. *Protein science*, pp.391–399. Available at: http://onlinelibrary.wiley.com/doi/10.1110/ps.03411904/full.

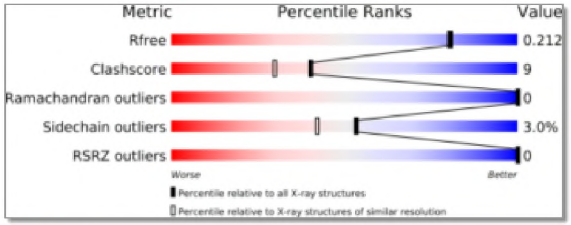
Zhou, H. & Zhou, Y., 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein science : a publication of the Protein Society*, 11(11), pp.2714–26. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2373736&tool=pmcentrez&rendertype=abstract.

Supplementary Materials

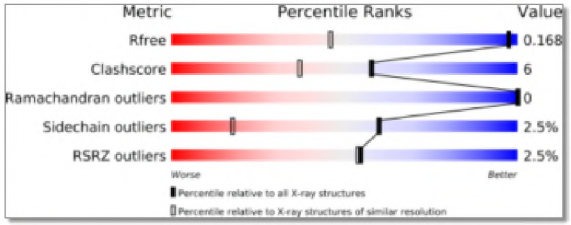
1. Multi-percentile validation reports of possible templates retrieved from PDB



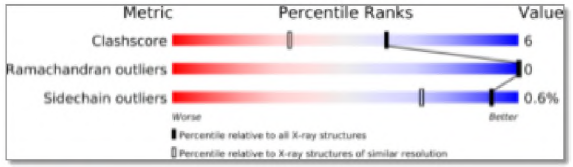
3pwm.png



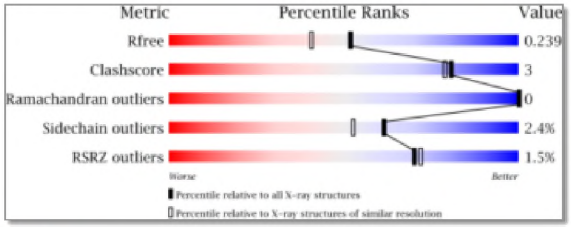
2wl0.png



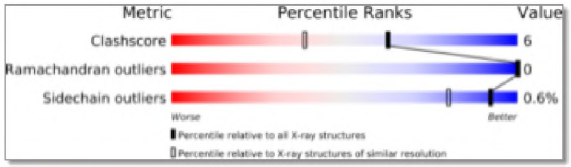
3bva.png



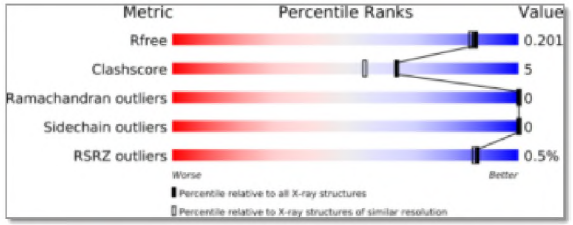
2aoc.png



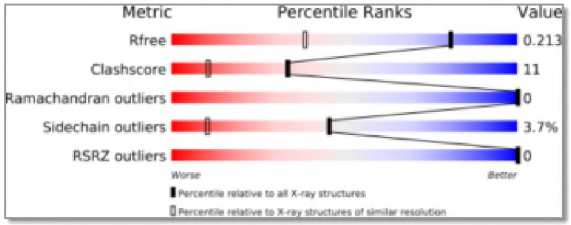
4qgi.png



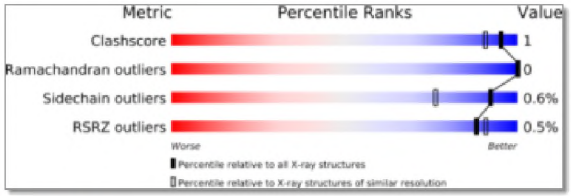
2i4v.png



2fxe.png

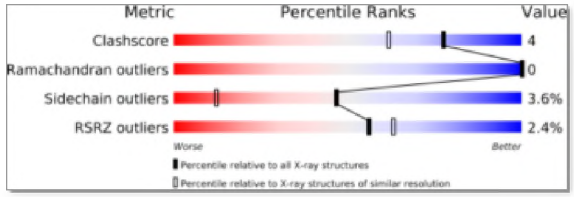


3cyw.png

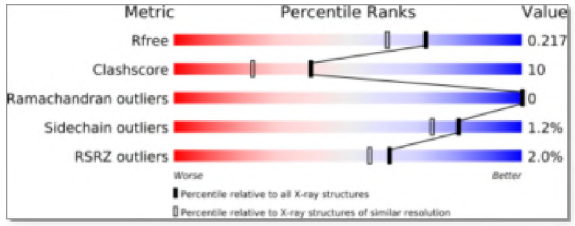


1sdv.png

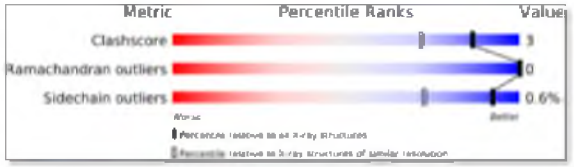
Supplementary Figure 1. Multi-percentile validation report summaries (filtered by: 2A resolution, 2 chains, no missing residues)



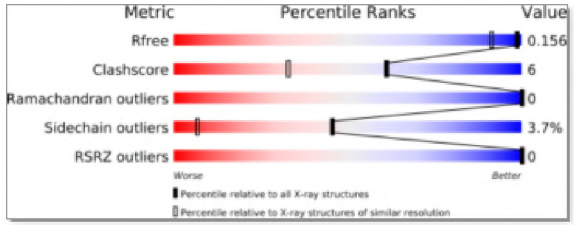
3b80.png



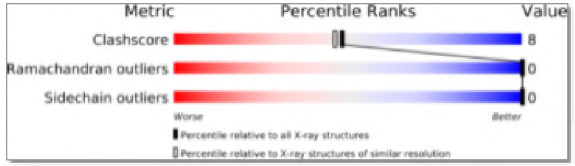
3jw2.png



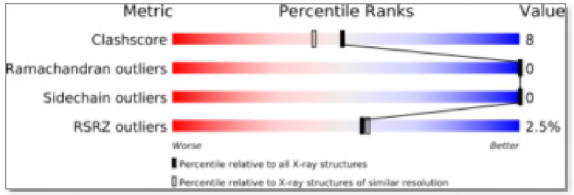
2o4l.png



2avq.png



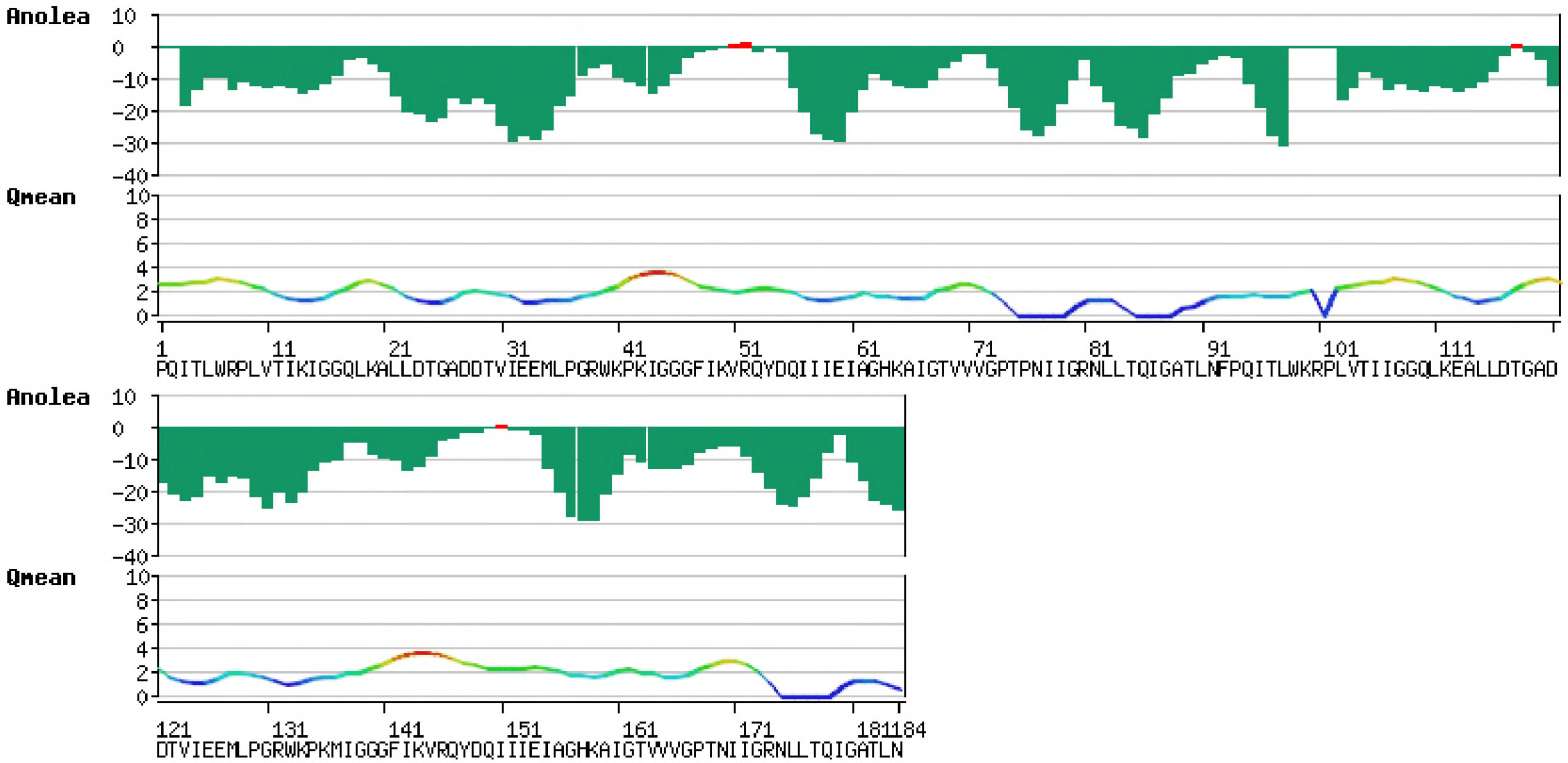
1mrx.png



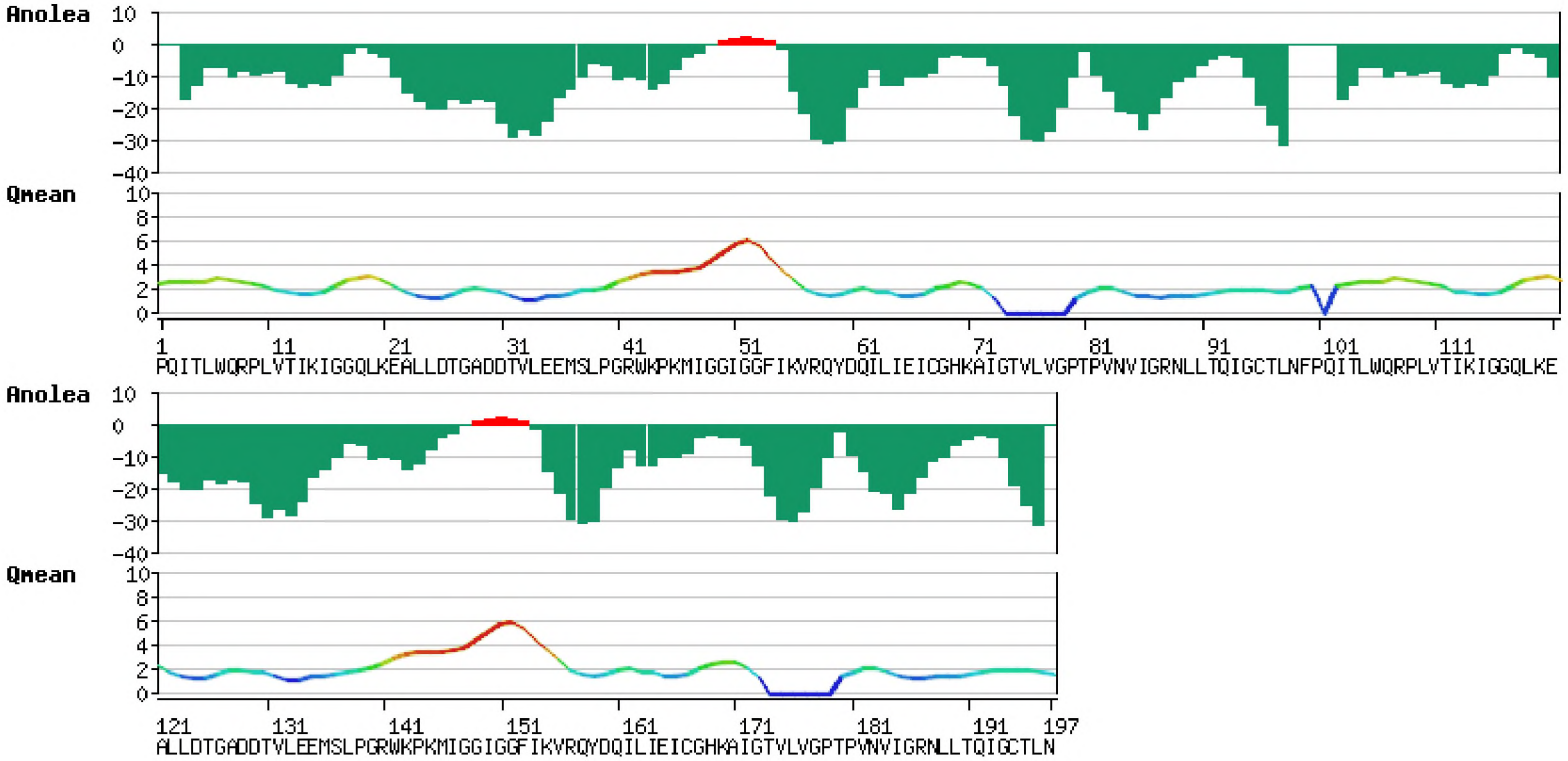
2fle.png

Supplementary Figure 2. Multi-percentile validation report summaries (filtered by: 2A resolution, 2 chains, no missing residues)

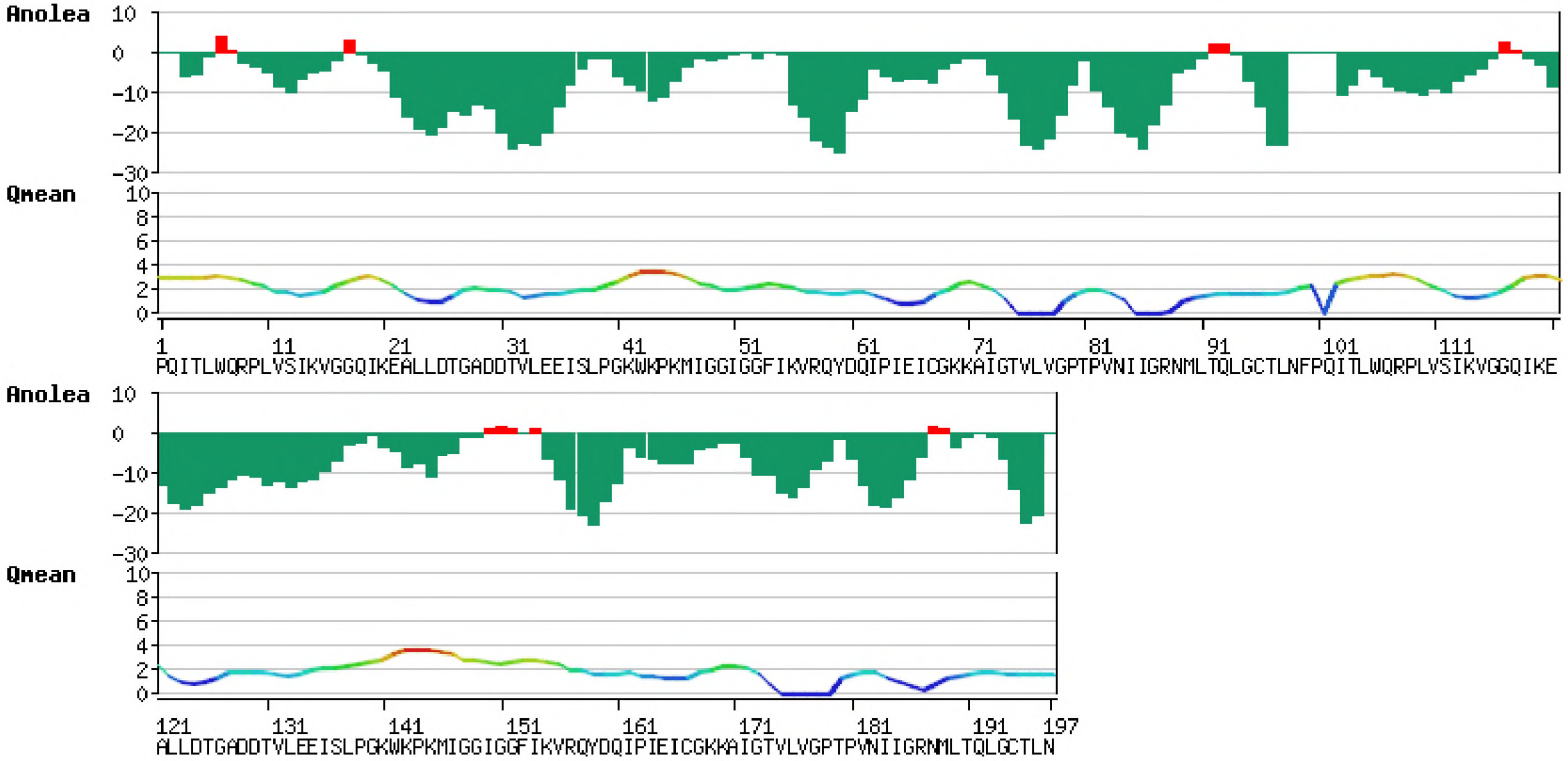
2. Template and model evaluation results from SWISSMODEL: ANOLEA and QMEAN scores



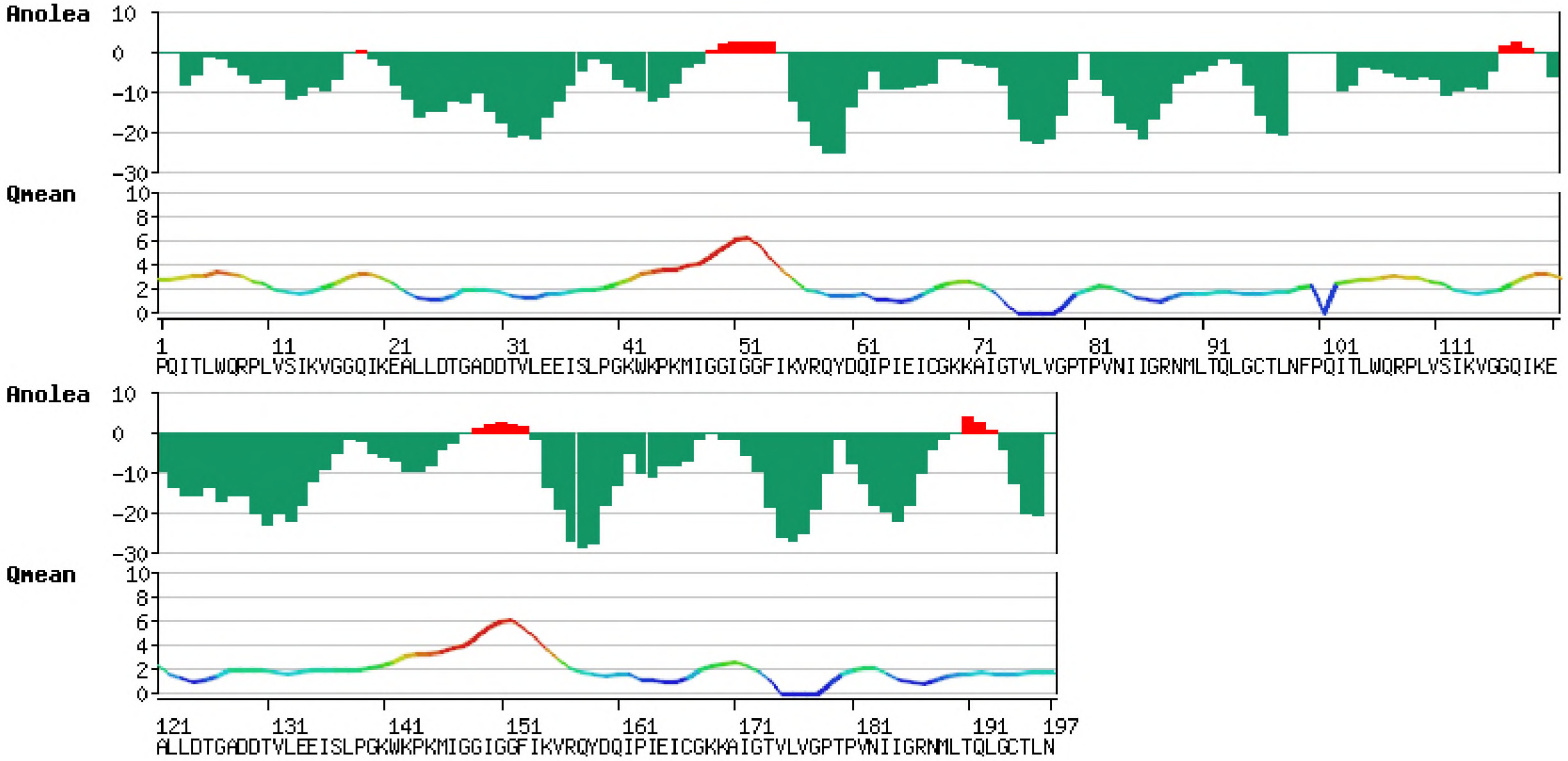
Supplementary Figure 3. ANOLEA and QMEAN score plots for open conformation template.



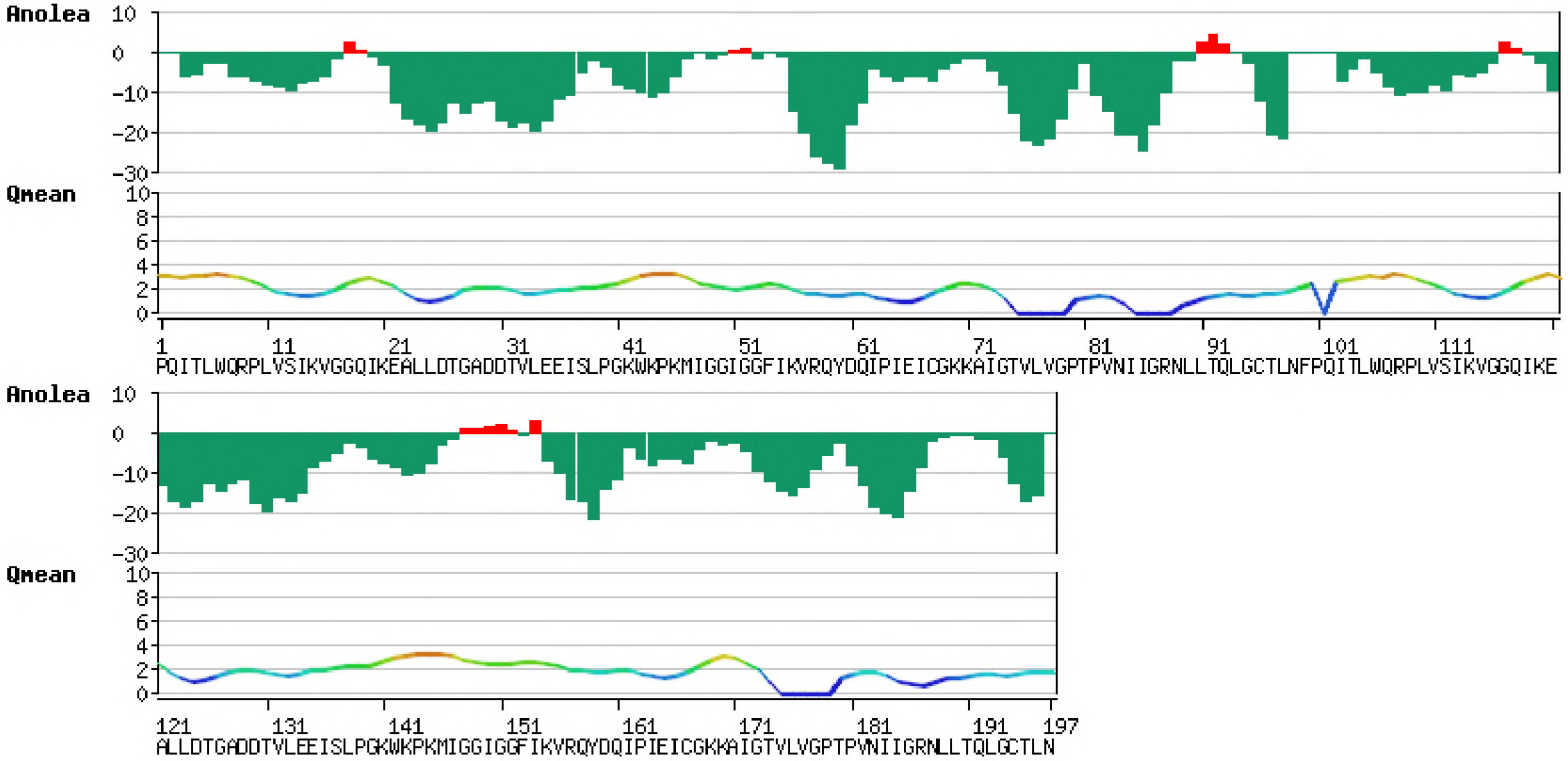
Supplementary Figure 4. ANOLEA and QMEAN score plots for closed conformation template.



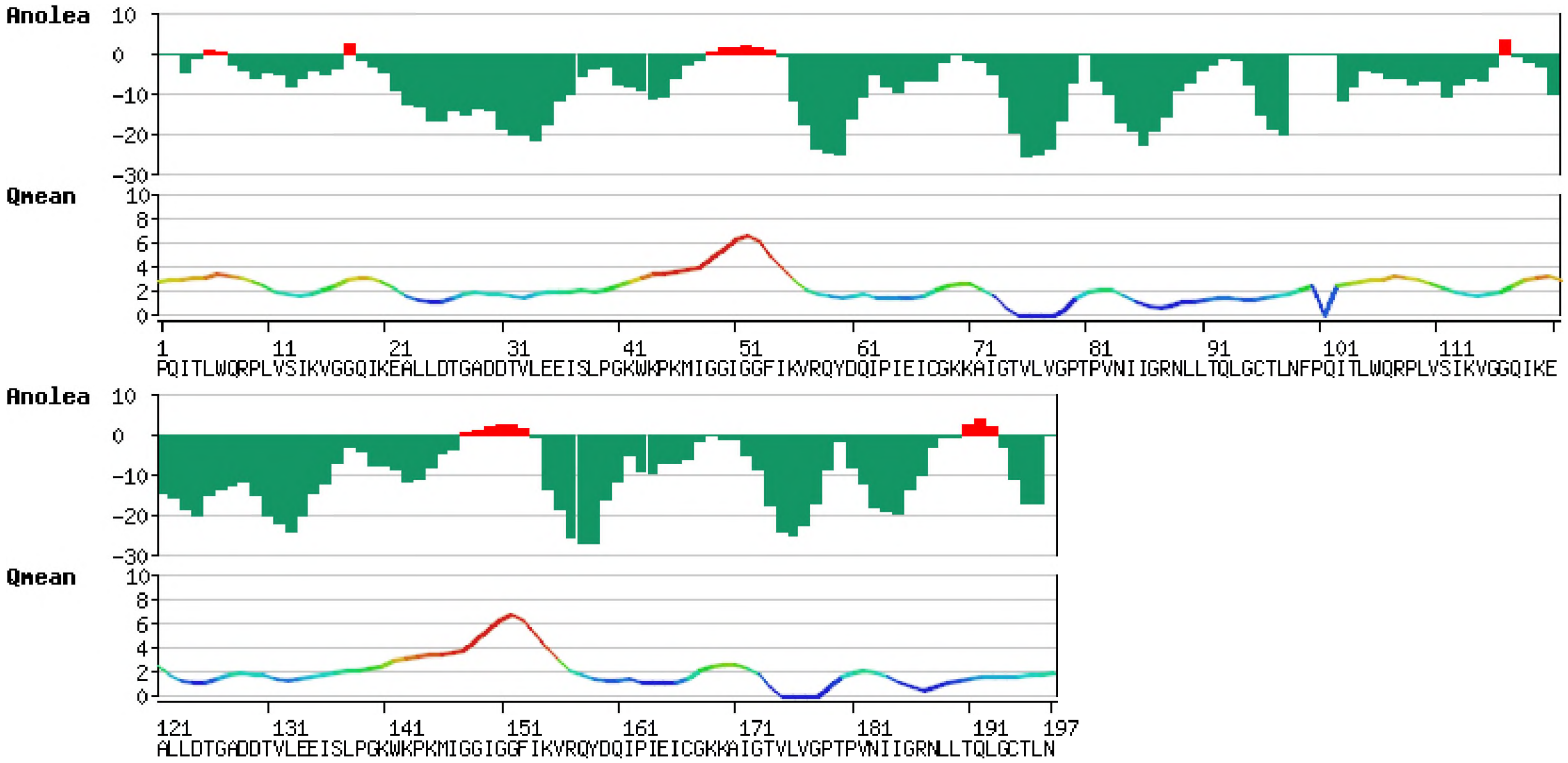
Supplementary Figure 5. ANOLEA and QMEAN score plots for model 1.



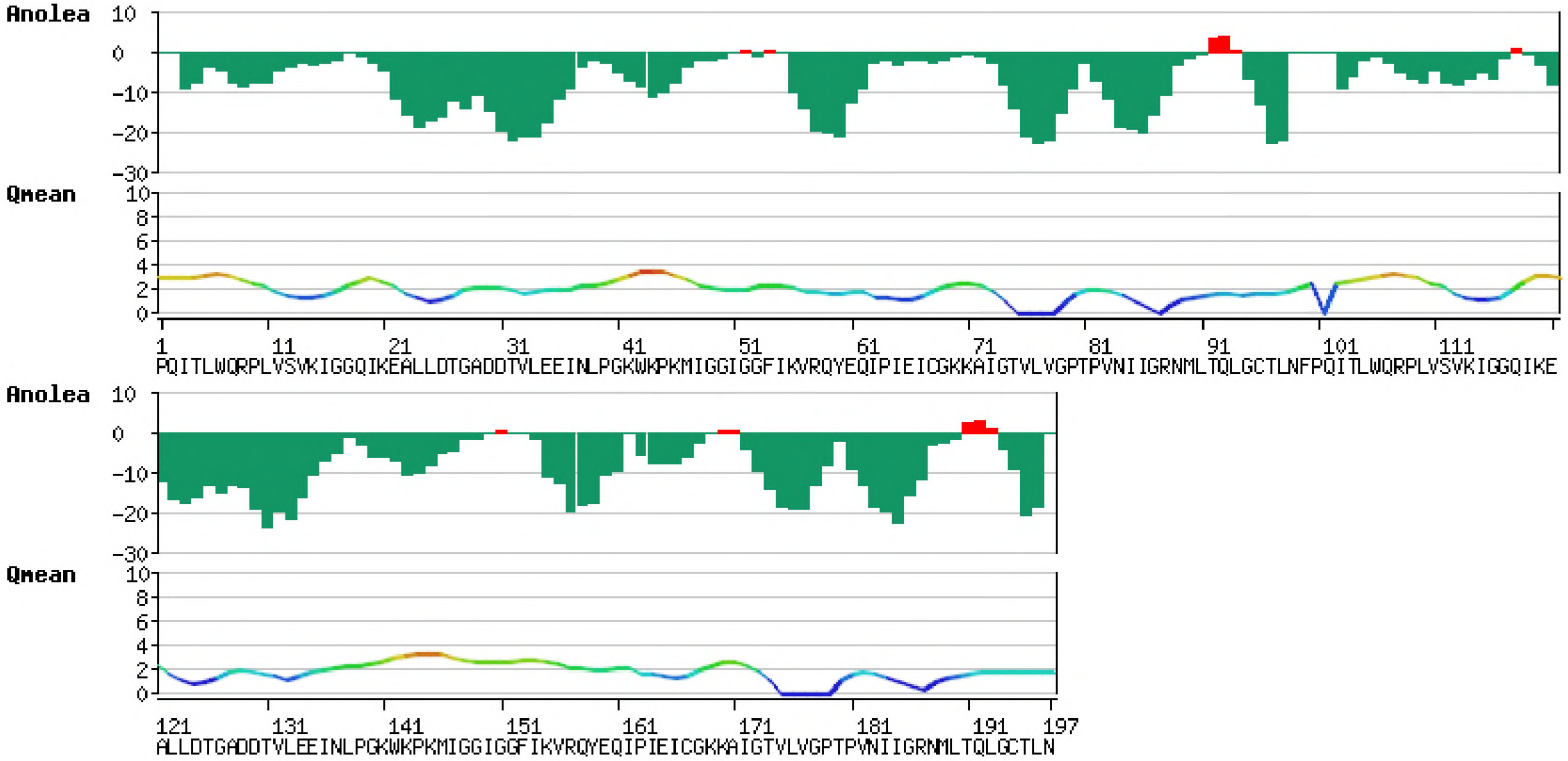
Supplementary Figure 6. ANOLEA and QMEAN score plots for model 2.



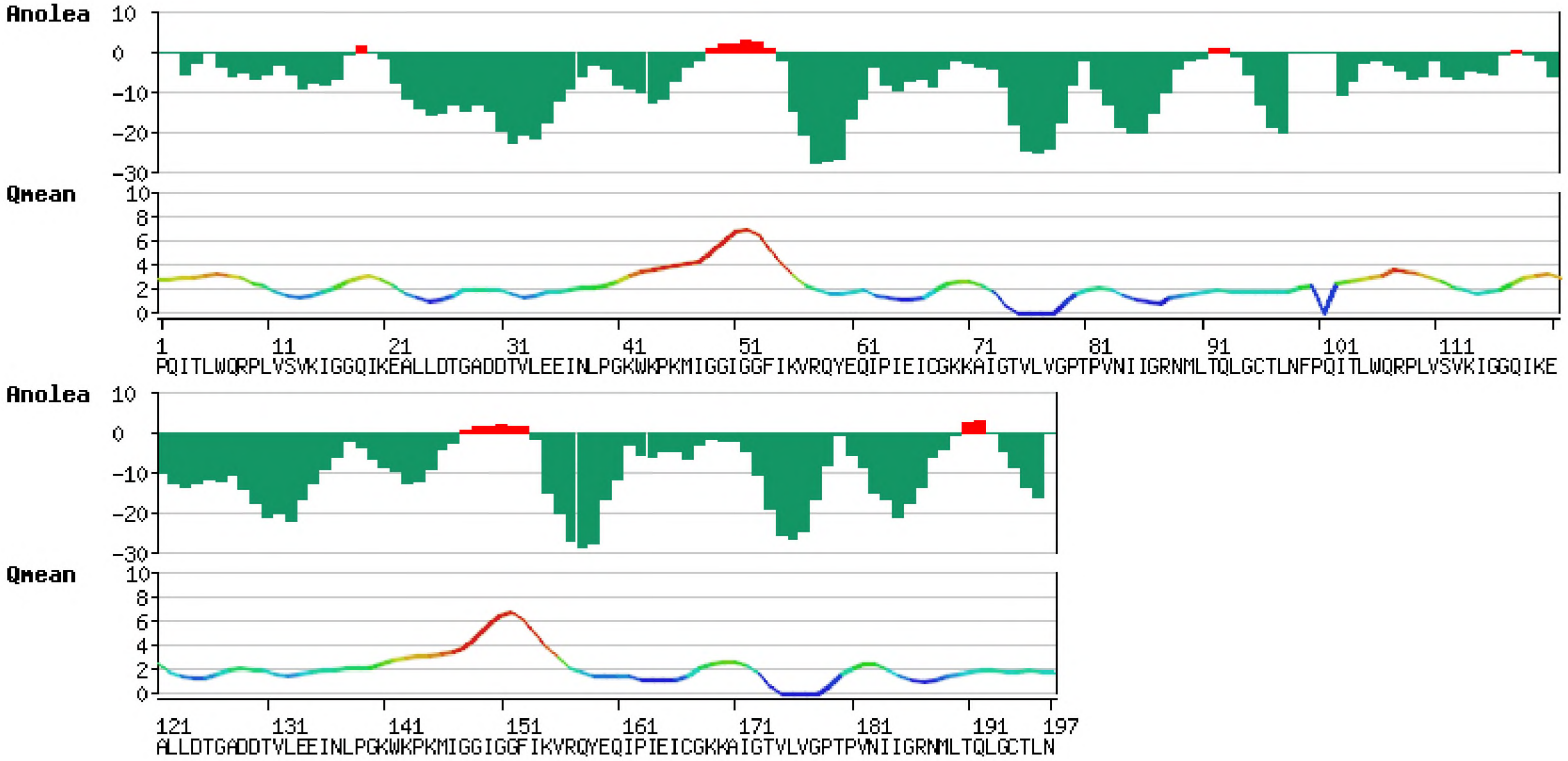
Supplementary Figure 7. ANOLEA and QMEAN score plots for model 3.



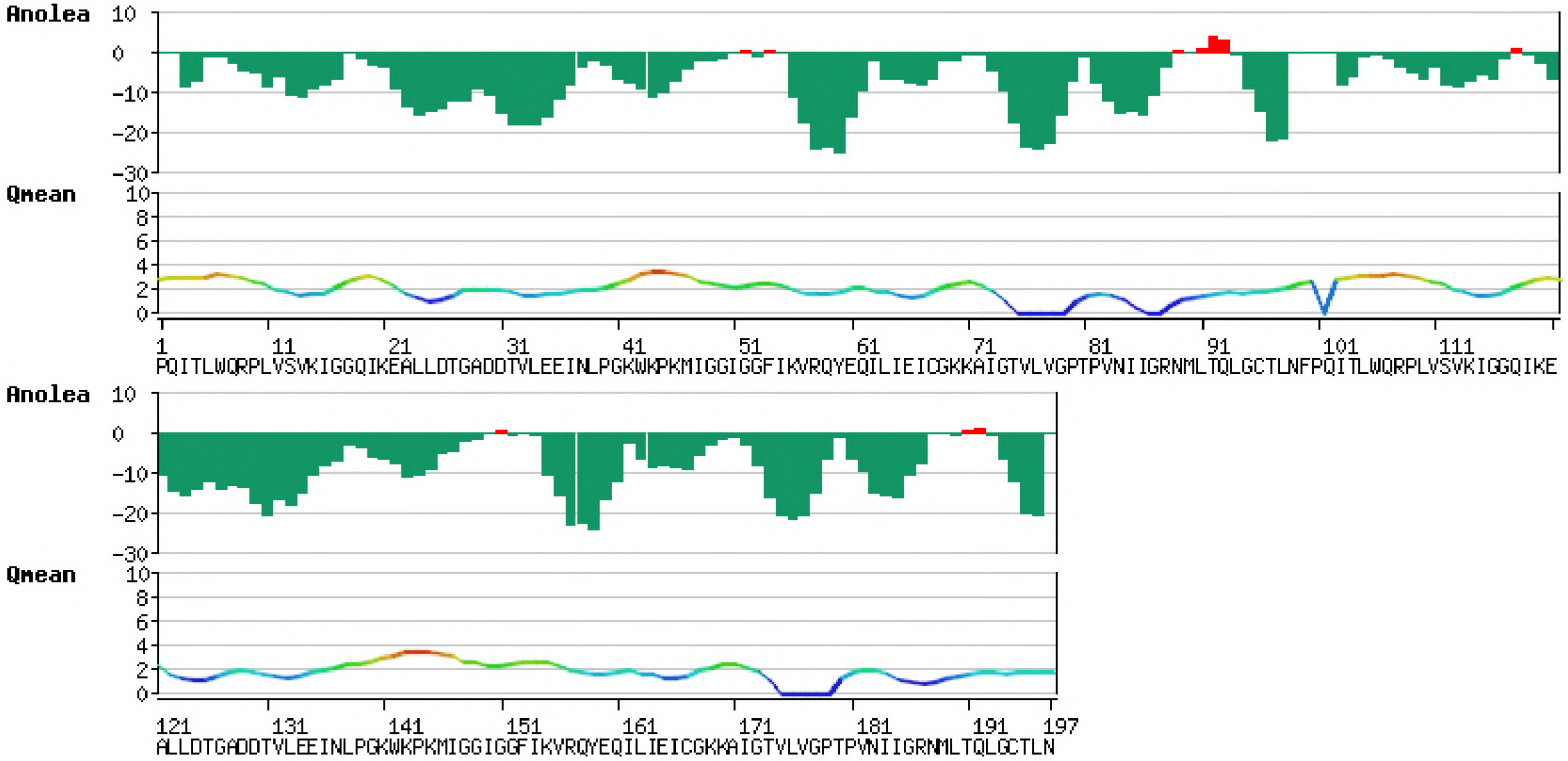
Supplementary Figure 8. ANOLEA and QMEAN score plots for model 4.



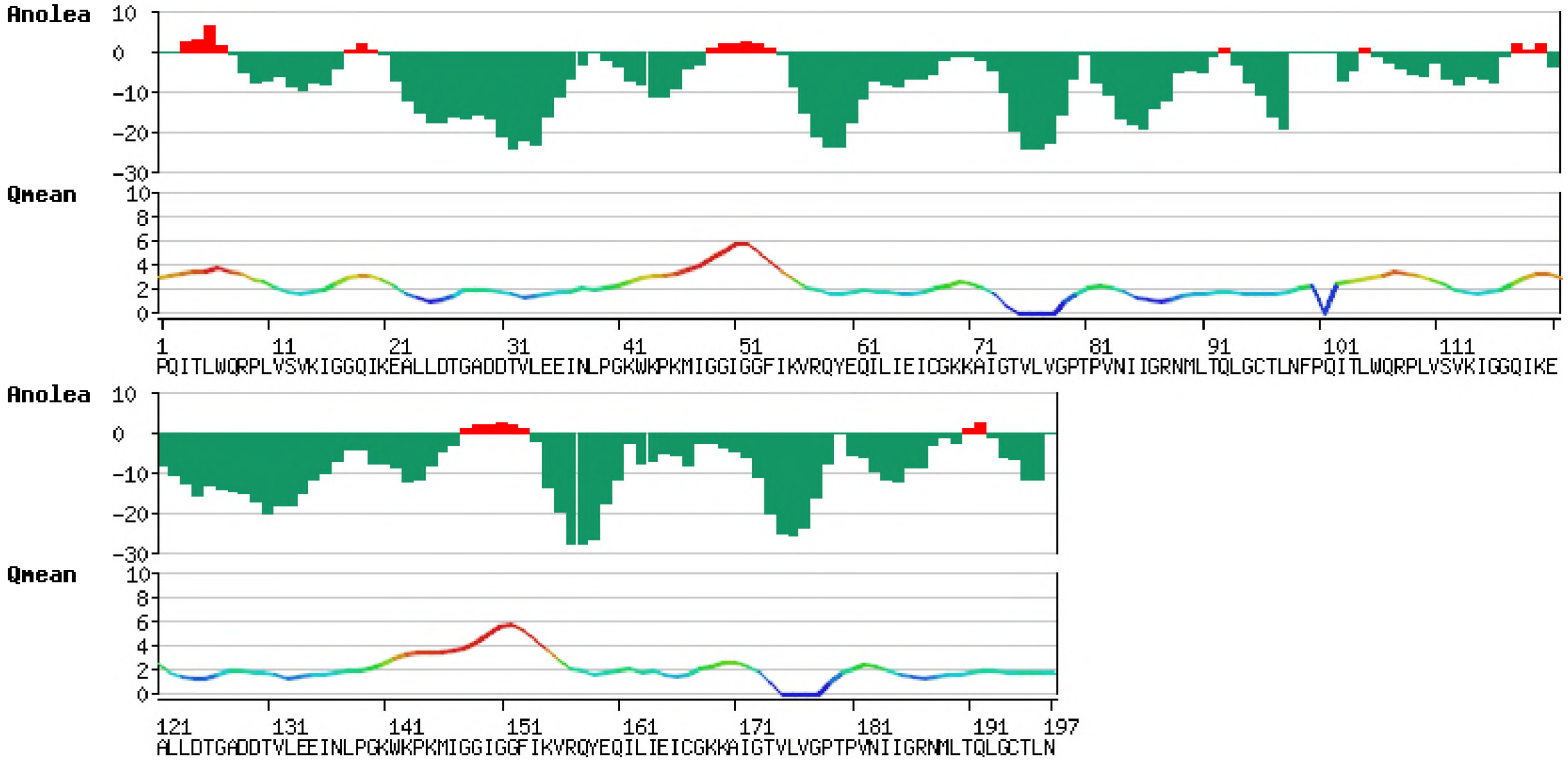
Supplementary Figure 9. ANOLEA and QMEAN score plots for model 5.



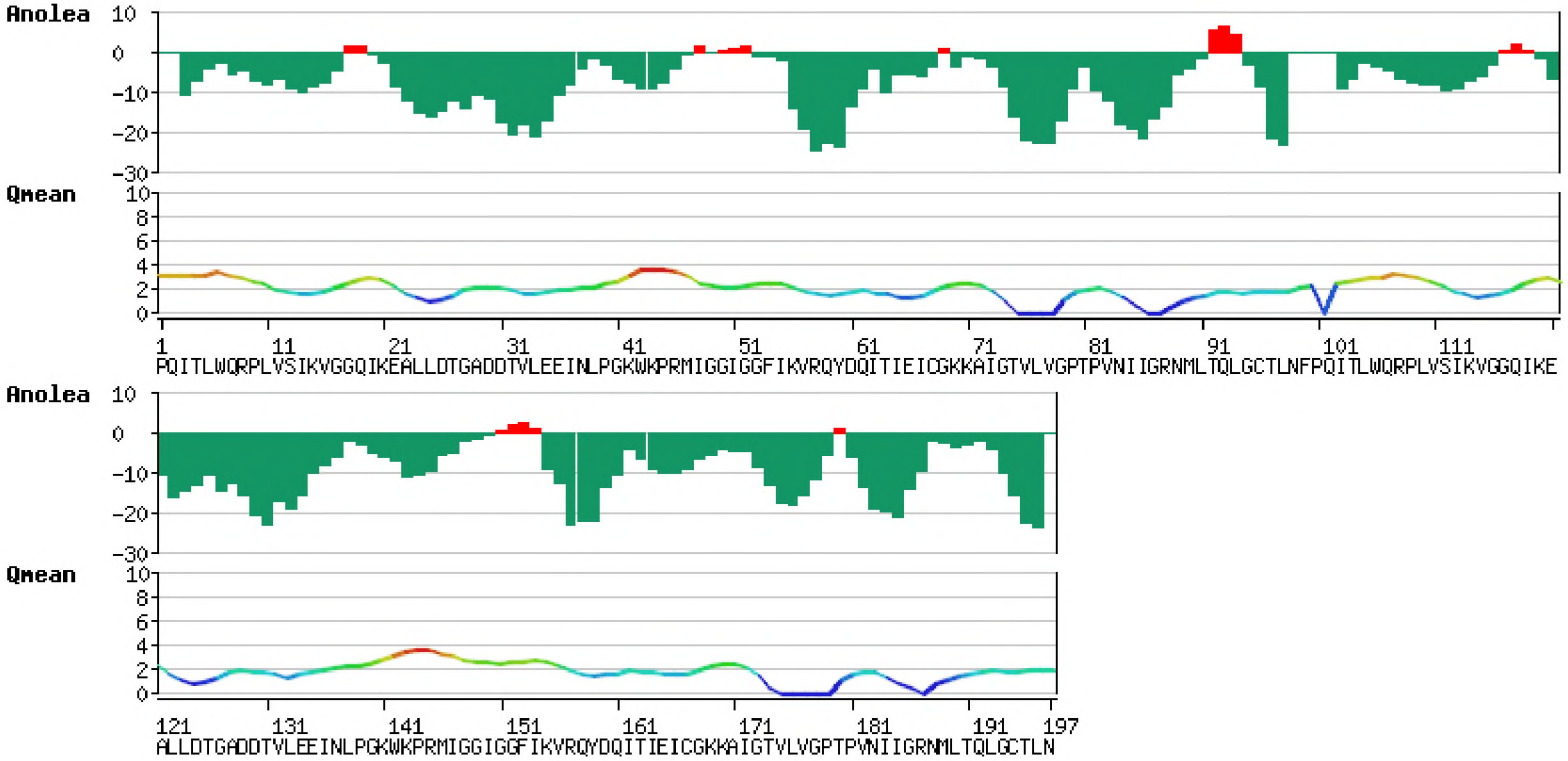
Supplementary Figure 10. ANOLEA and QMEAN score plots for model 6.



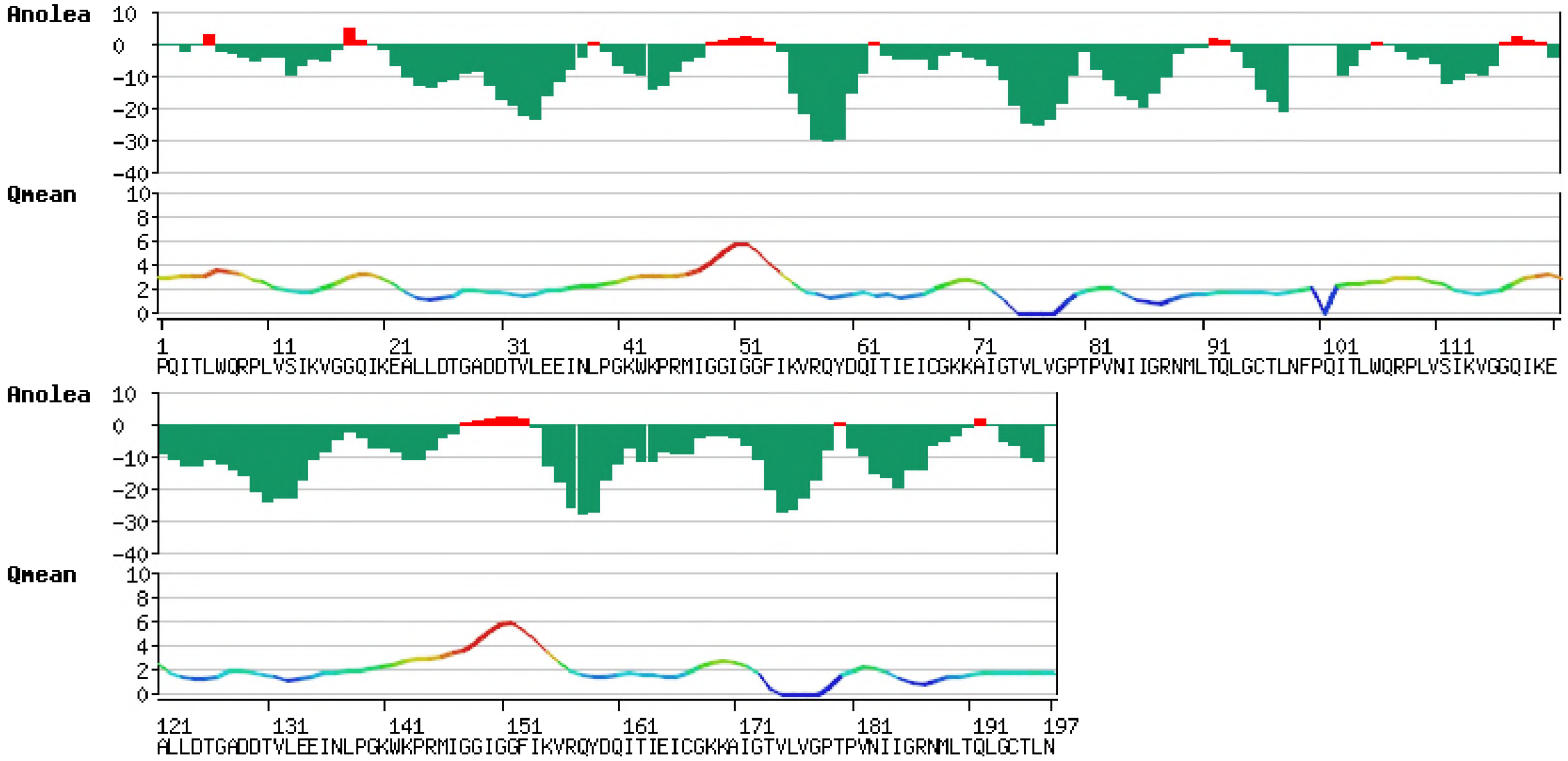
Supplementary Figure 11. ANOLEA and QMEAN score plots for model 7.



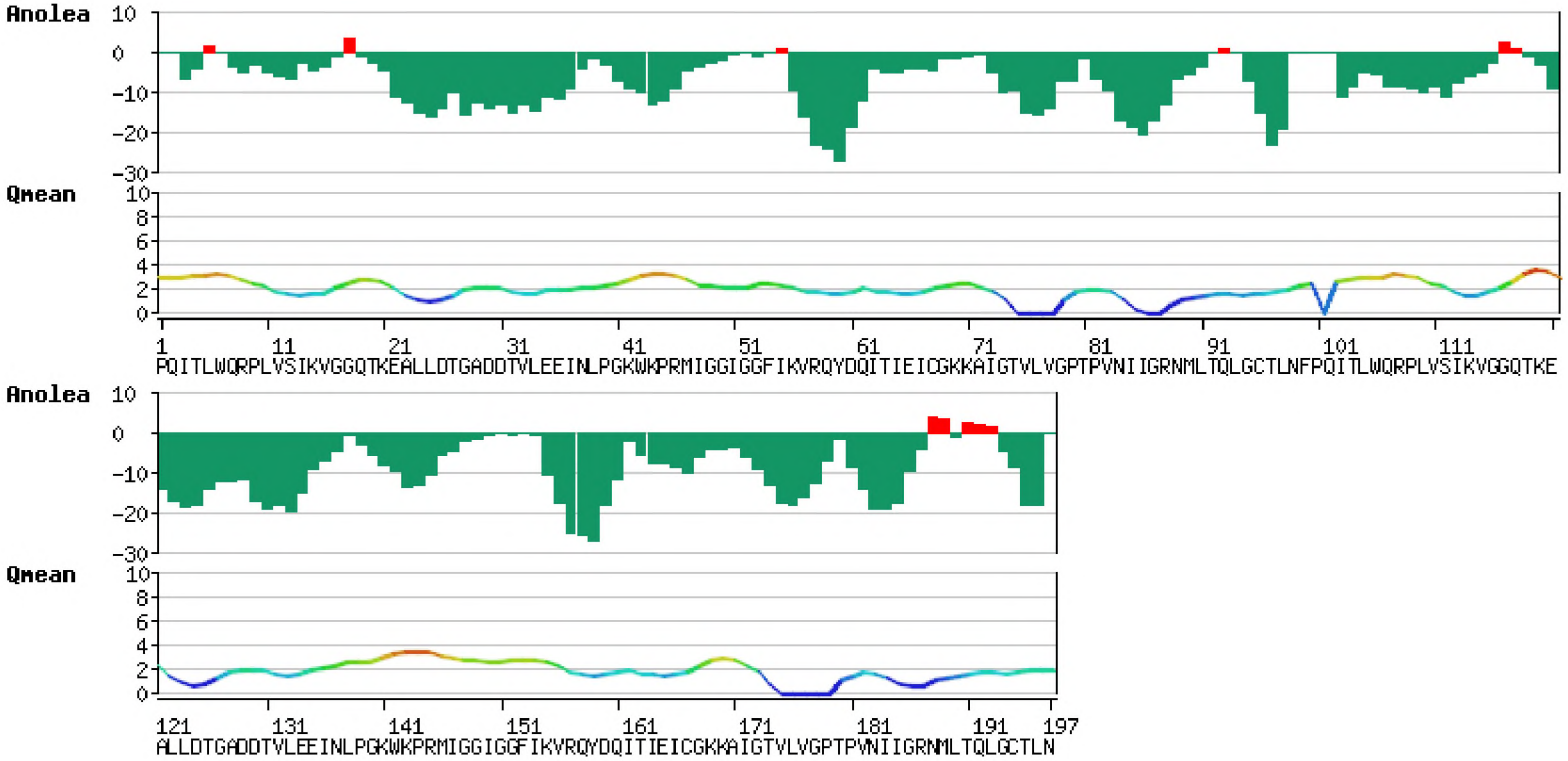
Supplementary Figure 12. ANOLEA and QMEAN score plots for model 8.



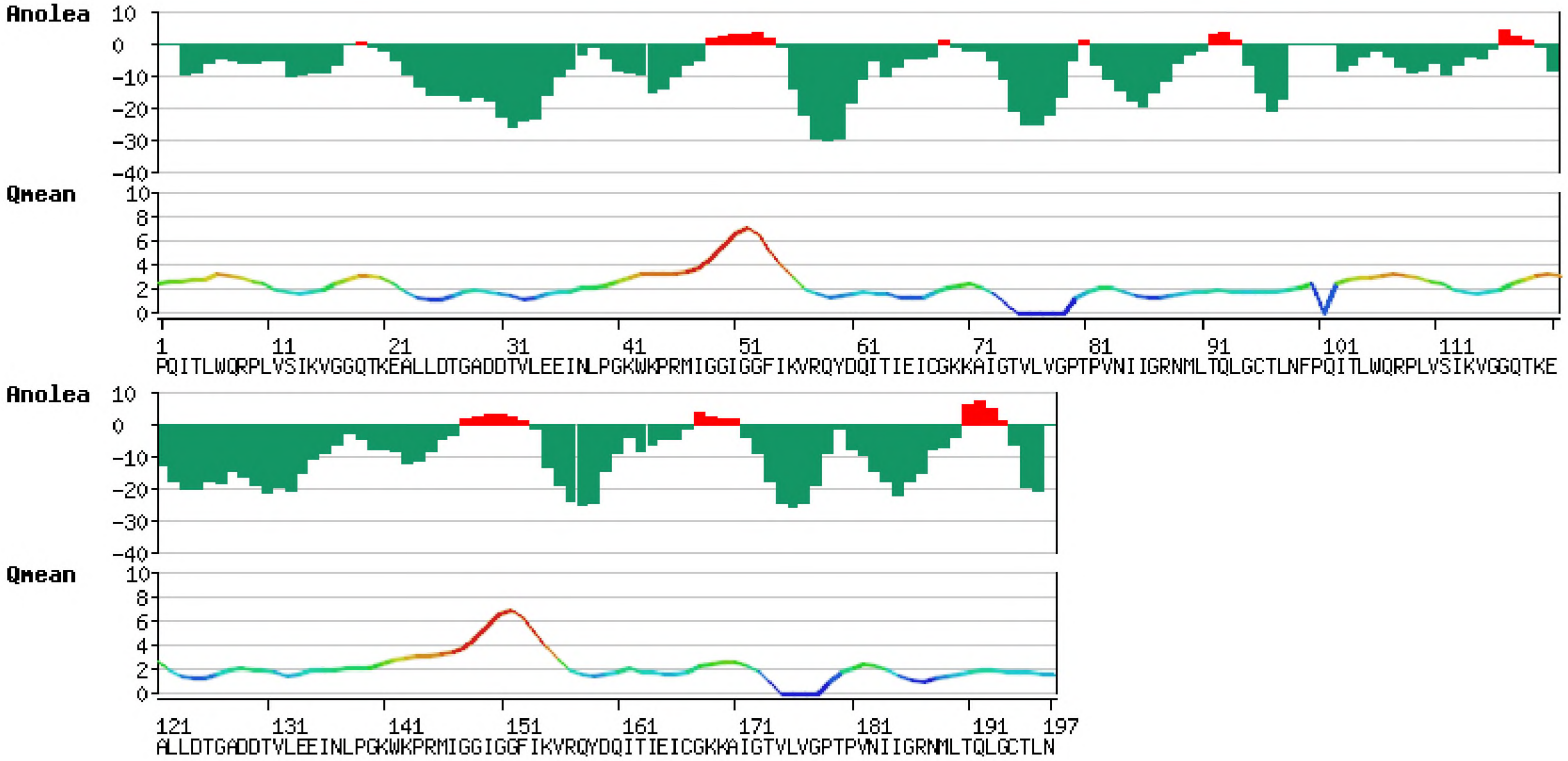
Supplementary Figure 13. ANOLEA and QMEAN score plots for model 9.



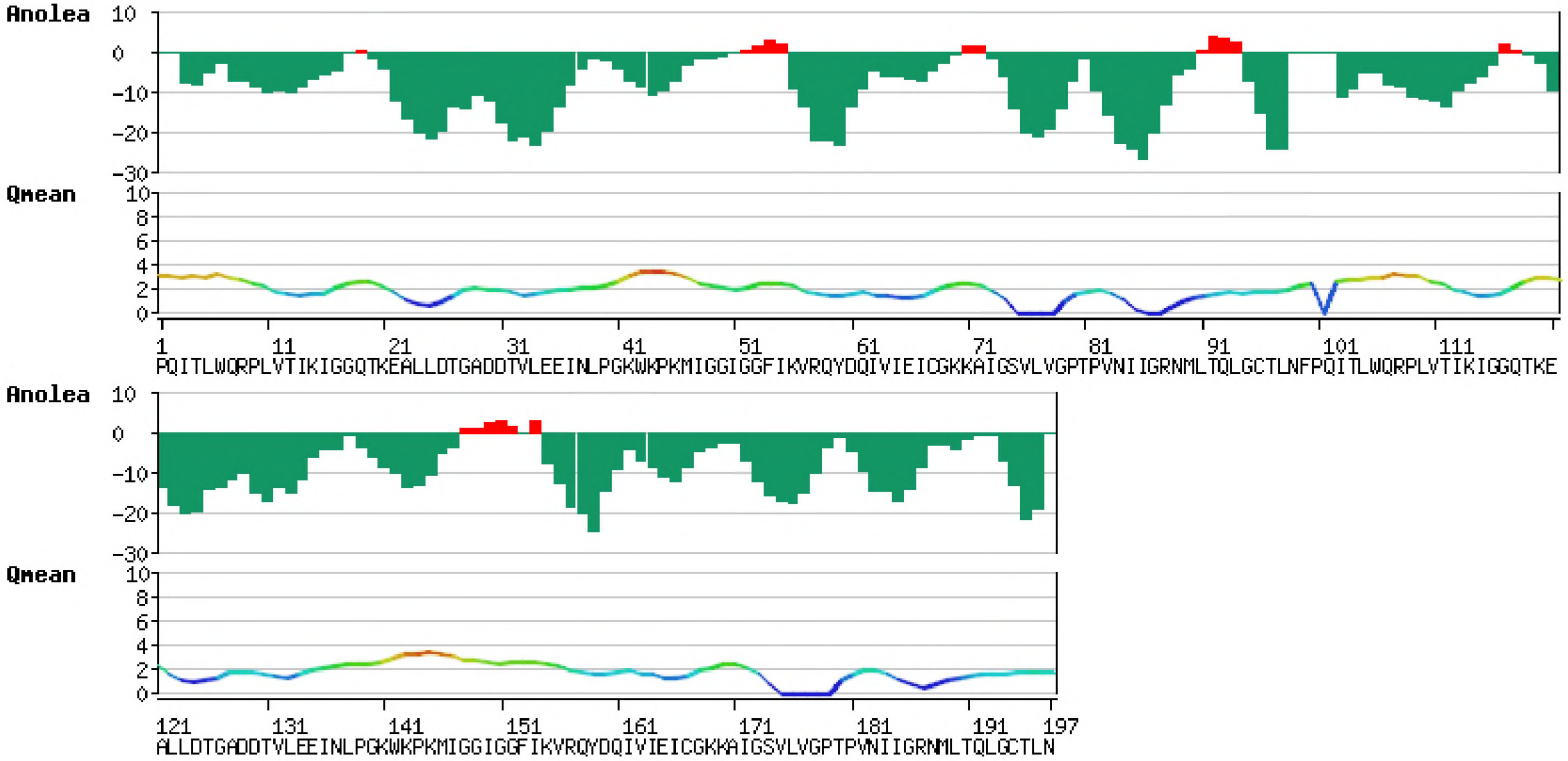
Supplementary Figure 14. ANOLEA and QMEAN score plots for model 10.



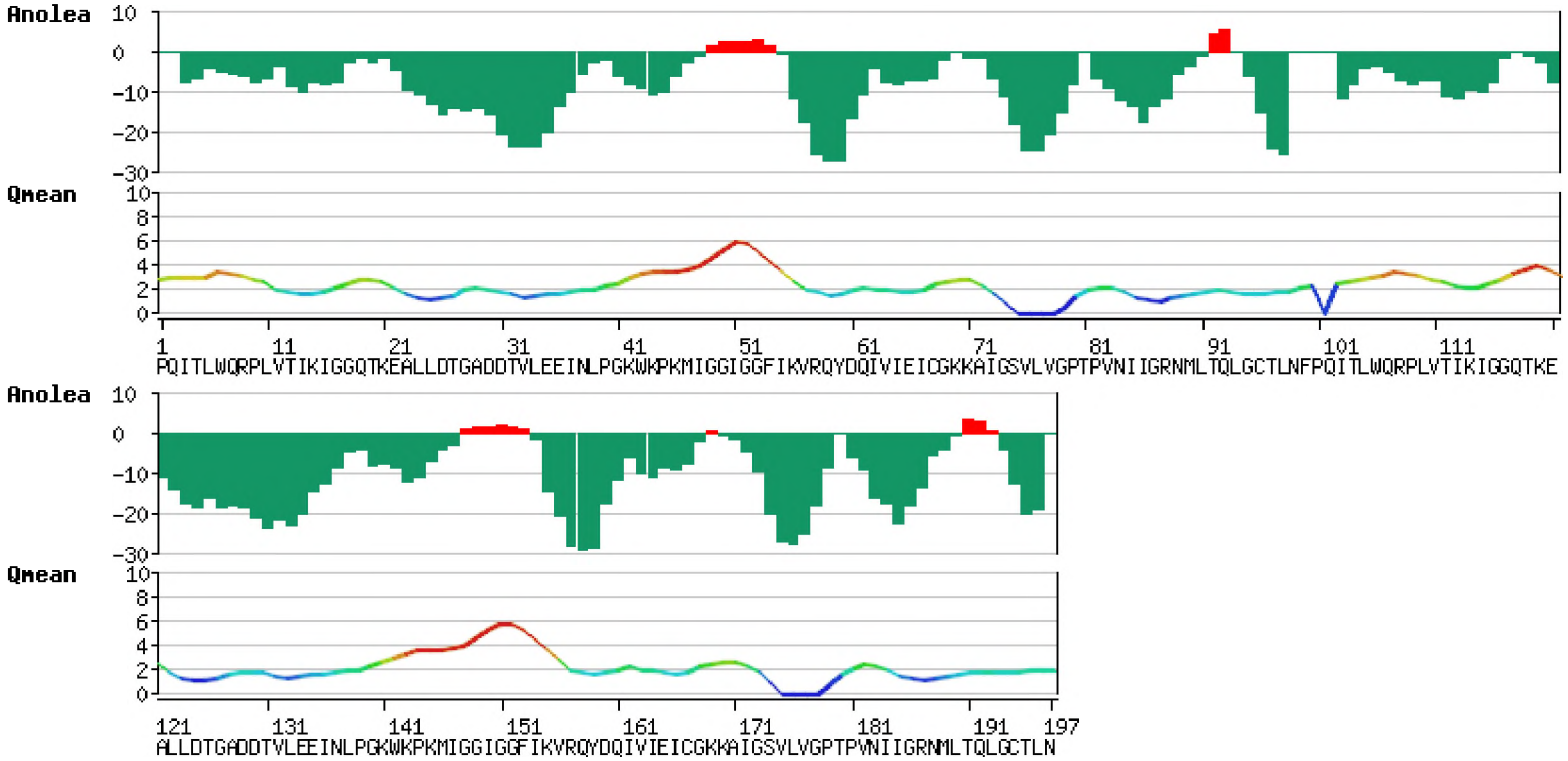
Supplementary Figure 15. ANOLEA and QMEAN score plots for model 11.



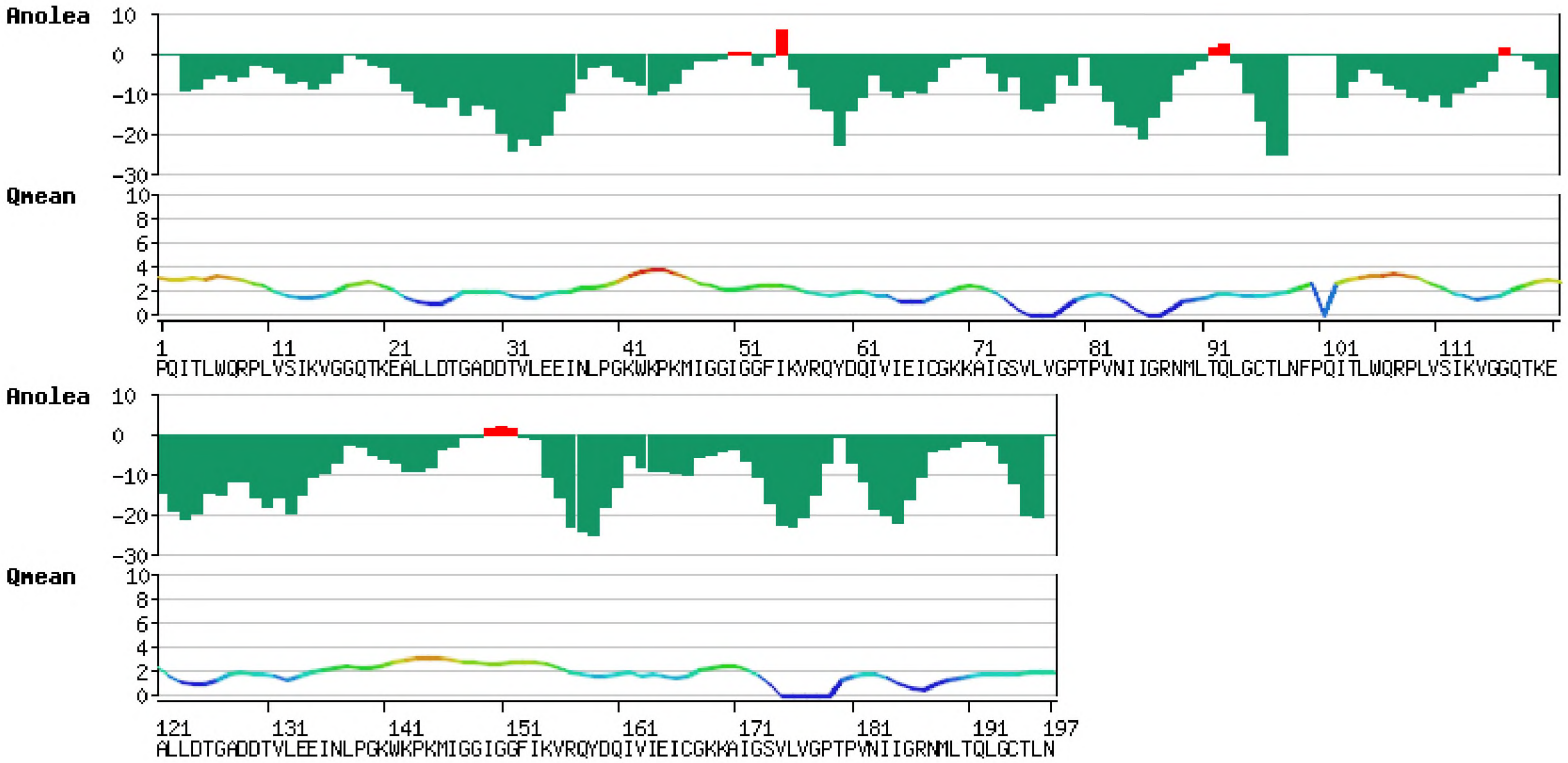
Supplementary Figure 16. ANOLEA and QMEAN score plots for model 12.



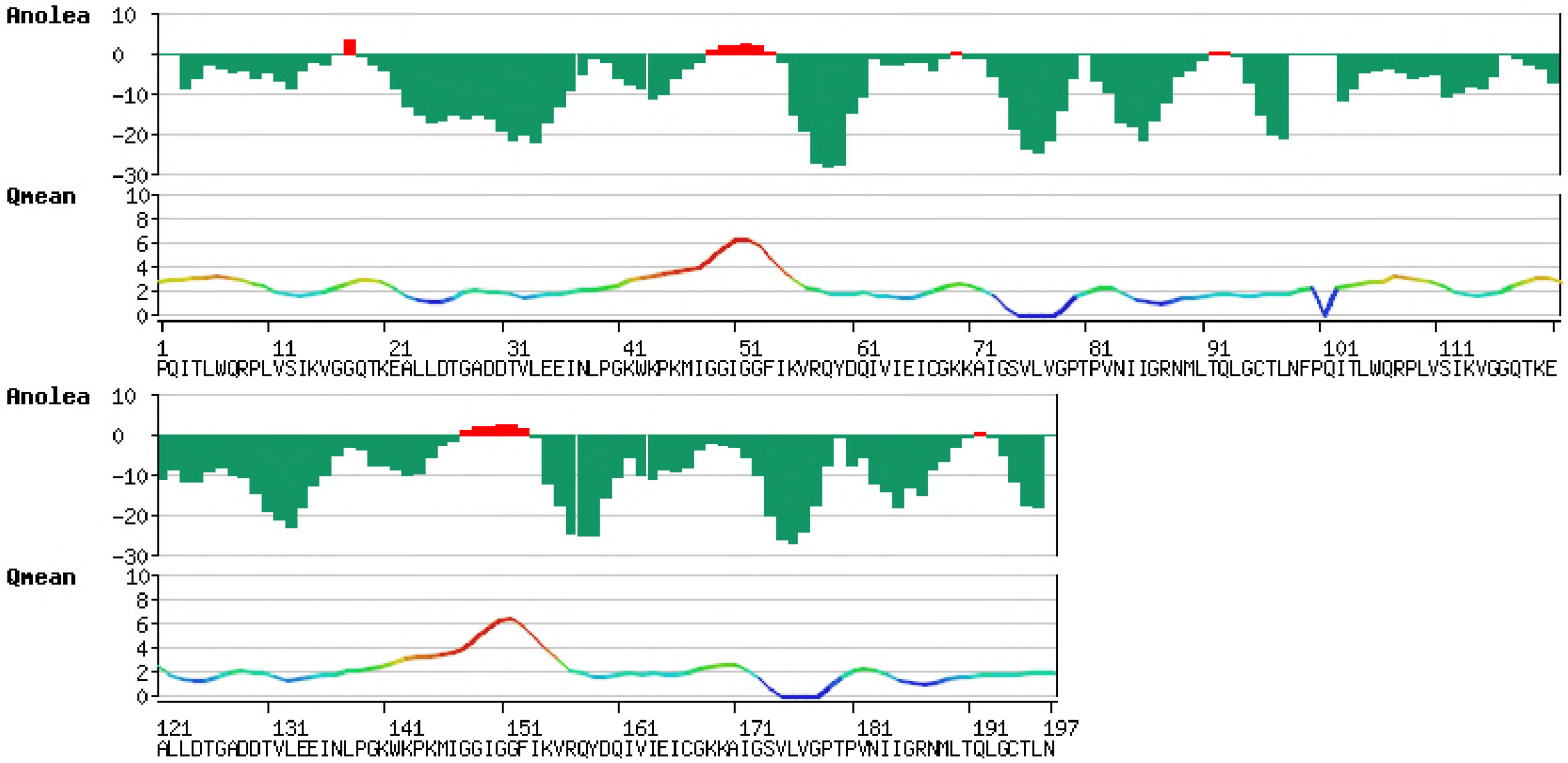
Supplementary Figure 17. ANOLEA and QMEAN score plots for model 13.



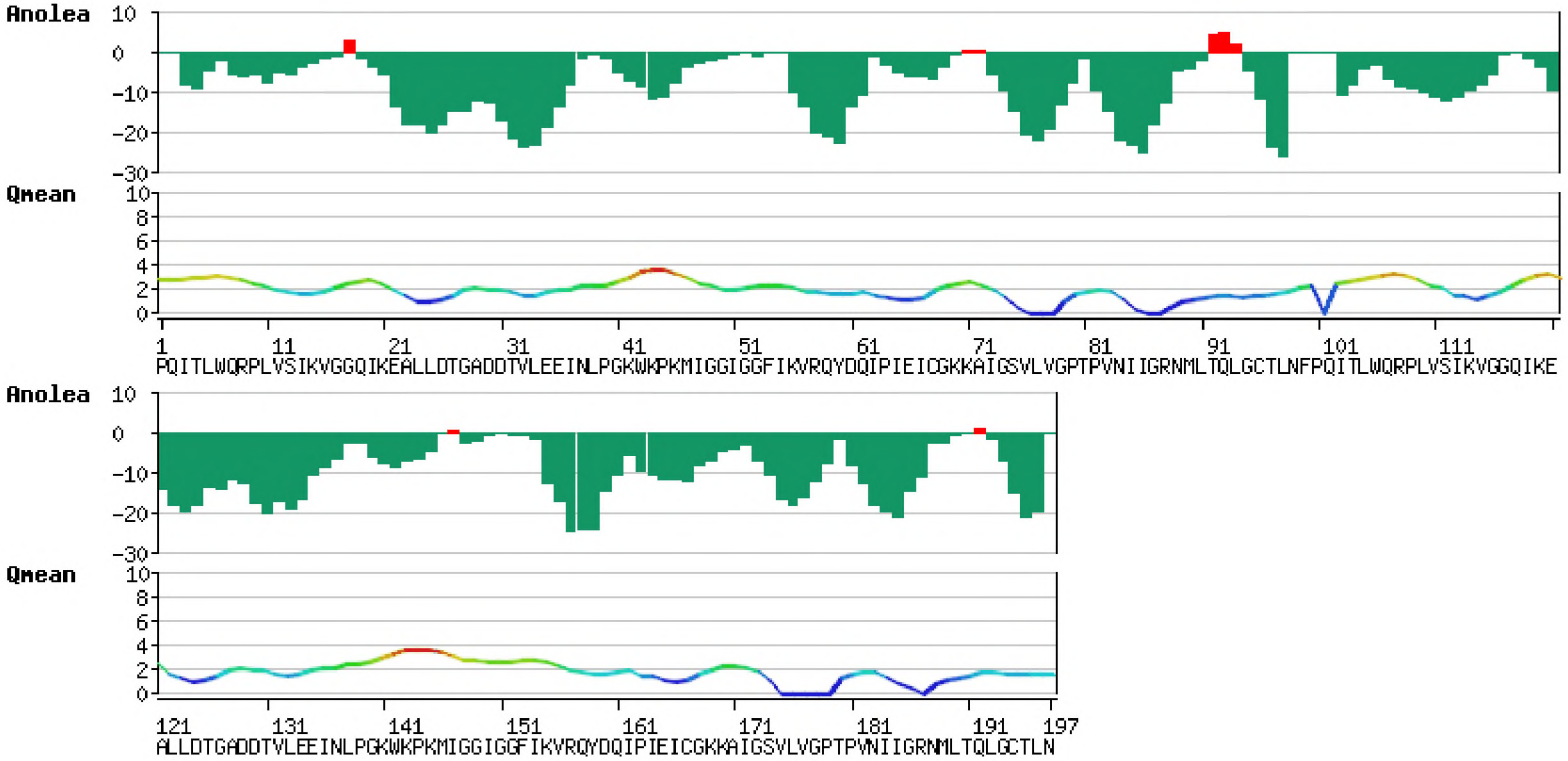
Supplementary Figure 18. ANOLEA and QMEAN score plots for model 14.



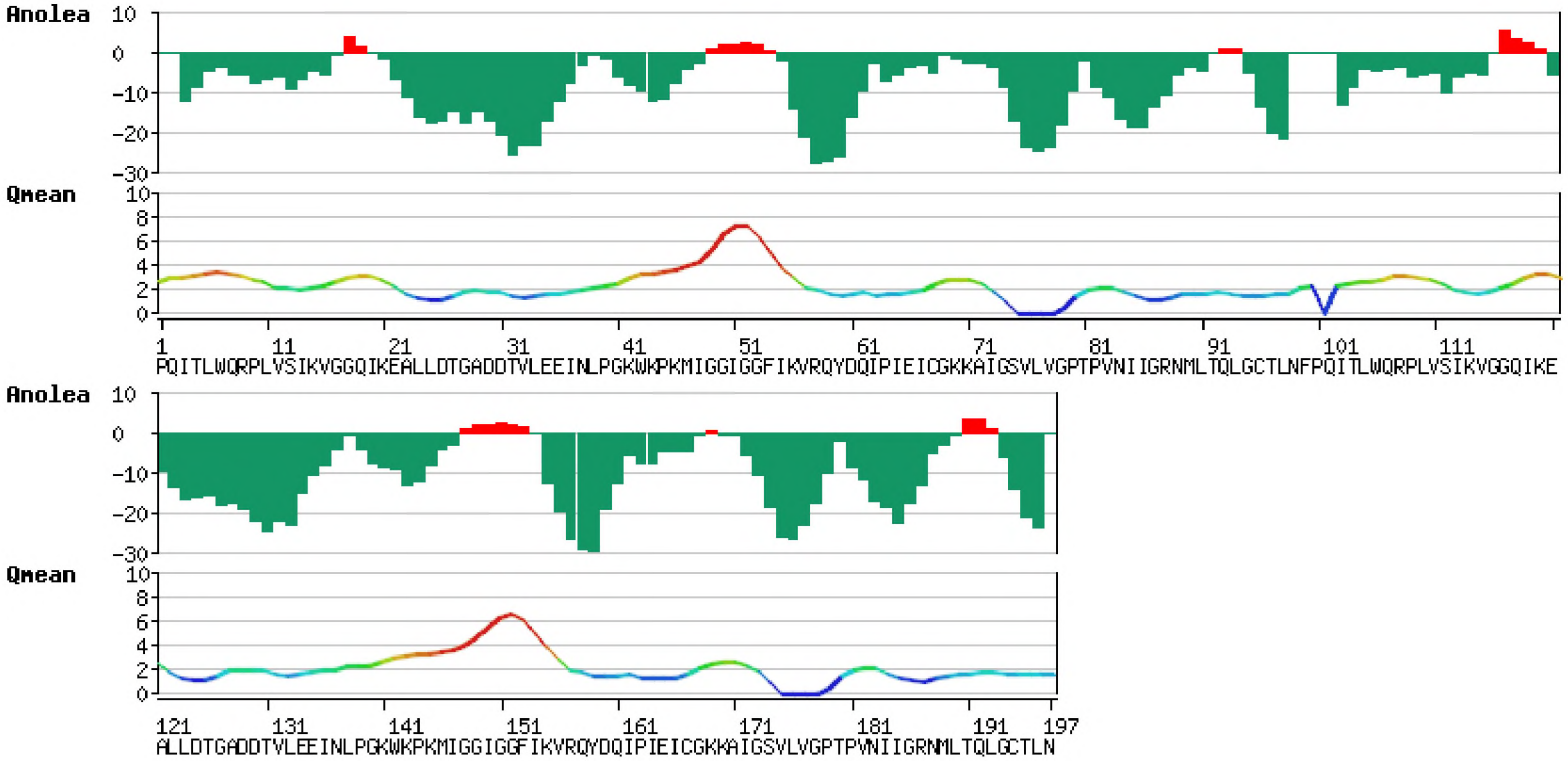
Supplementary Figure 19. ANOLEA and QMEAN score plots for model 15.



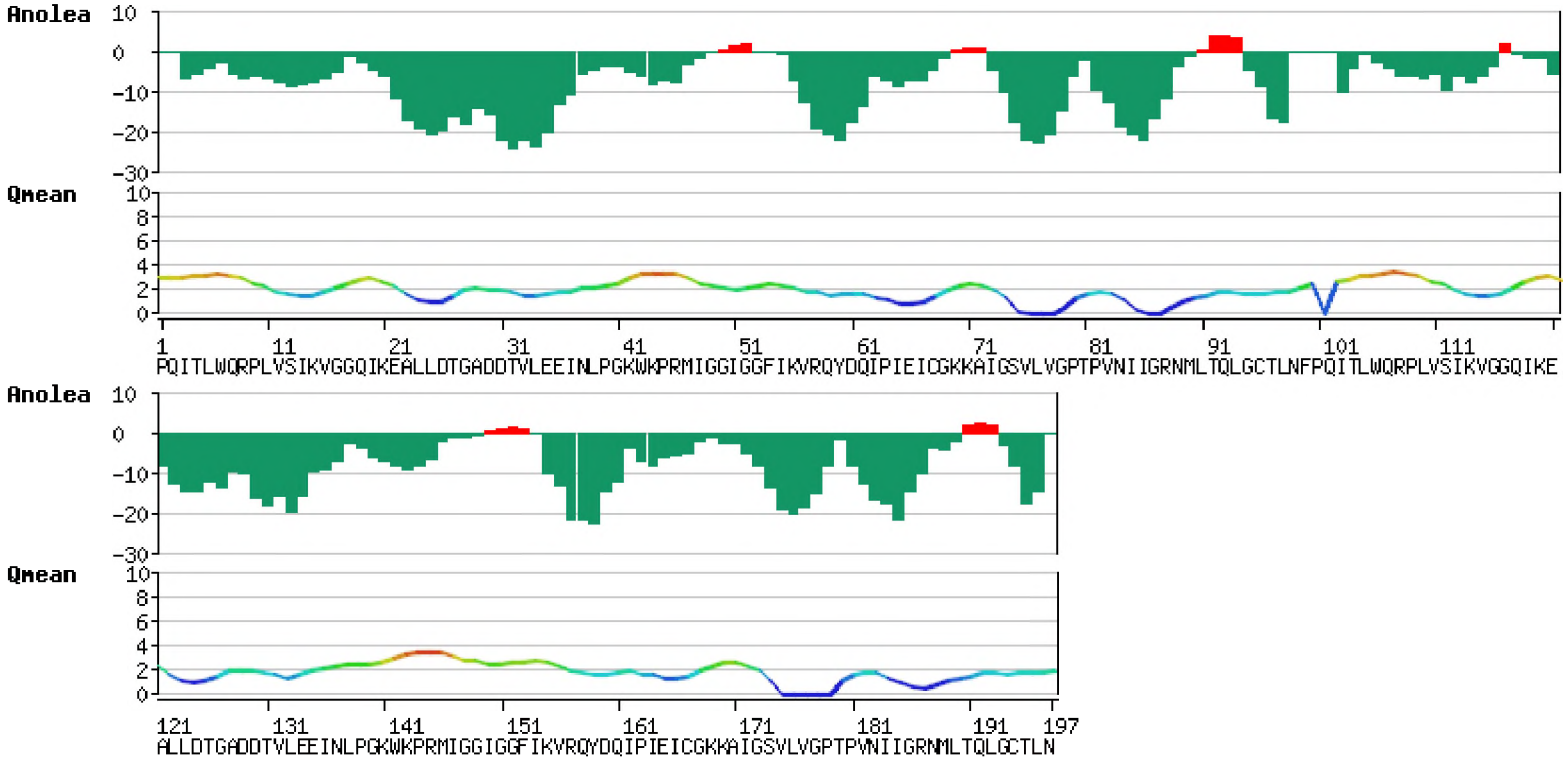
Supplementary Figure 20. ANOLEA and QMEAN score plots for model 16.



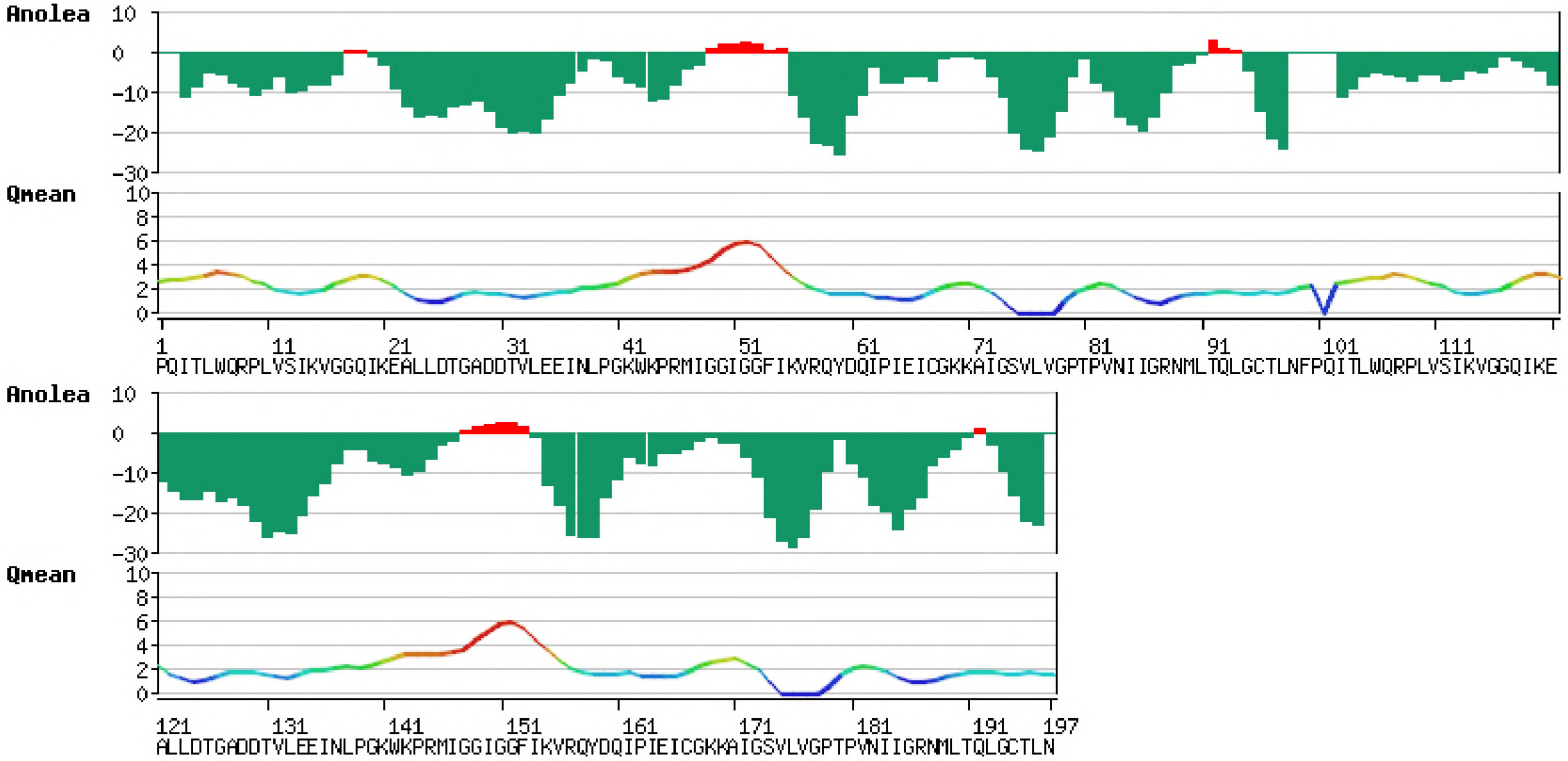
Supplementary Figure 21. ANOLEA and QMEAN score plots for model 17.



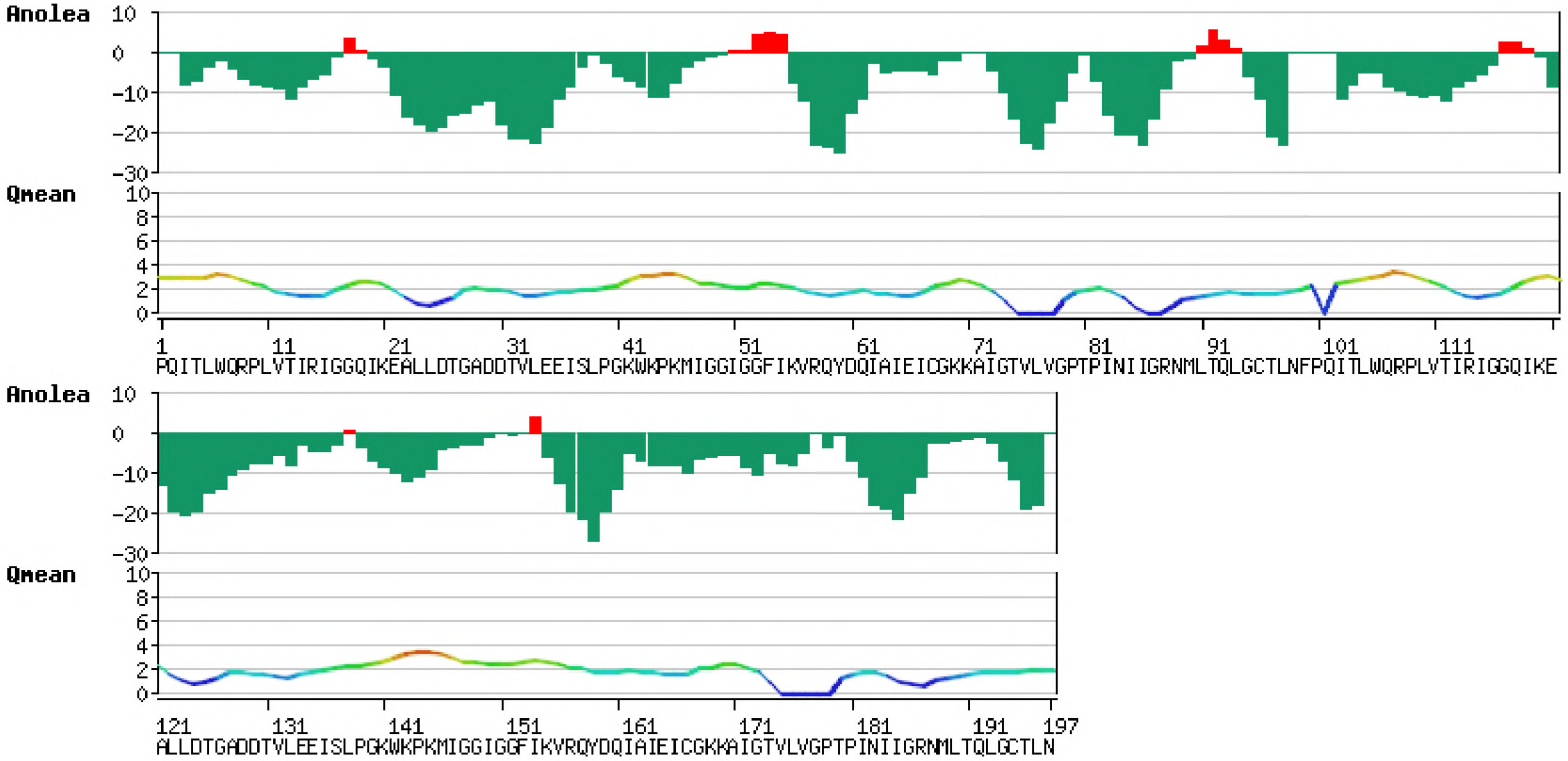
Supplementary Figure 22. ANOLEA and QMEAN score plots for model 18.



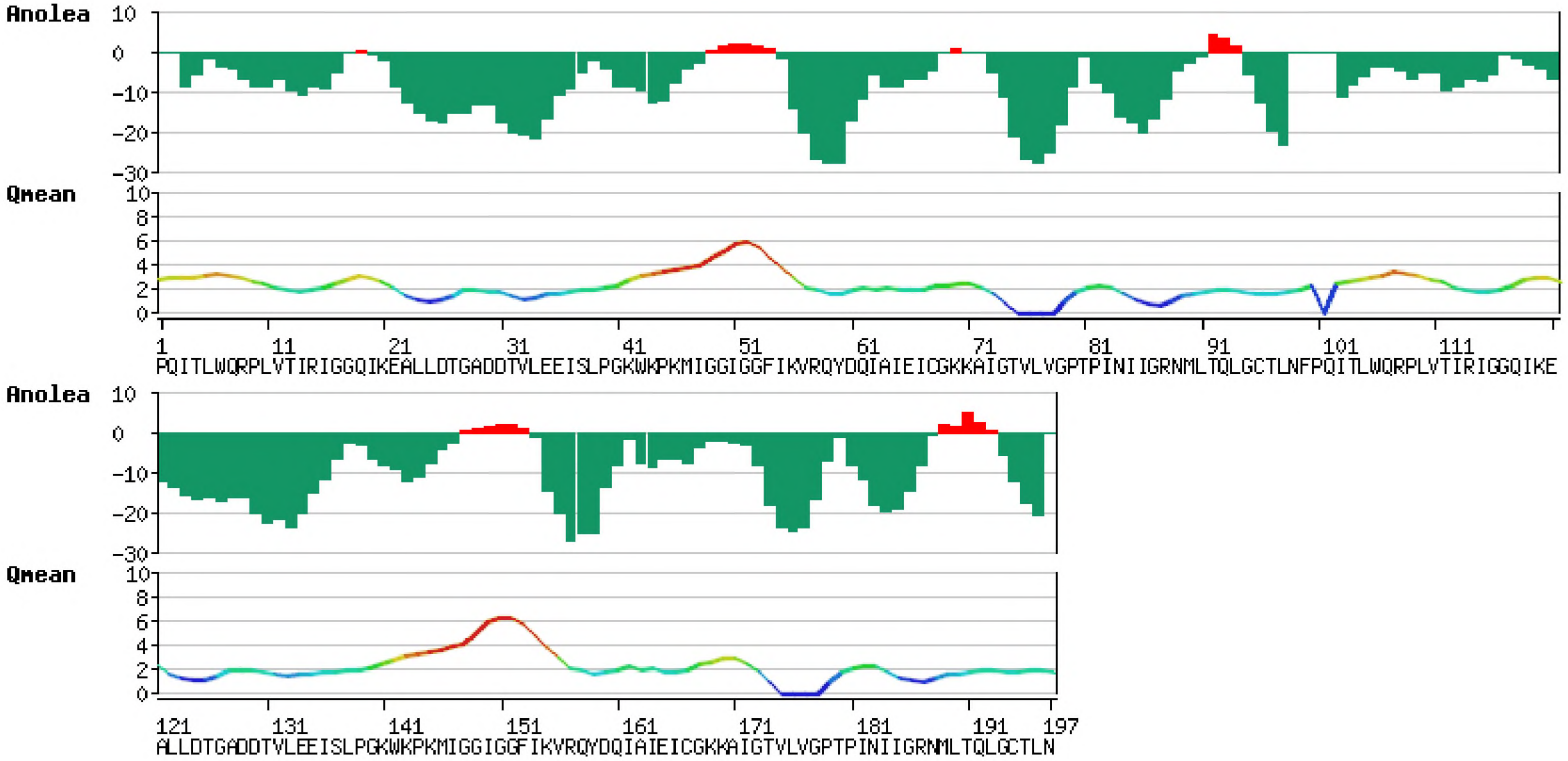
Supplementary Figure 23. ANOLEA and QMEAN score plots for model 19.



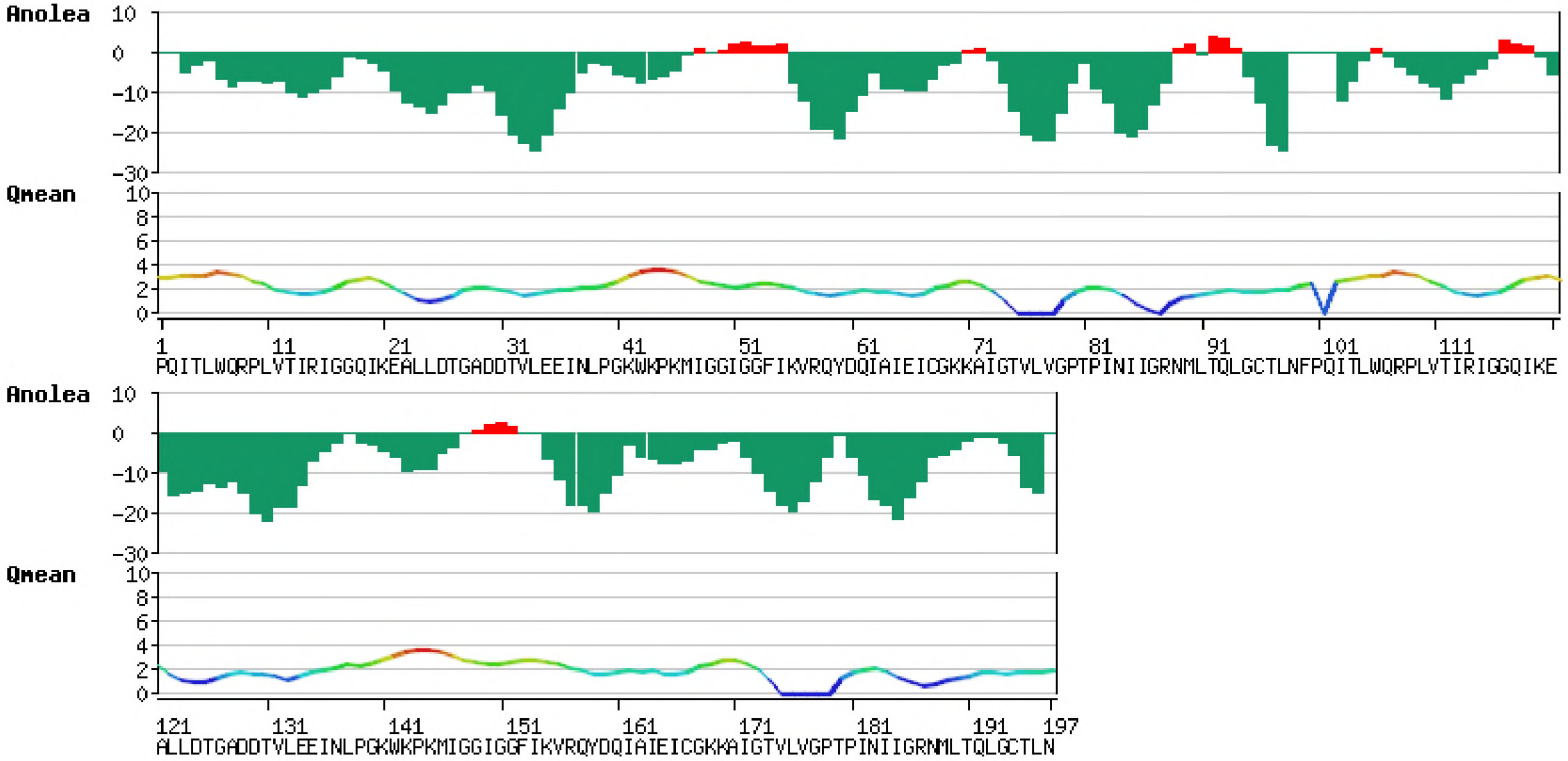
Supplementary Figure 24. ANOLEA and QMEAN score plots for model 20.



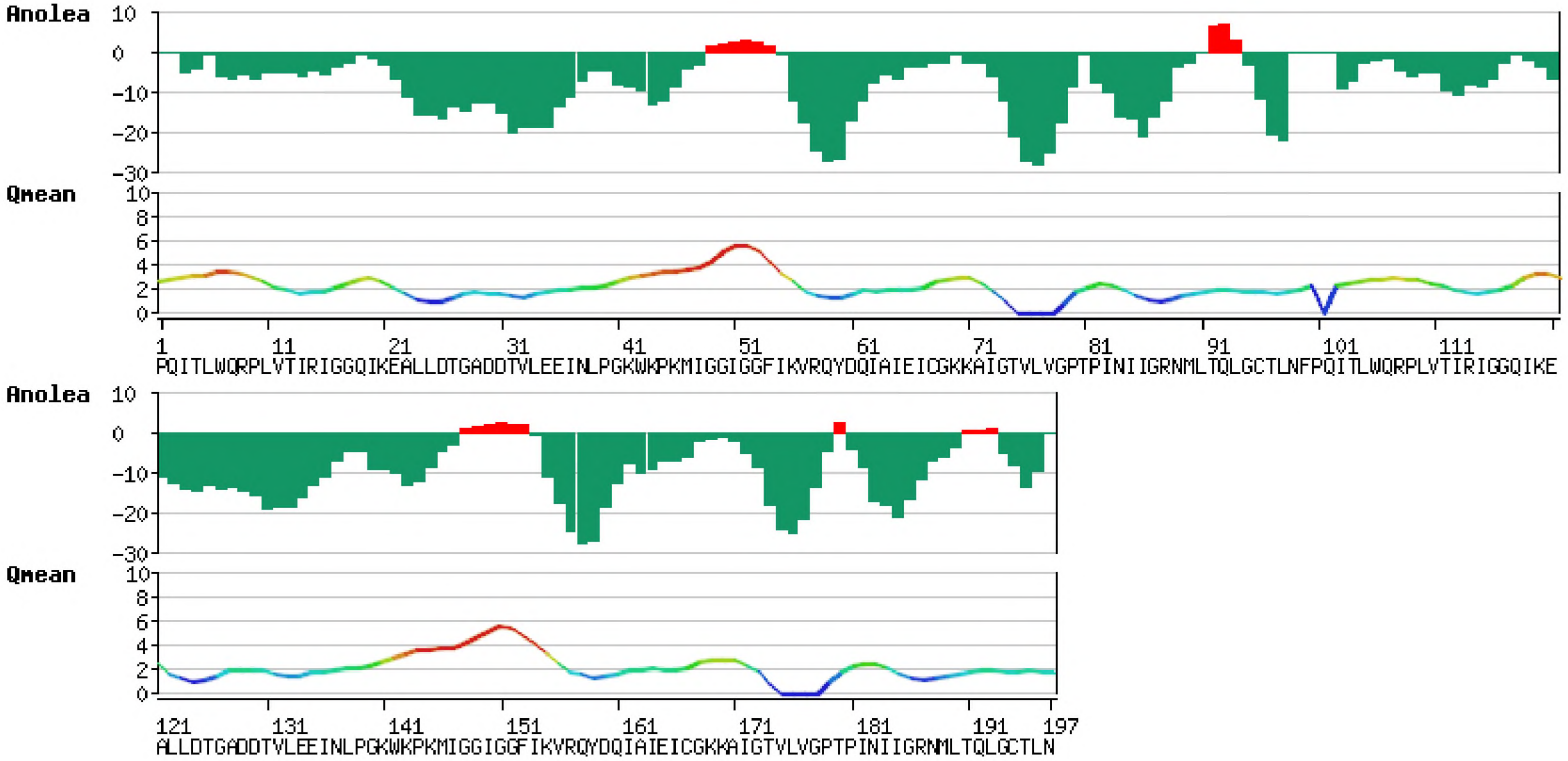
Supplementary Figure 25. ANOLEA and QMEAN score plots for model 21.



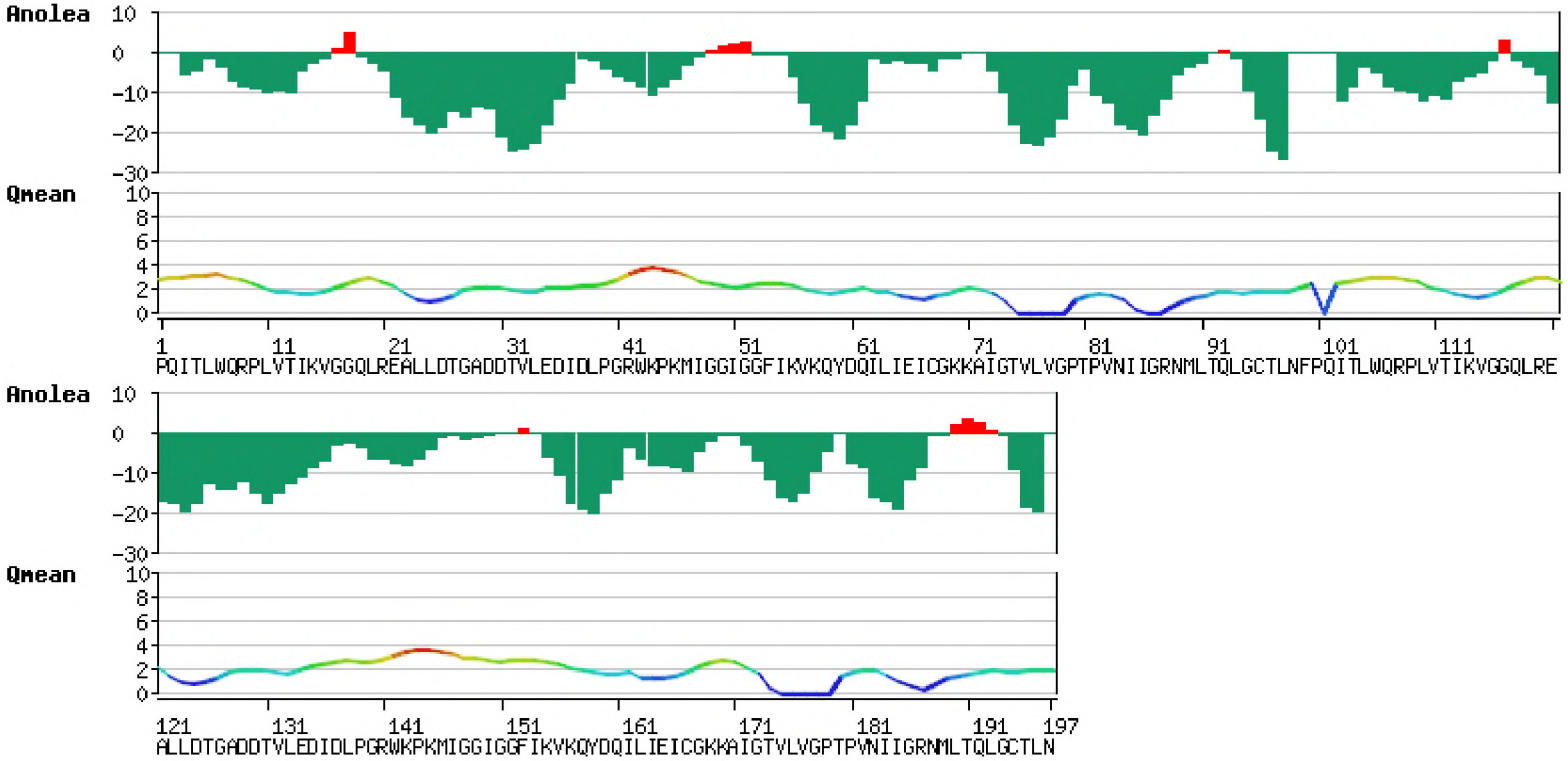
Supplementary Figure 26. ANOLEA and QMEAN score plots for model 22.



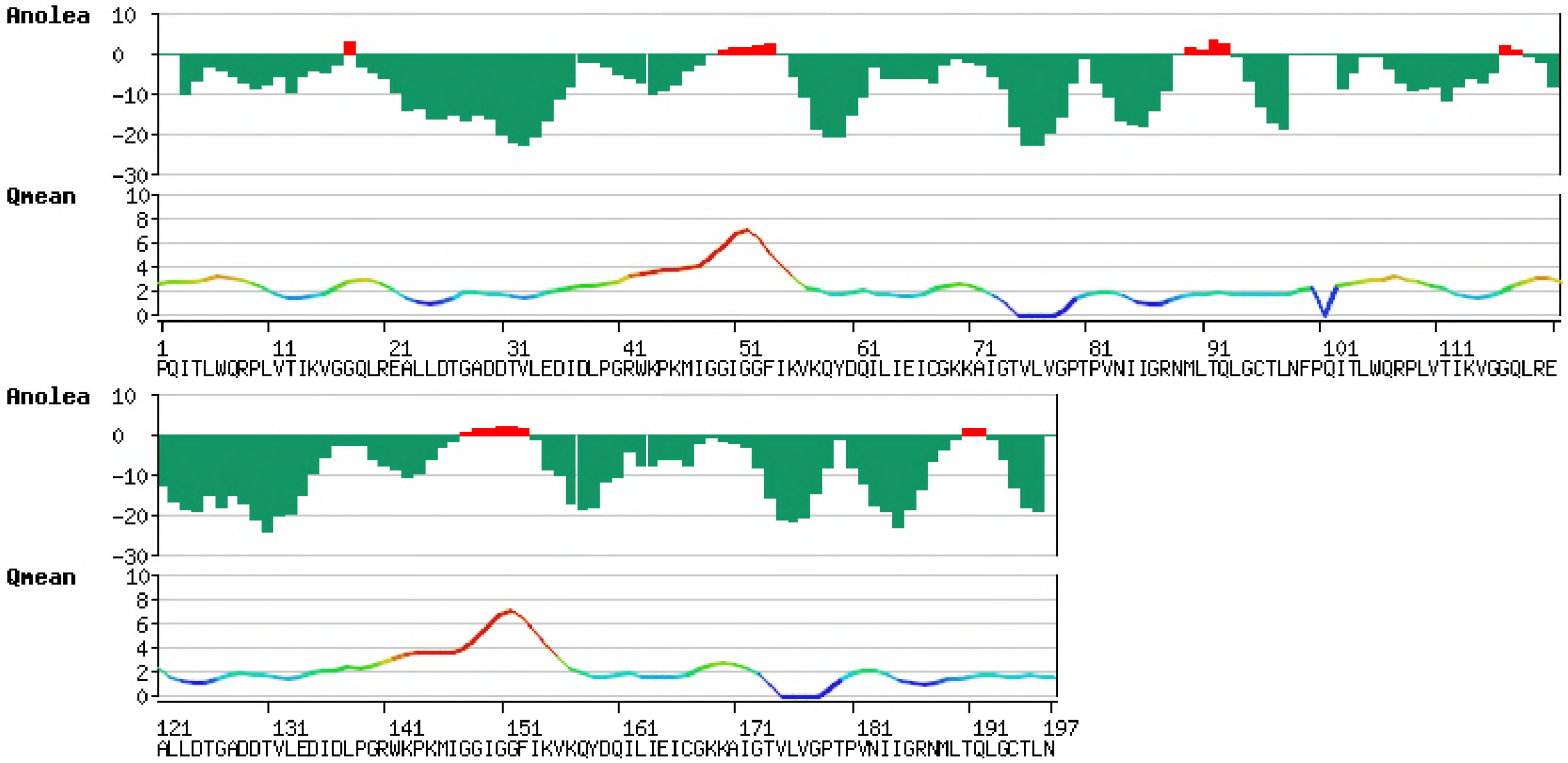
Supplementary Figure 27. ANOLEA and QMEAN score plots for model 23.



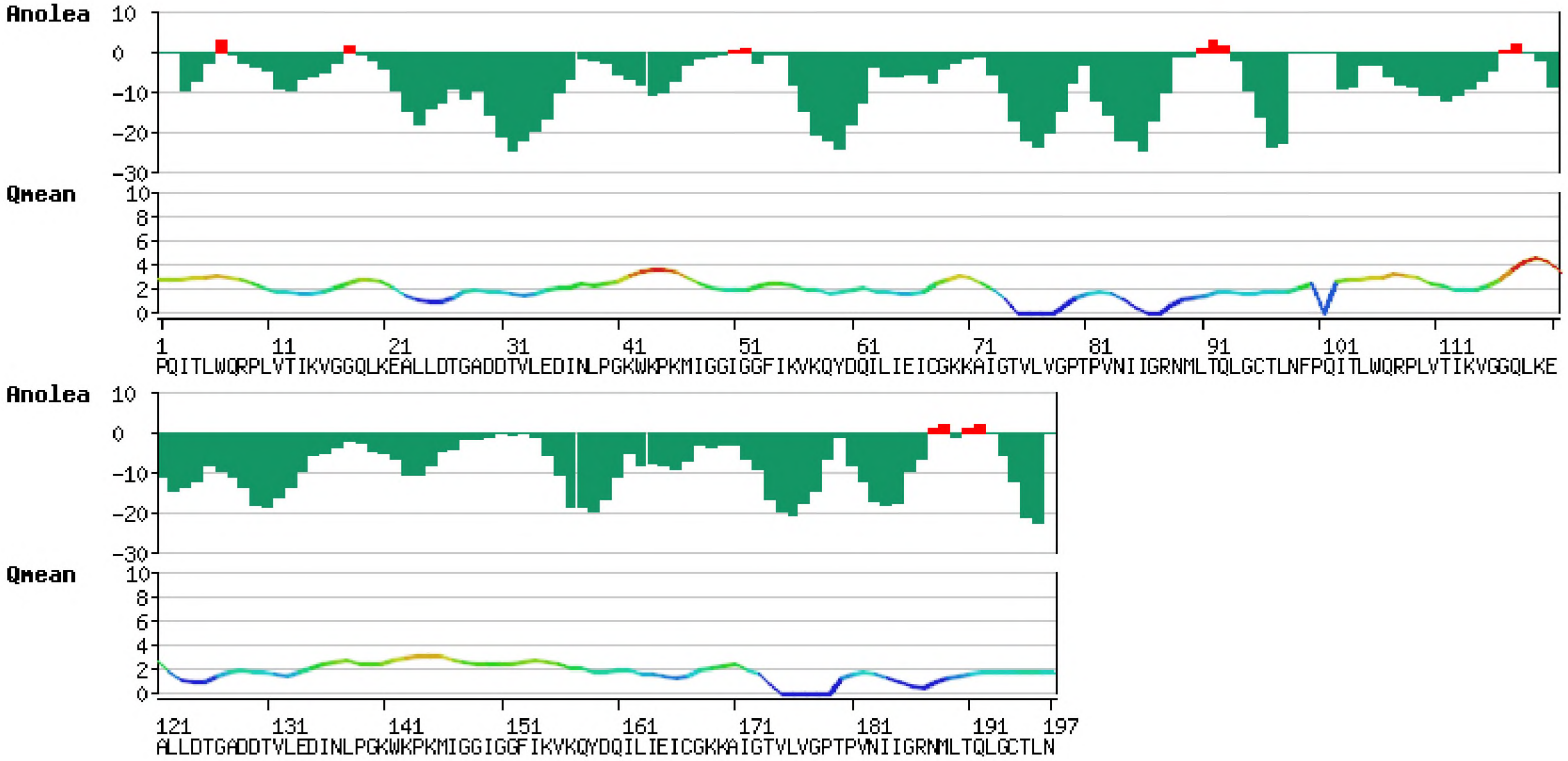
Supplementary Figure 28. ANOLEA and QMEAN score plots for model 24.



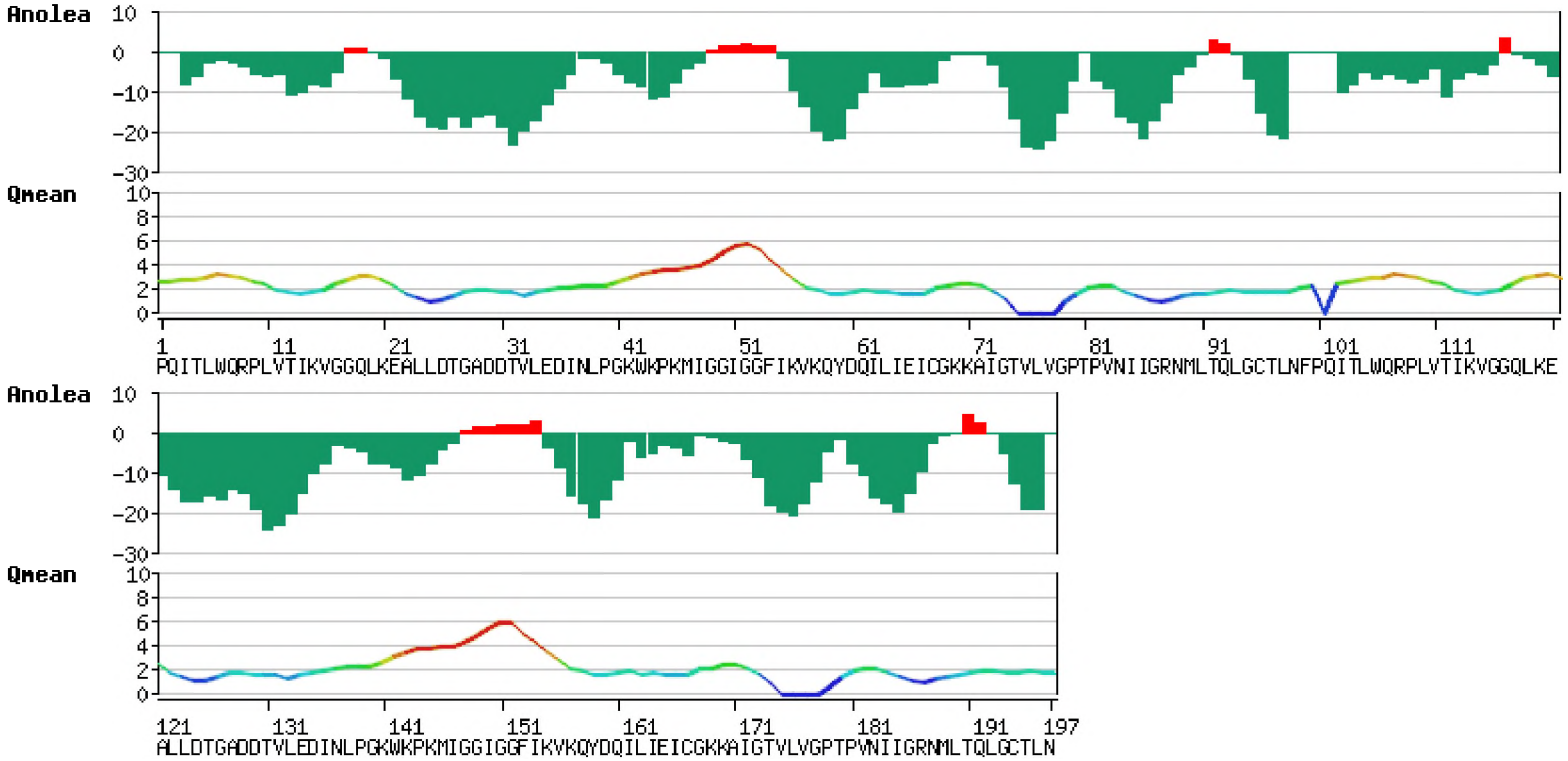
Supplementary Figure 29. ANOLEA and QMEAN score plots for model 25.



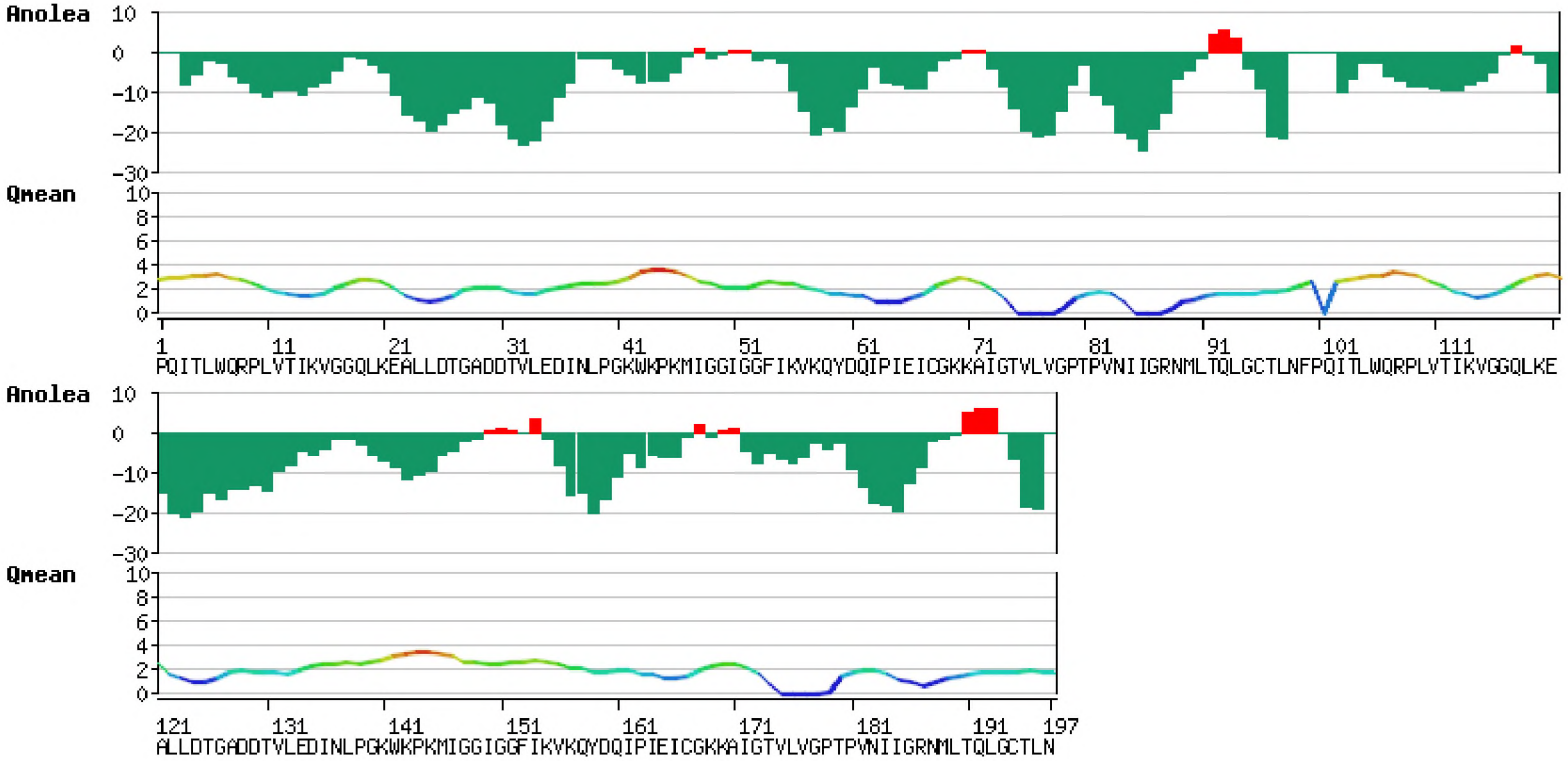
Supplementary Figure 30. ANOLEA and QMEAN score plots for model 26.



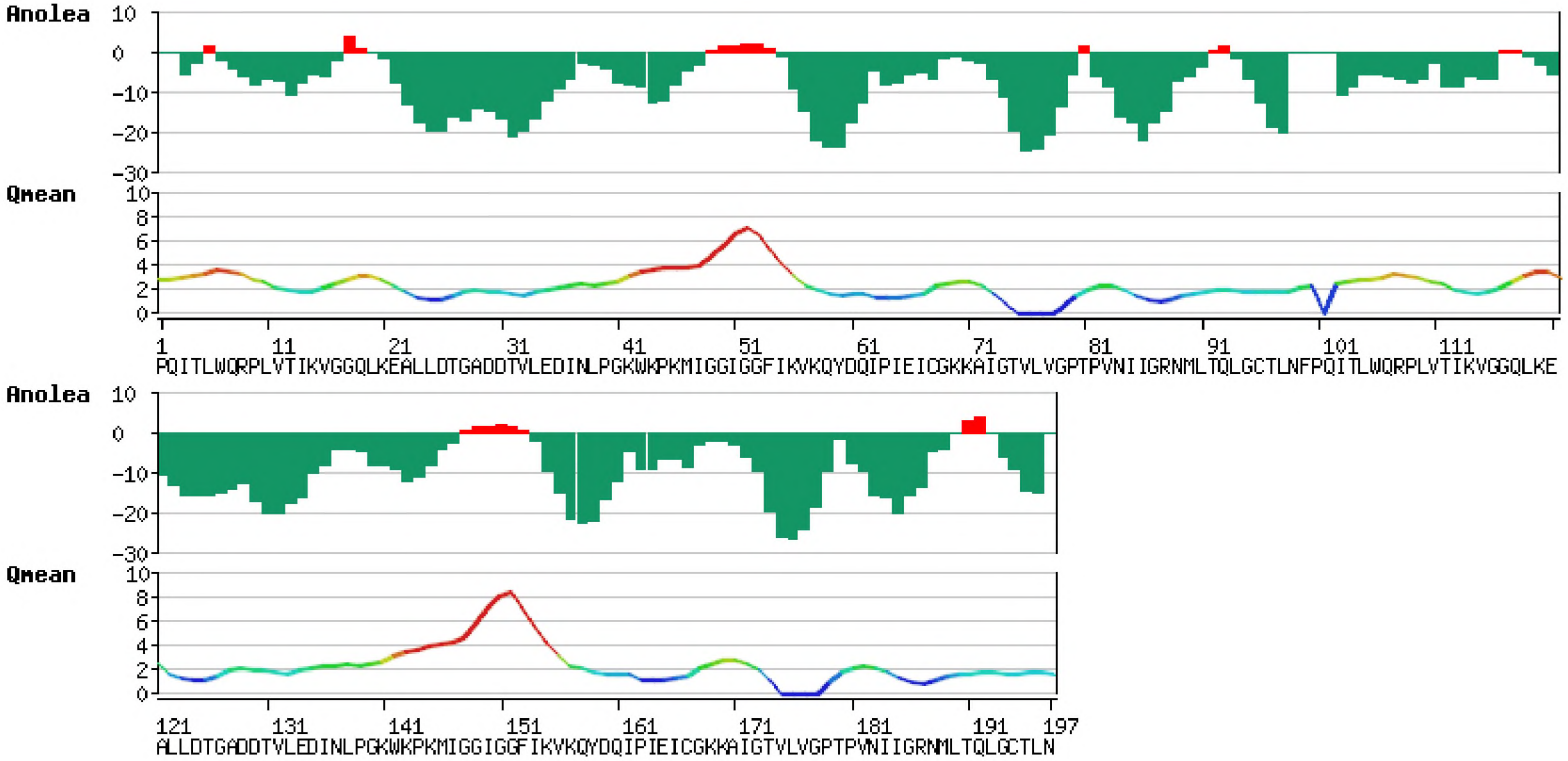
Supplementary Figure 31. ANOLEA and QMEAN score plots for model 27.



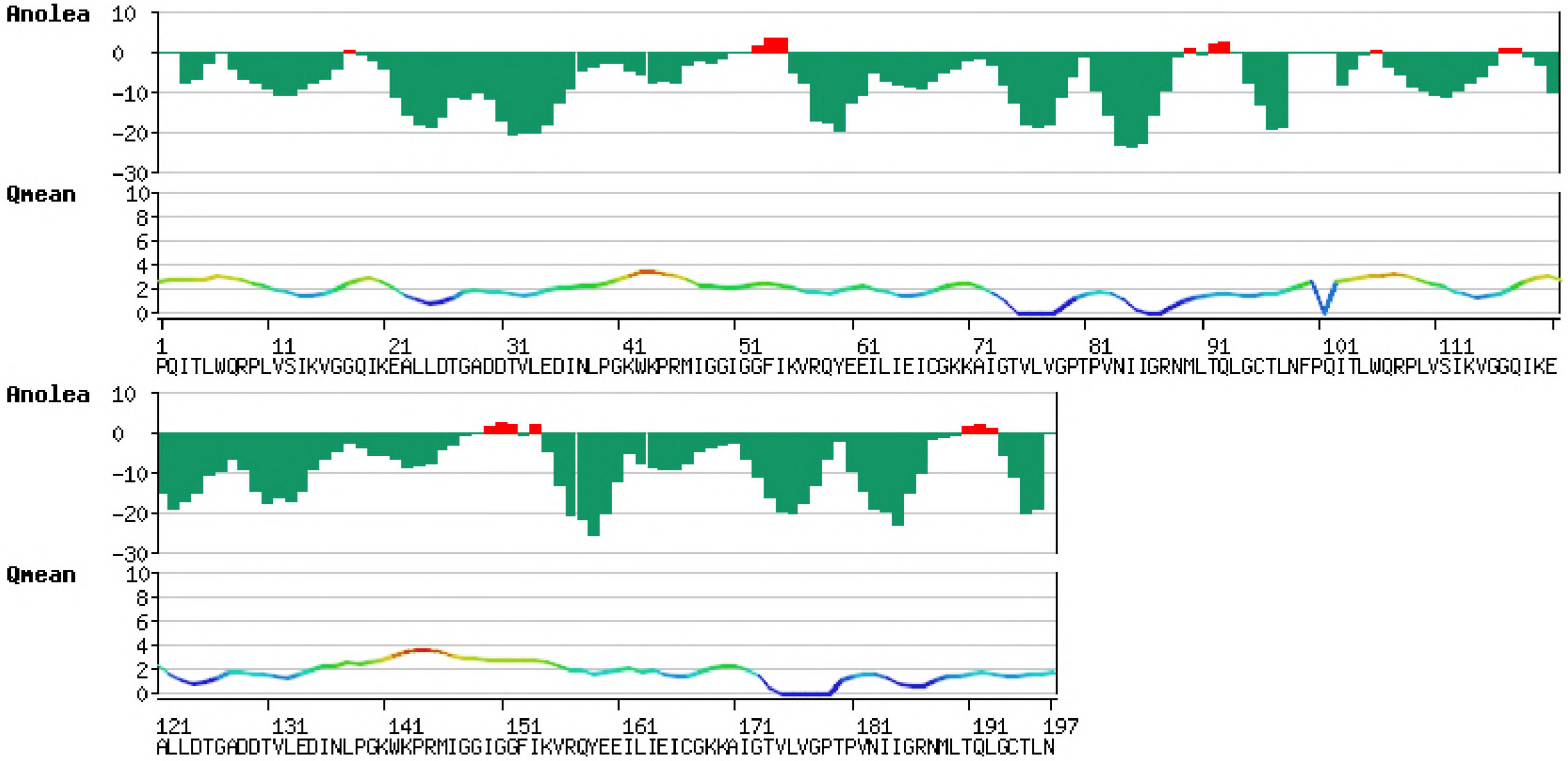
Supplementary Figure 32. ANOLEA and QMEAN score plots for model 28.



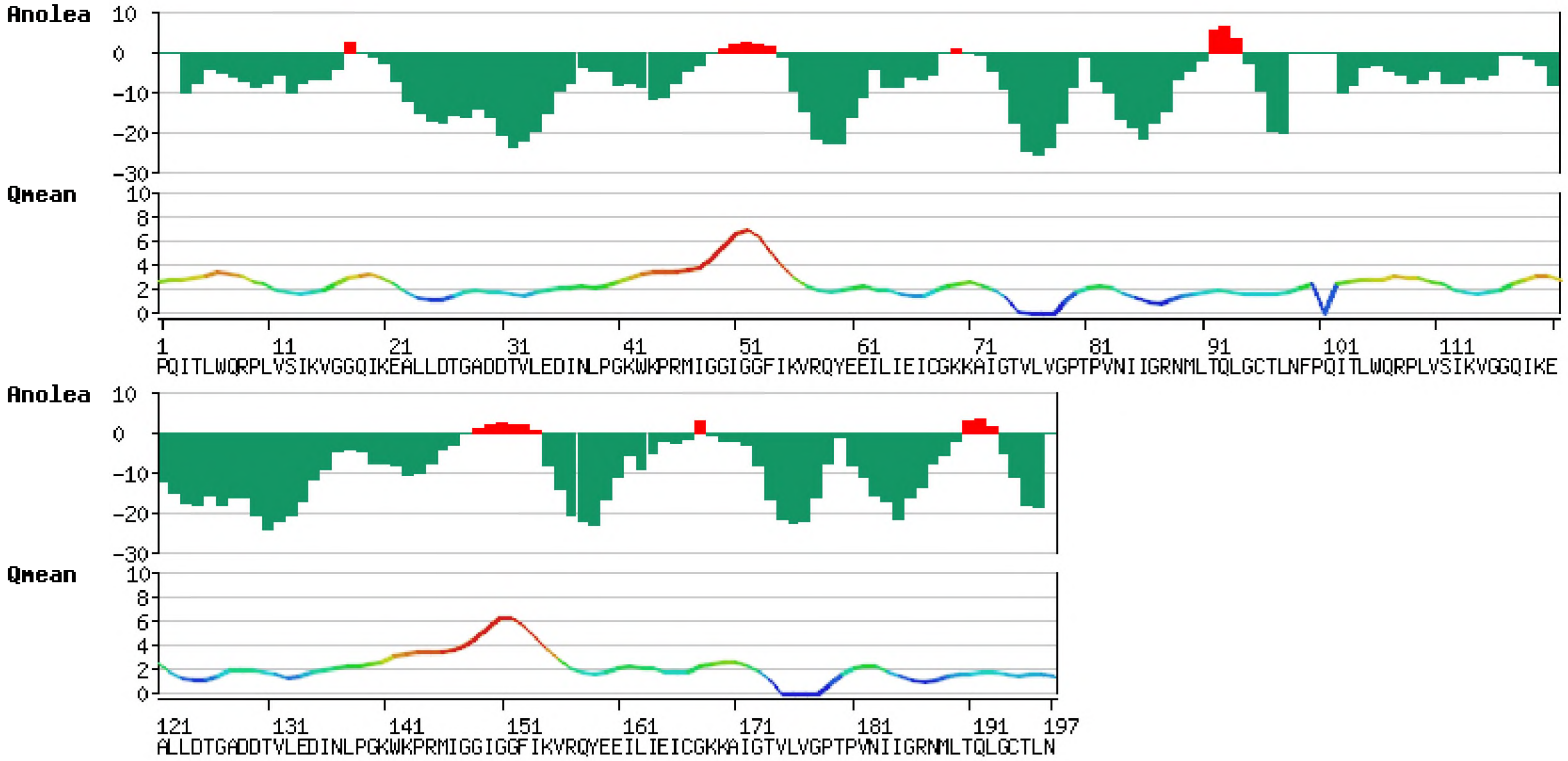
Supplementary Figure 33. ANOLEA and QMEAN score plots for model 29.



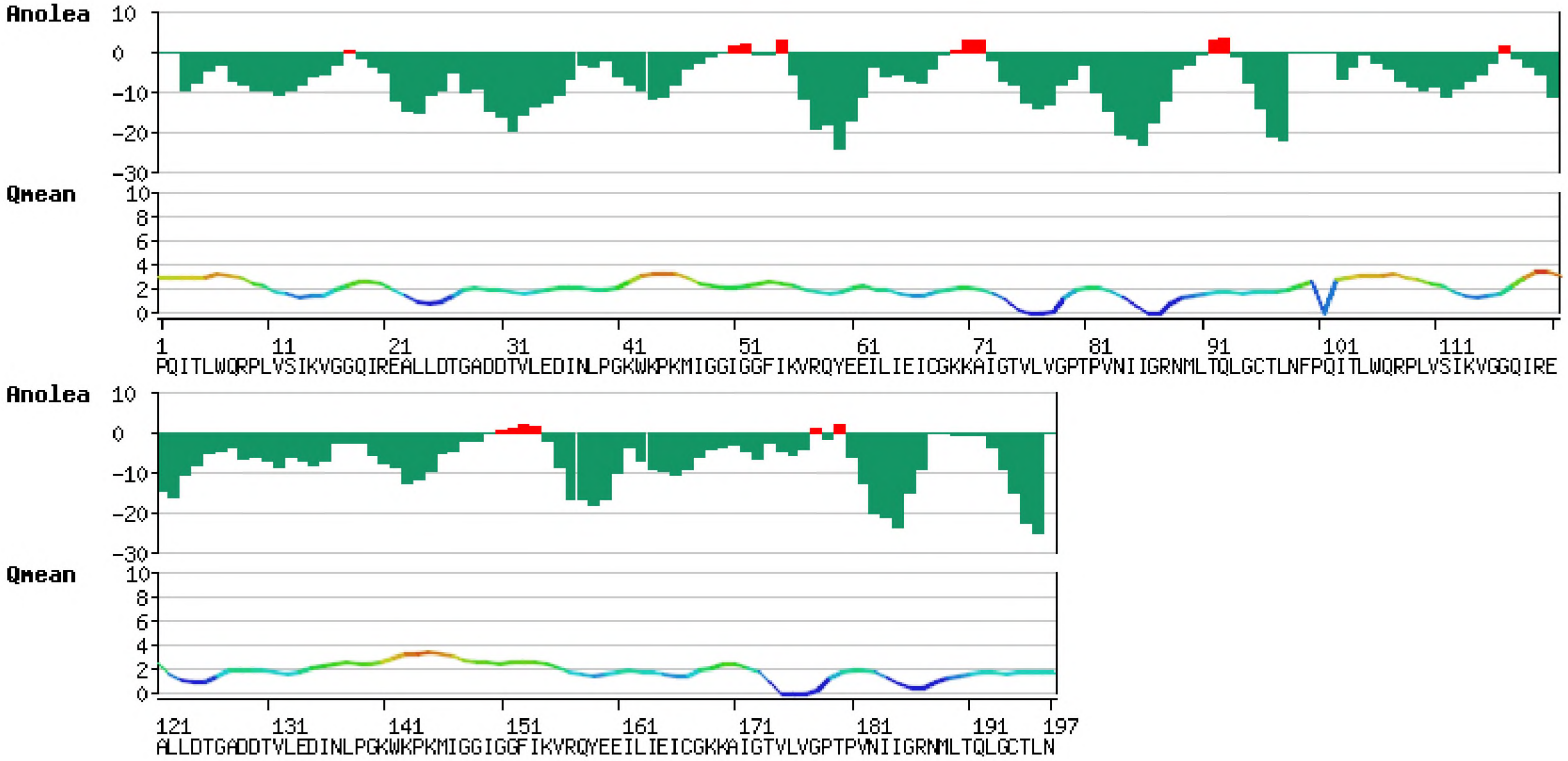
Supplementary Figure 34. ANOLEA and QMEAN score plots for model 30.



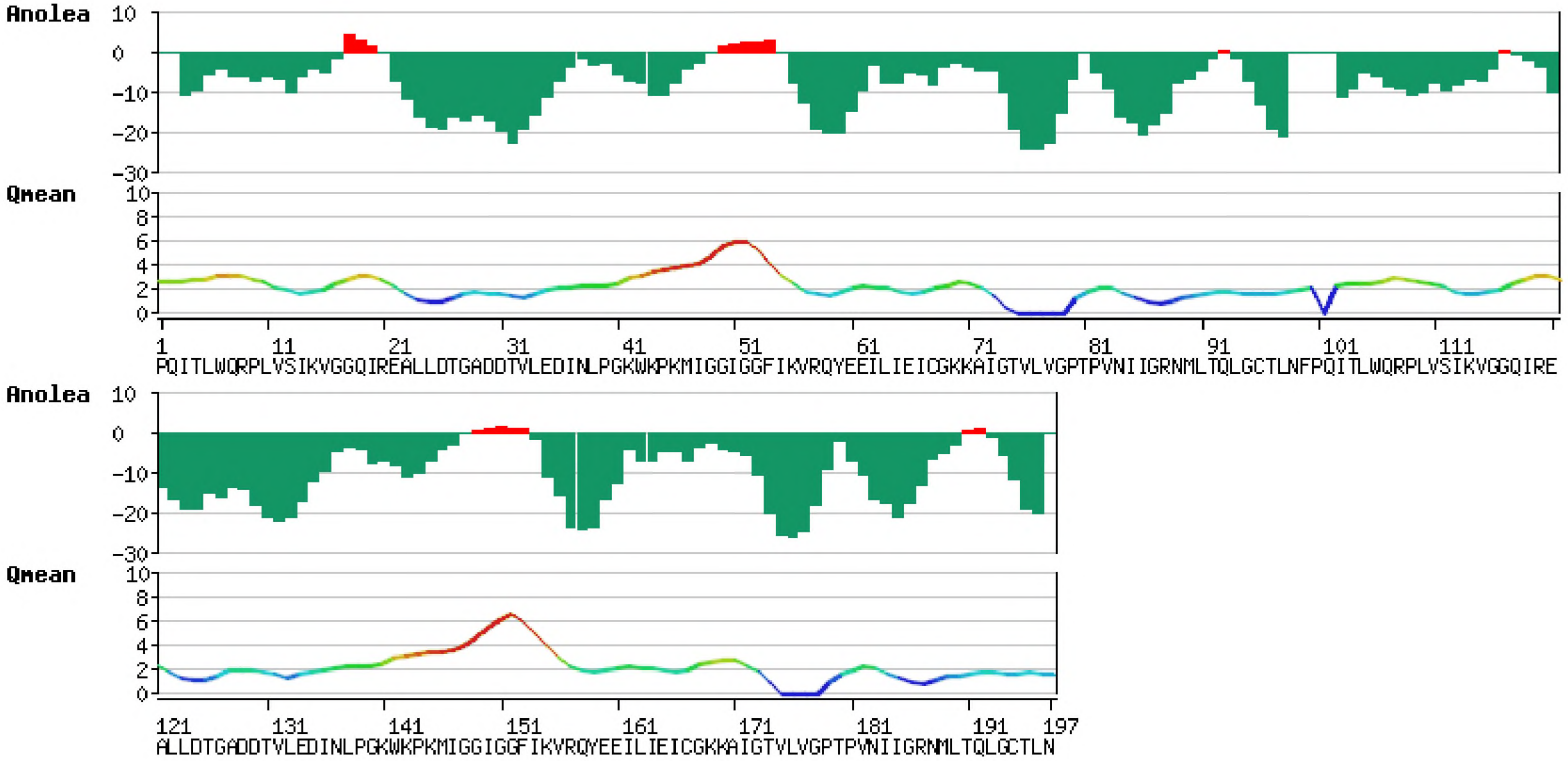
Supplementary Figure 35. ANOLEA and QMEAN score plots for model 31.



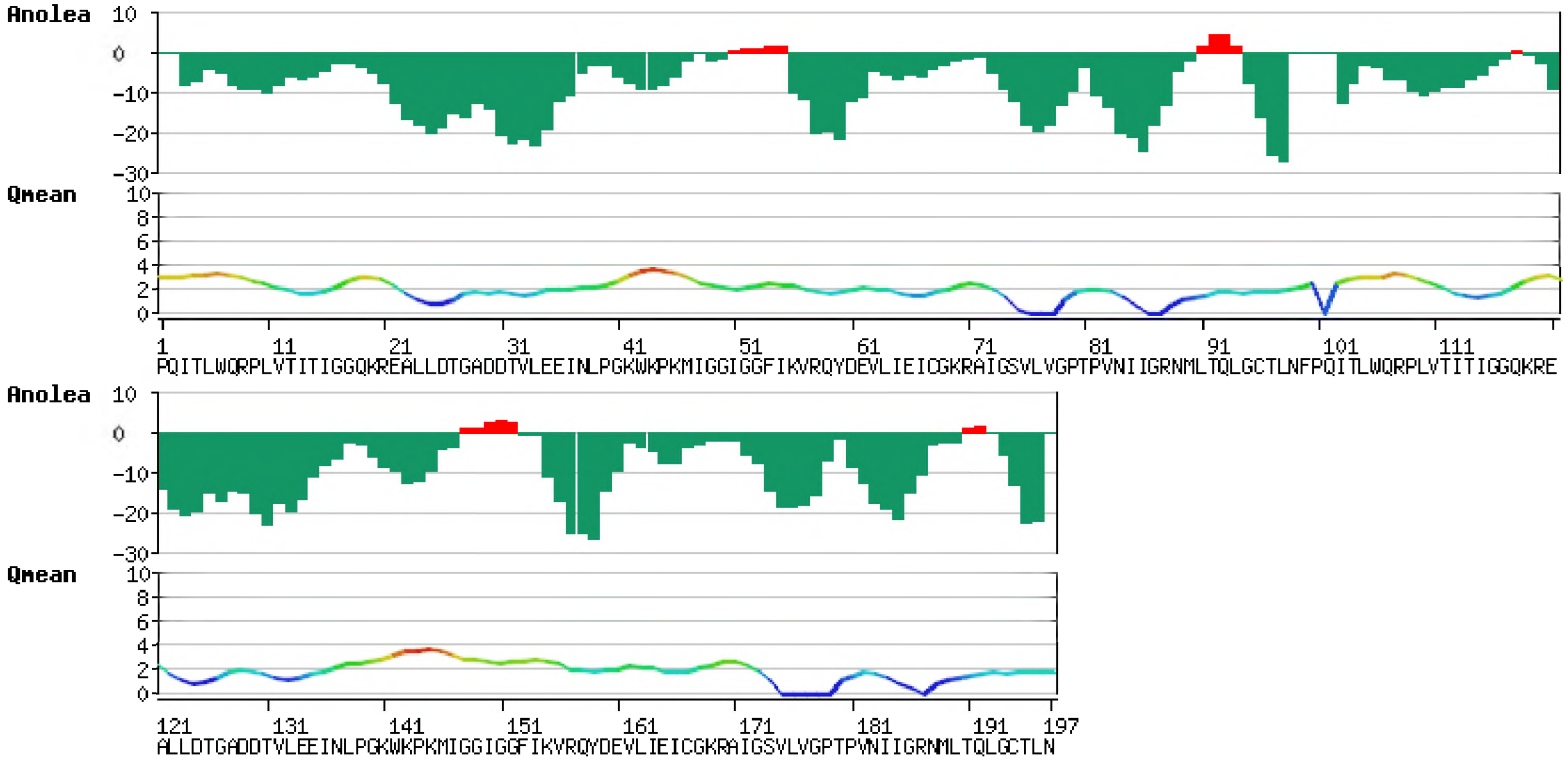
Supplementary Figure 36. ANOLEA and QMEAN score plots for model 32.



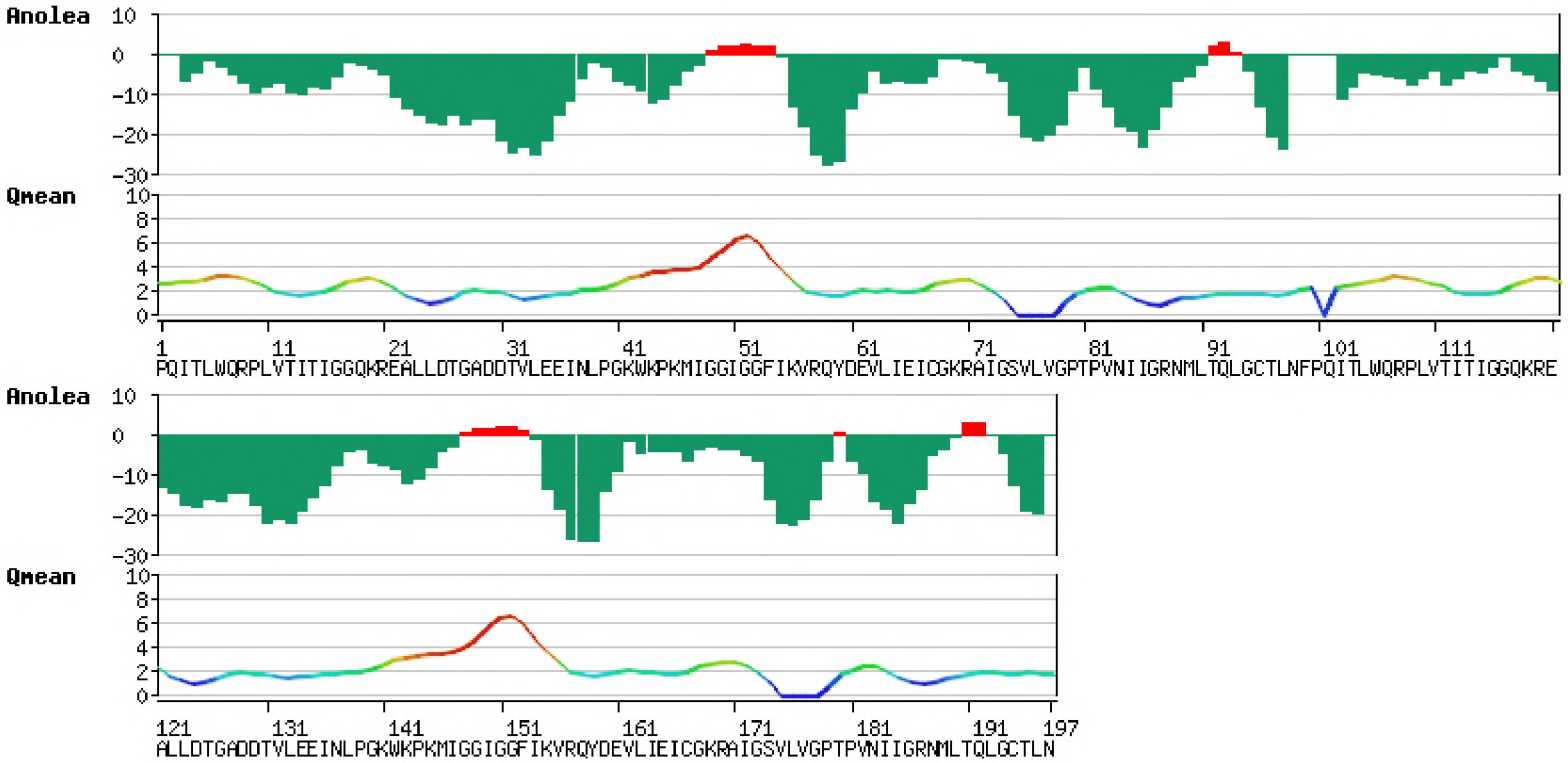
Supplementary Figure 37. ANOLEA and QMEAN score plots for model 33.



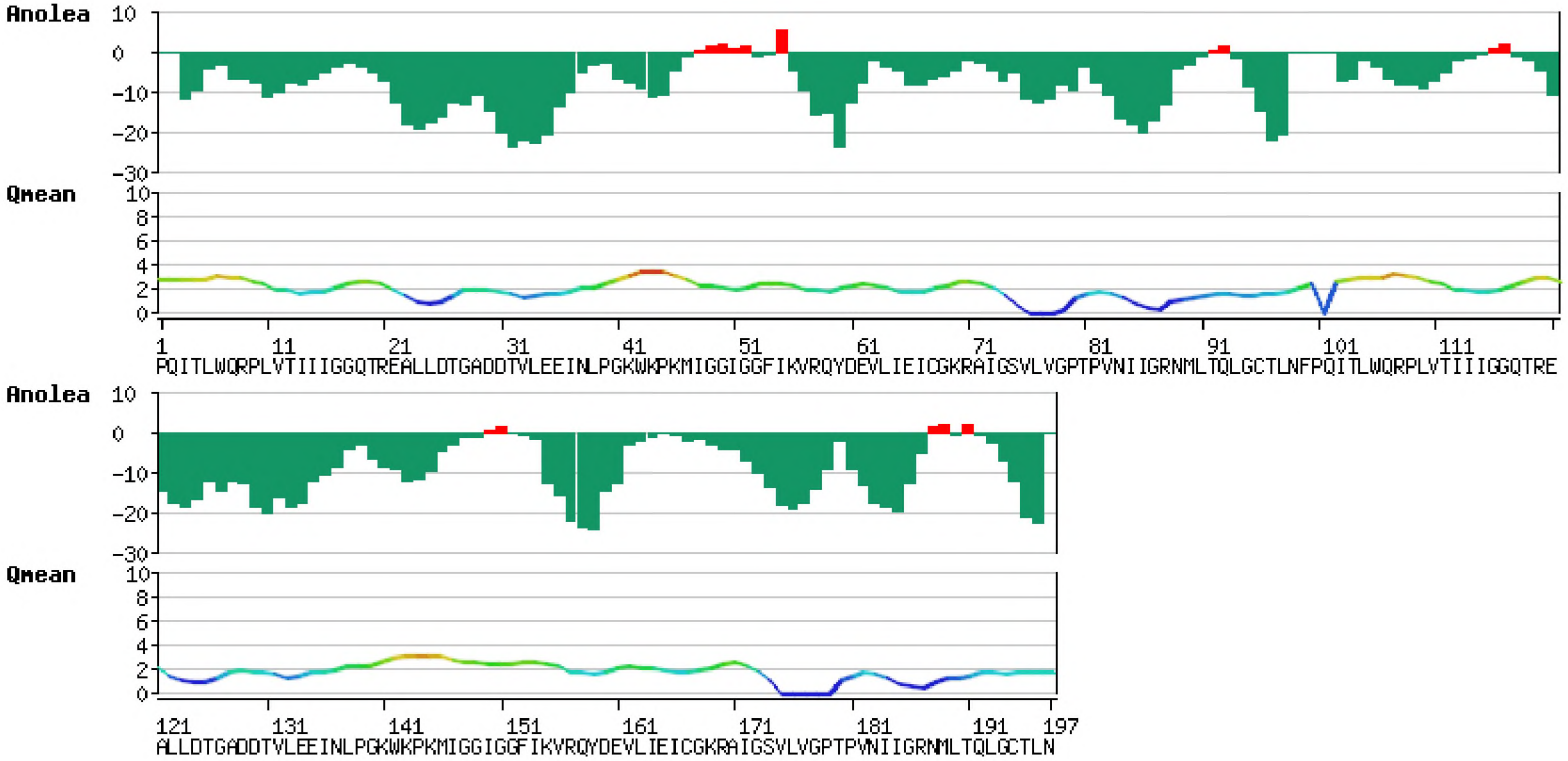
Supplementary Figure 38. ANOLEA and QMEAN score plots for model 34.



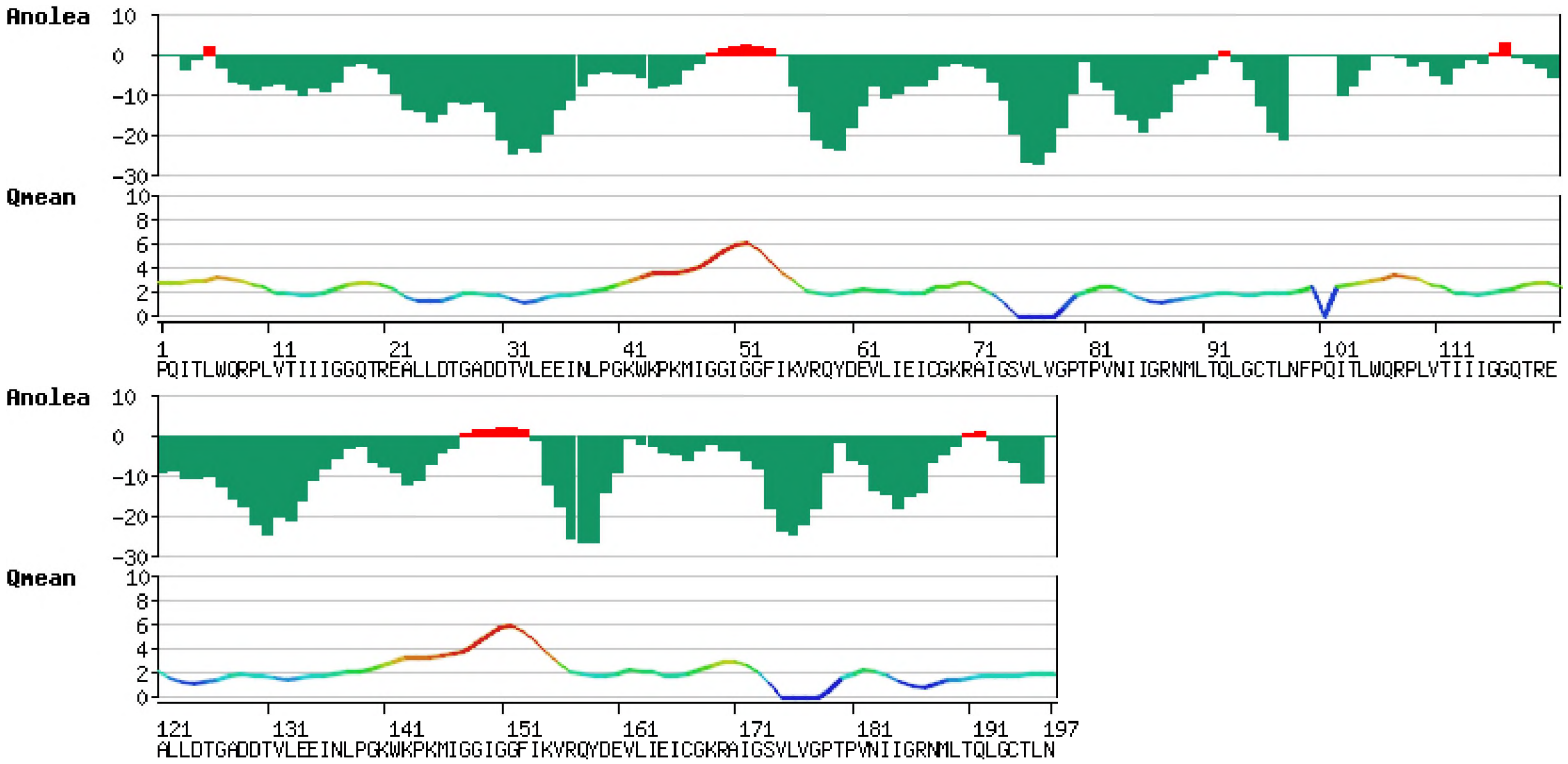
Supplementary Figure 39. ANOLEA and QMEAN score plots for model 35.



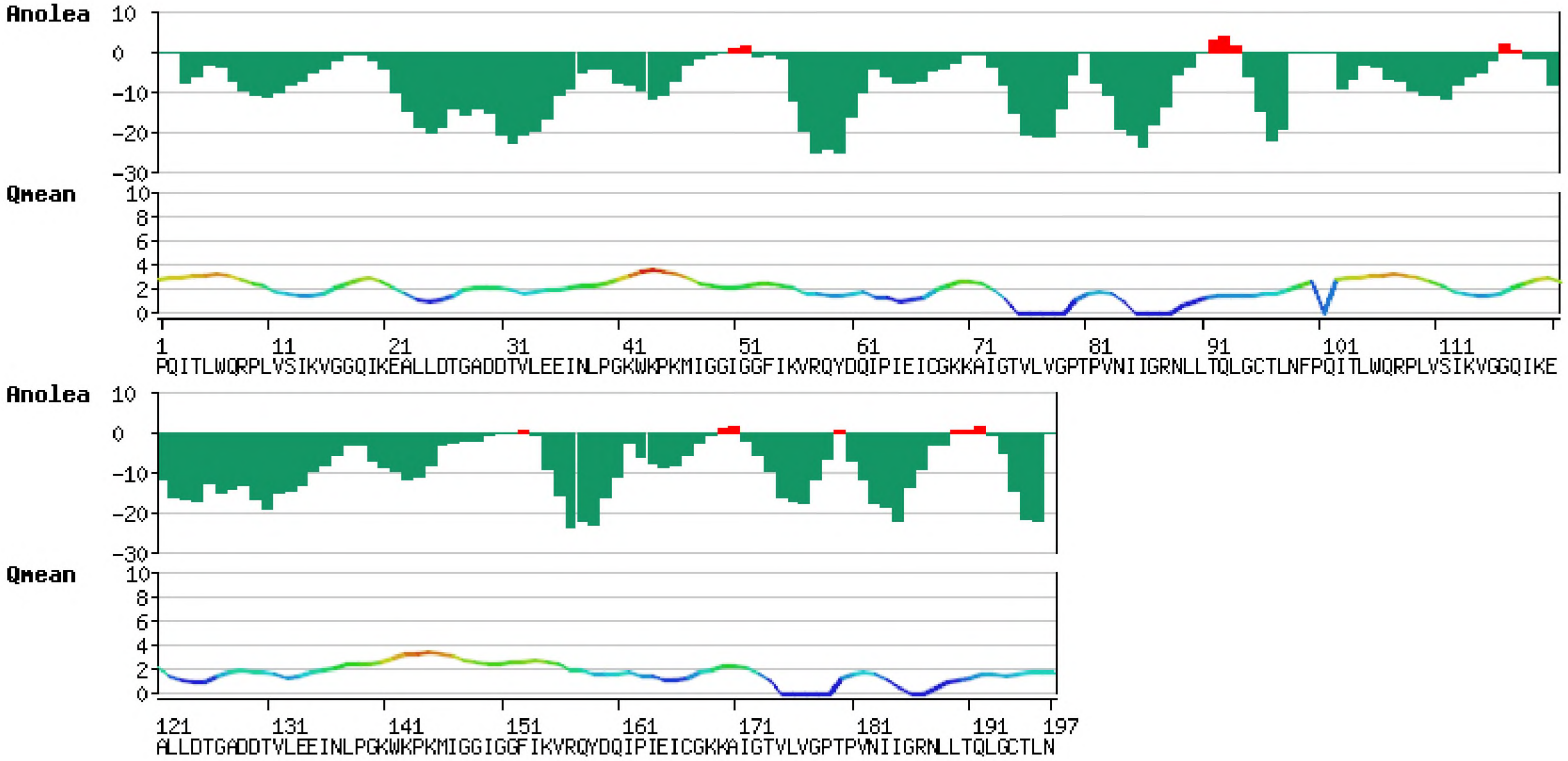
Supplementary Figure 40. ANOLEA and QMEAN score plots for model 36.



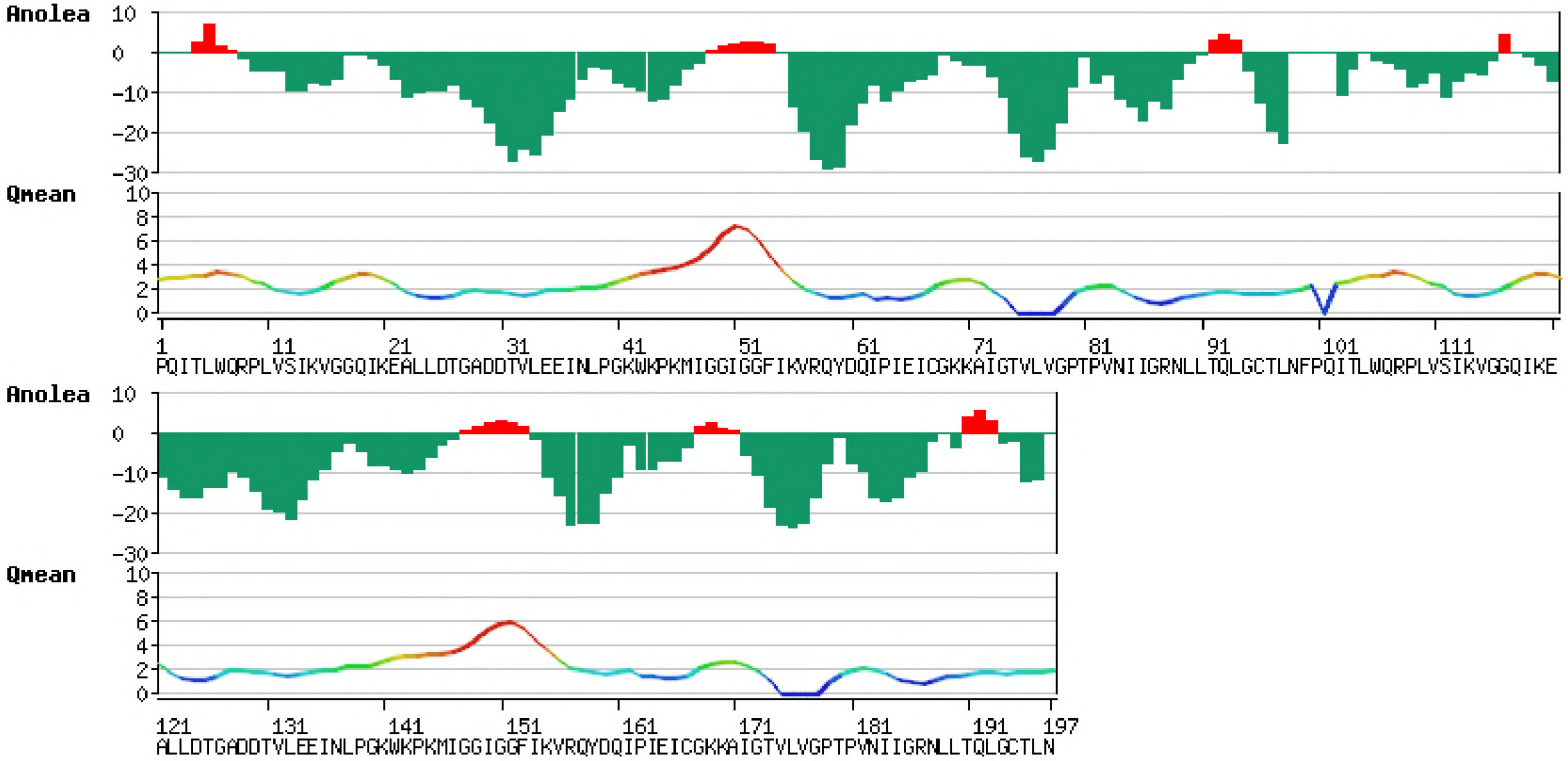
Supplementary Figure 41. ANOLEA and QMEAN score plots for model 37.



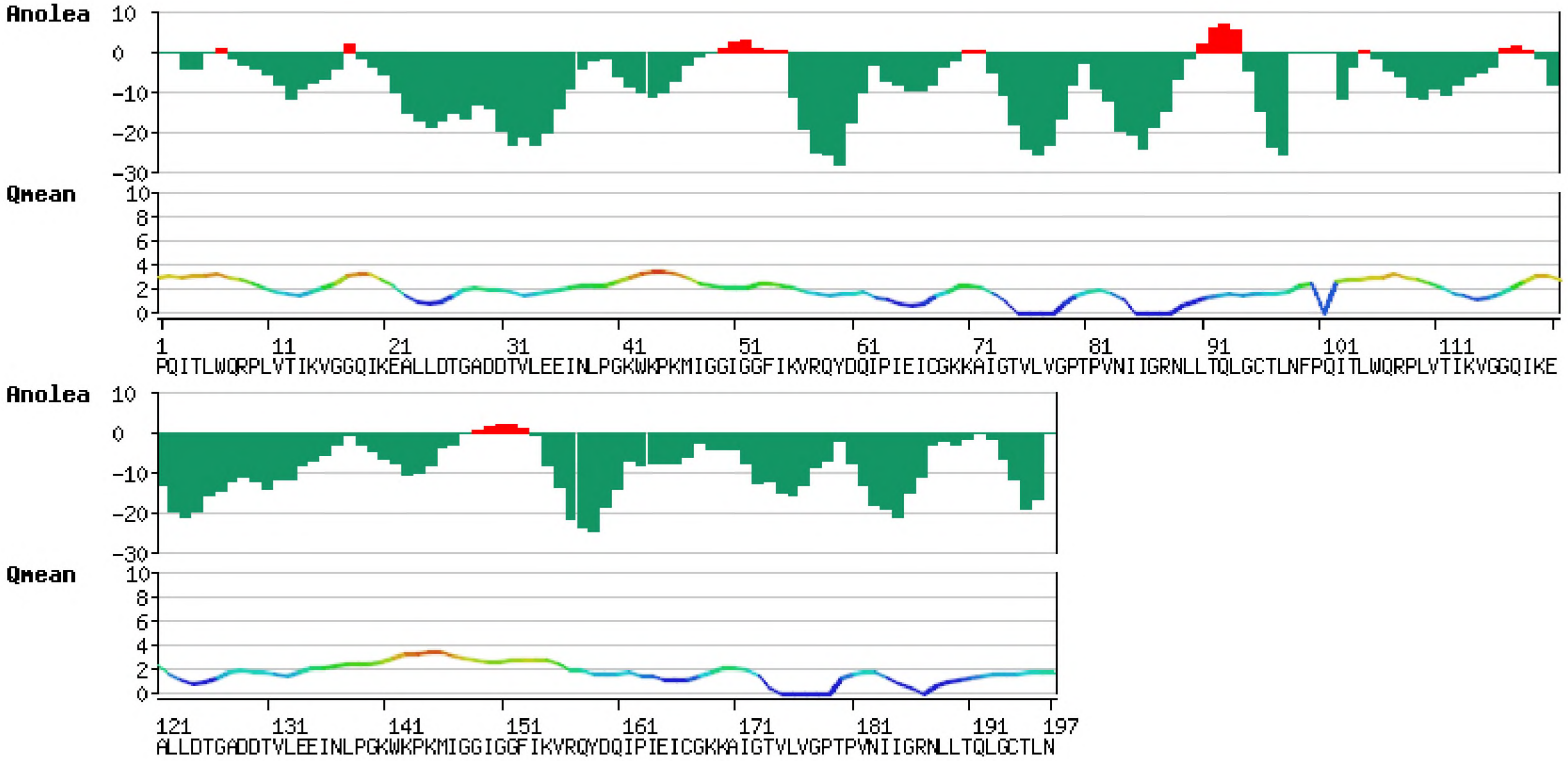
Supplementary Figure 42. ANOLEA and QMEAN score plots for model 38.



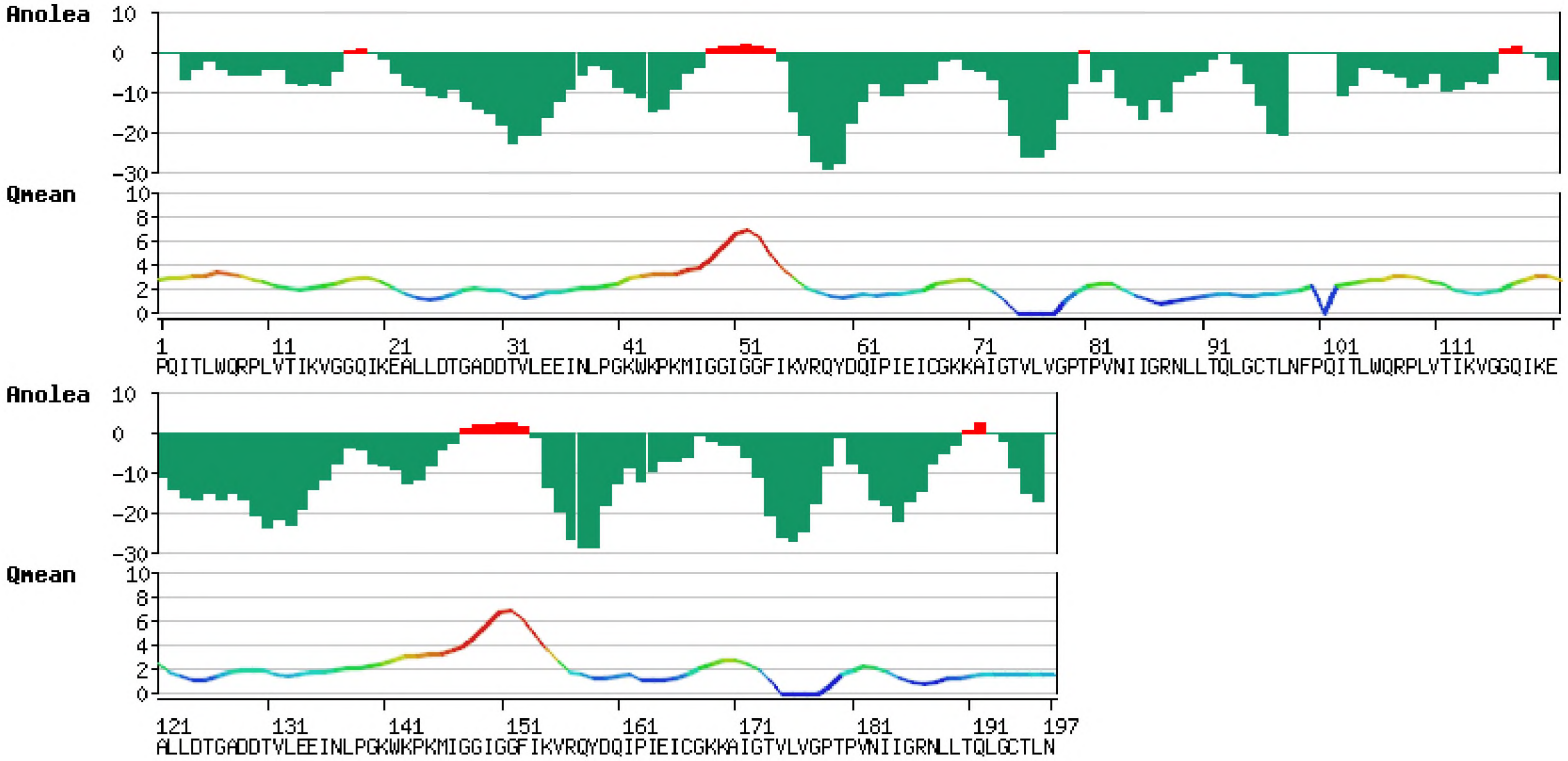
Supplementary Figure 43. ANOLEA and QMEAN score plots for model 39.



Supplementary Figure 44. ANOLEA and QMEAN score plots for model 40.



Supplementary Figure 45. ANOLEA and QMEAN score plots for model 41.



Supplementary Figure 46. ANOLEA and QMEAN score plots for model 42.

3. PDB files of SANCDB ligands corresponding to the network (ligand nodes) graph

1_apo_SANC00174.pdb

```
HETATM    1  C   LIG     1       8.667  21.359  20.704  1.00  0.00           C
HETATM    2  C   LIG     1       8.571  20.161  19.765  1.00  0.00           C
HETATM    3  C   LIG     1       7.826  19.023  20.479  1.00  0.00           C
HETATM    4  C   LIG     1       6.421  19.488  20.860  1.00  0.00           C
HETATM    5  C   LIG     1       5.580  19.651  19.580  1.00  0.00           C
HETATM    6  C   LIG     1       5.721  18.416  21.690  1.00  0.00           C
HETATM    7  C   LIG     1       5.586  17.249  21.114  1.00  0.00           C
HETATM    8  C   LIG     1       4.823  16.134  21.781  1.00  0.00           C
HETATM    9  C   LIG     1       3.674  16.801  22.552  1.00  0.00           C
HETATM   10  C   LIG     1       2.761  15.729  23.142  1.00  0.00           C
HETATM   11  C   LIG     1       3.539  14.840  24.105  1.00  0.00           C
HETATM   12  C   LIG     1       4.118  15.685  25.236  1.00  0.00           C
HETATM   13  C   LIG     1       2.998  16.273  26.095  1.00  0.00           C
HETATM   14  C   LIG     1       4.989  14.780  26.123  1.00  0.00           C
HETATM   15  C   LIG     1       4.998  16.808  24.684  1.00  0.00           C
HETATM   16  C   LIG     1       4.290  17.666  23.646  1.00  0.00           C
HETATM   17  C   LIG     1       5.231  18.728  23.065  1.00  0.00           C
HETATM   18  C   LIG     1       6.448  18.784  24.002  1.00  0.00           C
HETATM   19  C   LIG     1       4.544  20.087  23.131  1.00  0.00           C
HETATM   20  C   LIG     1       5.137  21.151  22.211  1.00  0.00           C
HETATM   21  C   LIG     1       6.501  20.760  21.685  1.00  0.00           C
HETATM   22  C   LIG     1       7.276  21.902  21.029  1.00  0.00           C
HETATM   23  C   LIG     1       6.548  22.475  19.823  1.00  0.00           C
HETATM   24  C   LIG     1       7.436  23.030  22.067  1.00  0.00           C
HETATM   25  C   LIG     1       8.319  24.132  21.485  1.00  0.00           C
HETATM   26  C   LIG     1       9.714  23.598  21.194  1.00  0.00           C
HETATM   27  C   LIG     1       9.689  22.381  20.269  1.00  0.00           C
HETATM   28  C   LIG     1      11.095  21.738  20.369  1.00  0.00           C
HETATM   29  C   LIG     1       9.506  22.829  18.821  1.00  0.00           C
HETATM   30  C   LIG     1       2.854  17.644  21.605  1.00  0.00           C
HETATM   31  O   LIG     1       2.194  18.563  22.032  1.00  0.00           O
HETATM   32  O   LIG     1       2.860  17.375  20.291  1.00  0.00           O
HETATM   33  H   LIG     1       2.016  17.451  19.823  1.00  0.00           H
HETATM   34  O   LIG     1      10.511  24.646  20.575  1.00  0.00           O
HETATM   35  C   LIG     1      11.488  25.188  21.320  1.00  0.00           C
HETATM   36  C   LIG     1      12.712  25.760  20.654  1.00  0.00           C
HETATM   37  O   LIG     1      11.386  25.217  22.523  1.00  0.00           O
```

23_apo_SANC00175.pdb

```
HETATM    1  C   LIG     1       6.380  19.919  22.500  1.00  0.00           C
HETATM    2  C   LIG     1       5.914  18.781  21.554  1.00  0.00           C
HETATM    3  C   LIG     1       7.006  18.685  20.451  1.00  0.00           C
HETATM    4  C   LIG     1       7.458  20.153  20.357  1.00  0.00           C
HETATM    5  C   LIG     1       6.331  20.976  19.729  1.00  0.00           C
HETATM    6  C   LIG     1       8.764  20.400  19.637  1.00  0.00           C
HETATM    7  C   LIG     1       9.141  21.881  19.788  1.00  0.00           C
HETATM    8  C   LIG     1       9.192  22.323  21.247  1.00  0.00           C
HETATM    9  C   LIG     1      10.385  21.604  21.890  1.00  0.00           C
HETATM   10  C   LIG     1       7.908  21.962  22.005  1.00  0.00           C
HETATM   11  C   LIG     1       8.026  22.371  23.474  1.00  0.00           C
HETATM   12  C   LIG     1       8.269  23.884  23.570  1.00  0.00           C
HETATM   13  C   LIG     1       9.538  24.239  22.801  1.00  0.00           C
HETATM   14  C   LIG     1       9.962  25.689  23.043  1.00  0.00           C
HETATM   15  C   LIG     1       8.825  26.657  22.748  1.00  0.00           C
HETATM   16  C   LIG     1      10.327  25.832  24.535  1.00  0.00           C
```

HETATM 17 C LIG 1 10.810 27.253 24.803 1.00 0.00 C
HETATM 18 C LIG 1 12.070 27.555 24.003 1.00 0.00 C
HETATM 19 C LIG 1 11.903 27.288 22.508 1.00 0.00 C
HETATM 20 C LIG 1 11.204 28.472 21.841 1.00 0.00 C
HETATM 21 C LIG 1 13.335 27.222 21.919 1.00 0.00 C
HETATM 22 C LIG 1 11.220 25.969 22.231 1.00 0.00 C
HETATM 23 C LIG 1 10.955 25.731 20.746 1.00 0.00 C
HETATM 24 C LIG 1 10.664 24.235 20.553 1.00 0.00 C
HETATM 25 C LIG 1 9.413 23.844 21.339 1.00 0.00 C
HETATM 26 C LIG 1 8.208 24.524 20.670 1.00 0.00 C
HETATM 27 C LIG 1 7.651 20.482 21.860 1.00 0.00 C
HETATM 28 C LIG 1 6.697 19.342 23.858 1.00 0.00 C
HETATM 29 C LIG 1 5.580 19.005 24.813 1.00 0.00 C
HETATM 30 C LIG 1 7.942 19.135 24.209 1.00 0.00 C
HETATM 31 O LIG 1 12.429 28.931 24.191 1.00 0.00 O
HETATM 32 H LIG 1 13.204 29.062 24.755 1.00 0.00 H

34_apo_SANC00264.pdb

HETATM 1 C LIG 1 4.511 25.429 14.853 1.00 0.00 C
HETATM 2 C LIG 1 5.520 26.271 15.617 1.00 0.00 C
HETATM 3 C LIG 1 5.208 26.477 17.091 1.00 0.00 C
HETATM 4 C LIG 1 6.036 27.660 17.622 1.00 0.00 C
HETATM 5 C LIG 1 3.743 26.785 17.283 1.00 0.00 C
HETATM 6 O LIG 1 3.388 27.715 17.969 1.00 0.00 O
HETATM 7 C LIG 1 2.729 25.898 16.600 1.00 0.00 C
HETATM 8 C LIG 1 3.073 25.831 15.109 1.00 0.00 C
HETATM 9 C LIG 1 2.868 27.254 14.530 1.00 0.00 C
HETATM 10 C LIG 1 2.059 24.925 14.412 1.00 0.00 C
HETATM 11 C LIG 1 2.376 23.445 14.593 1.00 0.00 C
HETATM 12 C LIG 1 3.791 23.150 14.097 1.00 0.00 C
HETATM 13 C LIG 1 3.871 23.539 12.620 1.00 0.00 C
HETATM 14 C LIG 1 4.120 21.682 14.276 1.00 0.00 C
HETATM 15 C LIG 1 3.138 20.848 14.642 1.00 0.00 C
HETATM 16 C LIG 1 3.356 19.488 14.903 1.00 0.00 C
HETATM 17 C LIG 1 4.653 18.976 14.939 1.00 0.00 C
HETATM 18 C LIG 1 4.894 17.632 15.308 1.00 0.00 C
HETATM 19 C LIG 1 3.757 16.676 15.562 1.00 0.00 C
HETATM 20 C LIG 1 6.171 17.187 15.453 1.00 0.00 C
HETATM 21 C LIG 1 7.257 18.063 15.225 1.00 0.00 C
HETATM 22 O LIG 1 8.405 17.680 15.400 1.00 0.00 O
HETATM 23 C LIG 1 7.014 19.379 14.783 1.00 0.00 C
HETATM 24 C LIG 1 5.741 19.822 14.618 1.00 0.00 C
HETATM 25 C LIG 1 5.520 21.201 14.048 1.00 0.00 C
HETATM 26 C LIG 1 5.776 21.059 12.541 1.00 0.00 C
HETATM 27 C LIG 1 6.557 22.158 14.619 1.00 0.00 C
HETATM 28 C LIG 1 6.205 23.627 14.425 1.00 0.00 C
HETATM 29 C LIG 1 4.797 23.943 14.939 1.00 0.00 C
HETATM 30 C LIG 1 4.720 23.451 16.390 1.00 0.00 C
HETATM 31 O LIG 1 6.401 15.896 15.816 1.00 0.00 O
HETATM 32 H LIG 1 6.559 15.300 15.071 1.00 0.00 H

3_apo_SANC00290.pdb

HETATM 1 C LIG 1 4.809 16.119 19.084 1.00 0.00 C
HETATM 2 C LIG 1 4.238 16.899 20.271 1.00 0.00 C
HETATM 3 C LIG 1 5.383 17.549 21.051 1.00 0.00 C
HETATM 4 C LIG 1 6.149 18.501 20.132 1.00 0.00 C
HETATM 5 C LIG 1 6.727 17.715 18.954 1.00 0.00 C
HETATM 6 C LIG 1 7.536 18.629 18.028 1.00 0.00 C
HETATM 7 C LIG 1 8.644 19.345 18.804 1.00 0.00 C
HETATM 8 C LIG 1 7.974 20.146 19.922 1.00 0.00 C

HETATM 9 C LIG 1 6.948 21.121 19.340 1.00 0.00 C
HETATM 10 C LIG 1 7.317 19.148 20.879 1.00 0.00 C
HETATM 11 C LIG 1 6.948 20.030 22.064 1.00 0.00 C
HETATM 12 C LIG 1 7.993 21.146 22.087 1.00 0.00 C
HETATM 13 O LIG 1 7.322 22.398 21.762 1.00 0.00 O
HETATM 14 C LIG 1 8.143 23.144 21.018 1.00 0.00 C
HETATM 15 O LIG 1 8.045 24.339 20.870 1.00 0.00 O
HETATM 16 C LIG 1 9.208 22.286 20.390 1.00 0.00 C
HETATM 17 C LIG 1 9.122 22.344 18.860 1.00 0.00 C
HETATM 18 C LIG 1 10.531 22.795 18.467 1.00 0.00 C
HETATM 19 C LIG 1 11.042 23.431 19.735 1.00 0.00 C
HETATM 20 O LIG 1 11.764 24.398 19.813 1.00 0.00 O
HETATM 21 O LIG 1 10.544 22.723 20.764 1.00 0.00 O
HETATM 22 C LIG 1 8.920 20.872 20.891 1.00 0.00 C
HETATM 23 C LIG 1 5.582 17.070 18.171 1.00 0.00 C
HETATM 24 C LIG 1 4.641 18.161 17.655 1.00 0.00 C
HETATM 25 C LIG 1 6.153 16.288 16.986 1.00 0.00 C
HETATM 26 C LIG 1 5.008 15.646 16.200 1.00 0.00 C
HETATM 27 C LIG 1 4.233 14.694 17.114 1.00 0.00 C
HETATM 28 C LIG 1 3.663 15.476 18.299 1.00 0.00 C
HETATM 29 N LIG 1 3.133 14.079 16.360 1.00 0.00 N
HETATM 30 H LIG 1 2.651 13.392 16.921 1.00 0.00 H
HETATM 31 H LIG 1 2.493 14.780 16.017 1.00 0.00 H

41_apo_SANC00342.pdb

HETATM 1 C LIG 1 7.793 21.888 20.630 1.00 0.00 C
HETATM 2 C LIG 1 6.932 22.776 21.188 1.00 0.00 C
HETATM 3 O LIG 1 7.313 23.992 21.591 1.00 0.00 O
HETATM 4 C LIG 1 8.578 24.446 21.481 1.00 0.00 C
HETATM 5 C LIG 1 8.903 25.723 21.924 1.00 0.00 C
HETATM 6 C LIG 1 10.198 26.188 21.809 1.00 0.00 C
HETATM 7 O LIG 1 10.536 27.431 22.237 1.00 0.00 O
HETATM 8 C LIG 1 11.845 27.600 22.779 1.00 0.00 C
HETATM 9 C LIG 1 12.138 29.093 22.938 1.00 0.00 C
HETATM 10 C LIG 1 11.916 26.924 24.150 1.00 0.00 C
HETATM 11 C LIG 1 12.878 26.987 21.870 1.00 0.00 C
HETATM 12 C LIG 1 12.569 25.917 21.139 1.00 0.00 C
HETATM 13 C LIG 1 11.199 25.376 21.244 1.00 0.00 C
HETATM 14 C LIG 1 10.891 24.095 20.798 1.00 0.00 C
HETATM 15 C LIG 1 9.575 23.630 20.917 1.00 0.00 C
HETATM 16 C LIG 1 9.204 22.279 20.466 1.00 0.00 C
HETATM 17 O LIG 1 10.024 21.521 19.976 1.00 0.00 O
HETATM 18 C LIG 1 7.323 20.549 20.199 1.00 0.00 C
HETATM 19 C LIG 1 7.472 20.152 18.874 1.00 0.00 C
HETATM 20 C LIG 1 7.028 18.895 18.481 1.00 0.00 C
HETATM 21 C LIG 1 7.166 18.432 17.083 1.00 0.00 C
HETATM 22 C LIG 1 6.433 18.032 19.414 1.00 0.00 C
HETATM 23 C LIG 1 6.440 17.386 16.694 1.00 0.00 C
HETATM 24 C LIG 1 5.502 16.719 17.666 1.00 0.00 C
HETATM 25 C LIG 1 4.134 17.402 17.597 1.00 0.00 C
HETATM 26 O LIG 1 6.009 16.812 18.998 1.00 0.00 O
HETATM 27 C LIG 1 5.348 15.245 17.285 1.00 0.00 C
HETATM 28 C LIG 1 6.288 18.431 20.730 1.00 0.00 C
HETATM 29 C LIG 1 6.730 19.683 21.126 1.00 0.00 C
HETATM 30 O LIG 1 6.587 20.073 22.419 1.00 0.00 O
HETATM 31 H LIG 1 5.685 19.990 22.758 1.00 0.00 H
HETATM 32 O LIG 1 11.852 23.307 20.253 1.00 0.00 O
HETATM 33 H LIG 1 12.001 23.463 19.311 1.00 0.00 H

7_apo_SANC00347.pdb

HETATM 1 C LIG 1 5.677 15.386 18.171 1.00 0.00 C
HETATM 2 C LIG 1 4.542 16.412 18.130 1.00 0.00 C
HETATM 3 C LIG 1 3.452 15.921 17.174 1.00 0.00 C
HETATM 4 C LIG 1 3.957 16.563 19.509 1.00 0.00 C
HETATM 5 C LIG 1 4.468 17.457 20.353 1.00 0.00 C
HETATM 6 C LIG 1 5.602 18.287 19.888 1.00 0.00 C
HETATM 7 C LIG 1 6.395 18.980 20.810 1.00 0.00 C
HETATM 8 C LIG 1 7.429 19.731 20.306 1.00 0.00 C
HETATM 9 C LIG 1 7.688 19.812 18.918 1.00 0.00 C
HETATM 10 C LIG 1 6.901 19.119 18.006 1.00 0.00 C
HETATM 11 C LIG 1 5.855 18.358 18.511 1.00 0.00 C
HETATM 12 O LIG 1 5.051 17.661 17.663 1.00 0.00 O
HETATM 13 O LIG 1 8.751 20.635 18.673 1.00 0.00 O
HETATM 14 C LIG 1 8.776 21.547 19.763 1.00 0.00 C
HETATM 15 O LIG 1 9.948 22.252 20.089 1.00 0.00 O
HETATM 16 C LIG 1 9.860 22.982 21.246 1.00 0.00 C
HETATM 17 C LIG 1 10.733 24.038 21.442 1.00 0.00 C
HETATM 18 C LIG 1 10.645 24.812 22.586 1.00 0.00 C
HETATM 19 O LIG 1 11.495 25.848 22.797 1.00 0.00 O
HETATM 20 C LIG 1 10.970 26.991 23.474 1.00 0.00 C
HETATM 21 C LIG 1 12.122 27.919 23.864 1.00 0.00 C
HETATM 22 C LIG 1 10.016 27.737 22.538 1.00 0.00 C
HETATM 23 C LIG 1 10.224 26.576 24.714 1.00 0.00 C
HETATM 24 C LIG 1 9.594 25.403 24.751 1.00 0.00 C
HETATM 25 C LIG 1 9.664 24.538 23.555 1.00 0.00 C
HETATM 26 C LIG 1 8.785 23.472 23.381 1.00 0.00 C
HETATM 27 C LIG 1 8.897 22.677 22.229 1.00 0.00 C
HETATM 28 C LIG 1 8.015 21.502 22.057 1.00 0.00 C
HETATM 29 O LIG 1 7.027 21.302 22.734 1.00 0.00 O
HETATM 30 C LIG 1 8.466 20.580 20.955 1.00 0.00 C
HETATM 31 O LIG 1 9.601 19.808 21.353 1.00 0.00 O
HETATM 32 H LIG 1 10.060 20.154 22.130 1.00 0.00 H
HETATM 33 O LIG 1 7.836 23.205 24.311 1.00 0.00 O
HETATM 34 H LIG 1 7.756 23.881 24.997 1.00 0.00 H

28_apo_SANC00380.pdb

HETATM 1 C LIG 1 7.204 21.655 15.393 1.00 0.00 C
HETATM 2 C LIG 1 8.656 21.836 15.834 1.00 0.00 C
HETATM 3 C LIG 1 8.662 22.743 17.050 1.00 0.00 C
HETATM 4 C LIG 1 7.923 23.910 17.029 1.00 0.00 C
HETATM 5 C LIG 1 7.860 24.695 18.161 1.00 0.00 C
HETATM 6 C LIG 1 8.522 24.308 19.322 1.00 0.00 C
HETATM 7 C LIG 1 9.339 23.184 19.303 1.00 0.00 C
HETATM 8 C LIG 1 9.397 22.399 18.167 1.00 0.00 C
HETATM 9 O LIG 1 8.253 24.945 20.469 1.00 0.00 O
HETATM 10 C LIG 1 9.095 25.596 21.282 1.00 0.00 C
HETATM 11 C LIG 1 10.346 26.023 20.871 1.00 0.00 C
HETATM 12 C LIG 1 11.167 26.733 21.739 1.00 0.00 C
HETATM 13 C LIG 1 10.687 27.041 23.003 1.00 0.00 C
HETATM 14 C LIG 1 9.443 26.618 23.419 1.00 0.00 C
HETATM 15 C LIG 1 8.902 27.081 24.755 1.00 0.00 C
HETATM 16 C LIG 1 7.371 27.116 24.600 1.00 0.00 C
HETATM 17 N LIG 1 6.933 25.746 24.315 1.00 0.00 N
HETATM 18 C LIG 1 5.480 25.613 24.482 1.00 0.00 C
HETATM 19 C LIG 1 7.361 25.228 23.022 1.00 0.00 C
HETATM 20 C LIG 1 7.519 23.703 23.136 1.00 0.00 C
HETATM 21 C LIG 1 7.198 23.062 21.812 1.00 0.00 C
HETATM 22 C LIG 1 7.980 22.034 21.317 1.00 0.00 C
HETATM 23 C LIG 1 7.739 21.517 20.059 1.00 0.00 C
HETATM 24 C LIG 1 6.699 22.025 19.296 1.00 0.00 C

HETATM 25 C LIG 1 5.851 22.987 19.842 1.00 0.00 C
HETATM 26 C LIG 1 6.099 23.505 21.094 1.00 0.00 C
HETATM 27 O LIG 1 6.550 21.705 18.008 1.00 0.00 O
HETATM 28 C LIG 1 6.334 20.526 17.432 1.00 0.00 C
HETATM 29 C LIG 1 5.839 19.428 18.110 1.00 0.00 C
HETATM 30 C LIG 1 5.612 18.236 17.426 1.00 0.00 C
HETATM 31 C LIG 1 5.888 18.183 16.071 1.00 0.00 C
HETATM 32 C LIG 1 6.394 19.269 15.381 1.00 0.00 C
HETATM 33 C LIG 1 6.680 19.118 13.917 1.00 0.00 C
HETATM 34 C LIG 1 7.649 20.225 13.481 1.00 0.00 C
HETATM 35 N LIG 1 7.123 21.520 13.935 1.00 0.00 N
HETATM 36 C LIG 1 5.749 21.727 13.460 1.00 0.00 C
HETATM 37 C LIG 1 6.624 20.443 16.067 1.00 0.00 C
HETATM 38 C LIG 1 8.673 25.825 22.600 1.00 0.00 C
HETATM 39 O LIG 1 10.773 25.747 19.608 1.00 0.00 O
HETATM 40 C LIG 1 10.774 26.827 18.672 1.00 0.00 C
HETATM 41 O LIG 1 12.413 27.117 21.356 1.00 0.00 O
HETATM 42 C LIG 1 12.614 28.511 21.108 1.00 0.00 C
HETATM 43 O LIG 1 5.576 19.511 19.443 1.00 0.00 O
HETATM 44 C LIG 1 4.206 19.644 19.827 1.00 0.00 C
HETATM 45 O LIG 1 5.127 17.145 18.078 1.00 0.00 O
HETATM 46 C LIG 1 4.607 16.086 17.272 1.00 0.00 C

6_apo_SANC00381.pdb

HETATM 1 C LIG 1 8.163 23.853 16.712 1.00 0.00 C
HETATM 2 C LIG 1 9.157 23.772 17.867 1.00 0.00 C
HETATM 3 C LIG 1 9.000 24.904 18.883 1.00 0.00 C
HETATM 4 C LIG 1 8.563 26.157 18.441 1.00 0.00 C
HETATM 5 C LIG 1 8.271 27.087 19.381 1.00 0.00 C
HETATM 6 C LIG 1 8.468 26.894 20.787 1.00 0.00 C
HETATM 7 O LIG 1 7.935 27.875 21.558 1.00 0.00 O
HETATM 8 C LIG 1 7.395 27.716 22.794 1.00 0.00 C
HETATM 9 C LIG 1 8.239 27.173 23.727 1.00 0.00 C
HETATM 10 O LIG 1 9.474 26.715 23.293 1.00 0.00 O
HETATM 11 C LIG 1 9.534 25.554 22.525 1.00 0.00 C
HETATM 12 C LIG 1 9.106 25.773 21.198 1.00 0.00 C
HETATM 13 C LIG 1 9.270 24.753 20.253 1.00 0.00 C
HETATM 14 C LIG 1 7.945 27.087 24.996 1.00 0.00 C
HETATM 15 C LIG 1 6.626 27.396 25.475 1.00 0.00 C
HETATM 16 C LIG 1 5.771 27.802 24.517 1.00 0.00 C
HETATM 17 C LIG 1 4.338 28.071 24.821 1.00 0.00 C
HETATM 18 C LIG 1 3.508 27.253 23.834 1.00 0.00 C
HETATM 19 N LIG 1 3.770 27.775 22.410 1.00 0.00 N
HETATM 20 C LIG 1 2.557 28.721 22.303 1.00 0.00 C
HETATM 21 C LIG 1 4.921 28.610 22.390 1.00 0.00 C
HETATM 22 C LIG 1 5.163 29.146 21.044 1.00 0.00 C
HETATM 23 C LIG 1 5.122 28.303 19.859 1.00 0.00 C
HETATM 24 C LIG 1 6.079 28.458 18.911 1.00 0.00 C
HETATM 25 C LIG 1 6.153 27.561 17.848 1.00 0.00 C
HETATM 26 C LIG 1 5.329 26.466 17.698 1.00 0.00 C
HETATM 27 O LIG 1 5.696 25.438 17.001 1.00 0.00 O
HETATM 28 C LIG 1 5.713 24.068 17.333 1.00 0.00 C
HETATM 29 C LIG 1 4.701 23.454 17.943 1.00 0.00 C
HETATM 30 O LIG 1 3.599 24.070 18.547 1.00 0.00 O
HETATM 31 C LIG 1 3.210 25.383 18.361 1.00 0.00 C
HETATM 32 C LIG 1 4.243 26.398 18.650 1.00 0.00 C
HETATM 33 C LIG 1 4.104 27.291 19.694 1.00 0.00 C
HETATM 34 C LIG 1 4.645 22.125 18.074 1.00 0.00 C
HETATM 35 C LIG 1 5.721 21.338 17.674 1.00 0.00 C

HETATM 36 C LIG 1 6.861 21.952 17.212 1.00 0.00 C
HETATM 37 C LIG 1 8.000 21.089 16.703 1.00 0.00 C
HETATM 38 C LIG 1 9.133 21.773 16.173 1.00 0.00 C
HETATM 39 N LIG 1 8.876 23.072 15.613 1.00 0.00 N
HETATM 40 C LIG 1 7.819 22.871 14.560 1.00 0.00 C
HETATM 41 C LIG 1 6.875 23.312 17.079 1.00 0.00 C
HETATM 42 C LIG 1 6.061 28.026 23.157 1.00 0.00 C
HETATM 44 C LIG 1 8.368 25.522 26.774 1.00 0.00 C
HETATM 45 O LIG 1 3.610 21.490 18.668 1.00 0.00 O
HETATM 46 C LIG 1 3.503 20.082 18.453 1.00 0.00 C

24_apo_SANC00383.pdb

HETATM 1 C LIG 1 8.686 26.106 23.933 1.00 0.00 C
HETATM 2 C LIG 1 9.575 26.158 22.681 1.00 0.00 C
HETATM 3 C LIG 1 9.182 25.040 21.750 1.00 0.00 C
HETATM 4 C LIG 1 9.825 23.819 21.847 1.00 0.00 C
HETATM 5 C LIG 1 9.416 22.765 21.059 1.00 0.00 C
HETATM 6 C LIG 1 8.347 22.923 20.178 1.00 0.00 C
HETATM 7 C LIG 1 7.794 24.182 19.996 1.00 0.00 C
HETATM 8 C LIG 1 8.210 25.237 20.787 1.00 0.00 C
HETATM 9 O LIG 1 7.791 21.821 19.665 1.00 0.00 O
HETATM 10 C LIG 1 7.539 21.474 18.404 1.00 0.00 C
HETATM 11 C LIG 1 8.251 22.000 17.340 1.00 0.00 C
HETATM 12 C LIG 1 8.023 21.553 16.044 1.00 0.00 C
HETATM 13 C LIG 1 7.087 20.552 15.828 1.00 0.00 C
HETATM 14 C LIG 1 6.370 20.041 16.890 1.00 0.00 C
HETATM 15 C LIG 1 5.486 18.832 16.673 1.00 0.00 C
HETATM 16 C LIG 1 5.525 18.039 17.989 1.00 0.00 C
HETATM 17 N LIG 1 4.970 18.871 19.048 1.00 0.00 N
HETATM 18 C LIG 1 4.775 18.095 20.280 1.00 0.00 C
HETATM 19 C LIG 1 5.663 20.119 19.323 1.00 0.00 C
HETATM 20 C LIG 1 4.602 21.184 19.629 1.00 0.00 C
HETATM 21 C LIG 1 4.989 22.032 20.805 1.00 0.00 C
HETATM 22 C LIG 1 6.235 21.887 21.394 1.00 0.00 C
HETATM 23 C LIG 1 6.644 22.749 22.391 1.00 0.00 C
HETATM 24 C LIG 1 5.786 23.738 22.846 1.00 0.00 C
HETATM 25 O LIG 1 6.206 24.689 23.706 1.00 0.00 O
HETATM 26 C LIG 1 6.192 26.013 23.425 1.00 0.00 C
HETATM 27 C LIG 1 5.059 26.666 22.956 1.00 0.00 C
HETATM 28 O LIG 1 3.888 26.078 22.640 1.00 0.00 O
HETATM 29 C LIG 1 3.534 24.761 22.994 1.00 0.00 C
HETATM 30 C LIG 1 4.483 23.777 22.364 1.00 0.00 C
HETATM 31 C LIG 1 4.091 22.961 21.324 1.00 0.00 C
HETATM 32 C LIG 1 5.089 28.058 22.791 1.00 0.00 C
HETATM 33 C LIG 1 6.237 28.784 23.044 1.00 0.00 C
HETATM 34 C LIG 1 7.380 28.122 23.436 1.00 0.00 C
HETATM 35 C LIG 1 8.657 28.865 23.697 1.00 0.00 C
HETATM 36 C LIG 1 9.471 28.267 24.822 1.00 0.00 C
HETATM 37 N LIG 1 9.343 26.835 25.015 1.00 0.00 N
HETATM 38 C LIG 1 8.718 26.532 26.310 1.00 0.00 C
HETATM 39 C LIG 1 7.360 26.744 23.599 1.00 0.00 C
HETATM 40 C LIG 1 6.507 20.553 18.163 1.00 0.00 C
HETATM 41 O LIG 1 9.183 22.964 17.565 1.00 0.00 O
HETATM 42 C LIG 1 10.554 22.557 17.554 1.00 0.00 C
HETATM 43 O LIG 1 8.708 22.092 15.001 1.00 0.00 O
HETATM 44 C LIG 1 8.017 23.053 14.199 1.00 0.00 C
HETATM 45 O LIG 1 3.968 28.705 22.374 1.00 0.00 O
HETATM 46 C LIG 1 3.996 29.272 21.063 1.00 0.00 C

38_apo_SANC00386.pdb

```
HETATM    1  C   LIG     1       8.878  25.767  23.475  1.00  0.00           C
HETATM    2  C   LIG     1       9.713  25.441  22.227  1.00  0.00           C
HETATM    3  C   LIG     1       9.184  24.180  21.595  1.00  0.00           C
HETATM    4  C   LIG     1       9.722  22.959  21.963  1.00  0.00           C
HETATM    5  C   LIG     1       9.189  21.794  21.455  1.00  0.00           C
HETATM    6  C   LIG     1       8.100  21.843  20.585  1.00  0.00           C
HETATM    7  C   LIG     1       7.652  23.072  20.123  1.00  0.00           C
HETATM    8  C   LIG     1       8.192  24.238  20.634  1.00  0.00           C
HETATM    9  O   LIG     1       7.429  20.710  20.359  1.00  0.00           O
HETATM   10  C   LIG     1       7.093  20.104  19.221  1.00  0.00           C
HETATM   11  C   LIG     1       7.802  20.298  18.048  1.00  0.00           C
HETATM   12  C   LIG     1       7.480  19.585  16.899  1.00  0.00           C
HETATM   13  C   LIG     1       6.452  18.654  16.946  1.00  0.00           C
HETATM   14  C   LIG     1       5.739  18.475  18.114  1.00  0.00           C
HETATM   15  C   LIG     1       4.745  17.339  18.208  1.00  0.00           C
HETATM   16  C   LIG     1       4.772  16.873  19.672  1.00  0.00           C
HETATM   17  N   LIG     1       4.339  17.978  20.517  1.00  0.00           N
HETATM   18  C   LIG     1       4.130  17.533  21.901  1.00  0.00           C
HETATM   19  C   LIG     1       5.148  19.185  20.475  1.00  0.00           C
HETATM   20  C   LIG     1       4.199  20.387  20.544  1.00  0.00           C
HETATM   21  C   LIG     1       4.708  21.445  21.477  1.00  0.00           C
HETATM   22  C   LIG     1       5.961  21.324  22.057  1.00  0.00           C
HETATM   23  C   LIG     1       6.487  22.352  22.813  1.00  0.00           C
HETATM   24  C   LIG     1       5.739  23.496  23.040  1.00  0.00           C
HETATM   25  O   LIG     1       6.276  24.577  23.643  1.00  0.00           O
HETATM   26  C   LIG     1       6.365  25.795  23.058  1.00  0.00           C
HETATM   27  C   LIG     1       5.275  26.425  22.472  1.00  0.00           C
HETATM   28  O   LIG     1       4.044  25.893  22.329  1.00  0.00           O
HETATM   29  C   LIG     1       3.592  24.734  22.991  1.00  0.00           C
HETATM   30  C   LIG     1       4.424  23.545  22.591  1.00  0.00           C
HETATM   31  C   LIG     1       3.917  22.550  21.782  1.00  0.00           C
HETATM   32  C   LIG     1       5.418  27.731  21.983  1.00  0.00           C
HETATM   33  C   LIG     1       6.634  28.383  22.032  1.00  0.00           C
HETATM   34  C   LIG     1       7.732  27.726  22.544  1.00  0.00           C
HETATM   35  C   LIG     1       9.079  28.385  22.595  1.00  0.00           C
HETATM   36  C   LIG     1       9.885  27.991  23.812  1.00  0.00           C
HETATM   37  N   LIG     1       9.642  26.663  24.340  1.00  0.00           N
HETATM   38  C   LIG     1       9.050  26.730  25.683  1.00  0.00           C
HETATM   39  C   LIG     1       7.599  26.432  23.028  1.00  0.00           C
HETATM   40  C   LIG     1       5.976  19.254  19.227  1.00  0.00           C
HETATM   41  O   LIG     1       8.823  21.195  18.019  1.00  0.00           O
HETATM   42  H   LIG     1       9.409  21.100  17.255  1.00  0.00           H
HETATM   43  O   LIG     1       8.164  19.799  15.744  1.00  0.00           O
HETATM   44  C   LIG     1       7.388  19.995  14.558  1.00  0.00           C
HETATM   45  O   LIG     1       4.340  28.366  21.449  1.00  0.00           O
HETATM   46  C   LIG     1       3.636  29.272  22.299  1.00  0.00           C
```

4_apo_SANC00400.pdb

```
HETATM    1  C   LIG     1       3.710  19.857  18.089  1.00  0.00           C
HETATM    2  C   LIG     1       2.530  20.409  18.896  1.00  0.00           C
HETATM    3  C   LIG     1       1.676  21.301  17.991  1.00  0.00           C
HETATM    4  O   LIG     1       2.485  22.351  17.458  1.00  0.00           O
HETATM    5  C   LIG     1       3.577  21.896  16.656  1.00  0.00           C
HETATM    6  C   LIG     1       4.508  21.027  17.504  1.00  0.00           C
HETATM    7  O   LIG     1       4.300  23.022  16.153  1.00  0.00           O
HETATM    8  C   LIG     1       3.545  24.234  16.126  1.00  0.00           C
HETATM    9  C   LIG     1       4.398  25.379  16.681  1.00  0.00           C
HETATM   10  C   LIG     1       3.658  26.675  16.522  1.00  0.00           C
HETATM   11  C   LIG     1       4.791  25.083  18.118  1.00  0.00           C
HETATM   12  C   LIG     1       5.675  25.452  15.841  1.00  0.00           C
```

```
HETATM  13  C   LIG     1       3.103  27.010  15.148  1.00  0.00           C
HETATM  14  C   LIG     1       3.469  27.524  17.488  1.00  0.00           C
HETATM  15  C   LIG     1       2.279  25.815  14.656  1.00  0.00           C
HETATM  16  C   LIG     1       3.143  24.553  14.685  1.00  0.00           C
HETATM  17  C   LIG     1       3.964  27.297  18.886  1.00  0.00           C
HETATM  18  C   LIG     1       5.190  26.377  18.844  1.00  0.00           C
HETATM  19  C   LIG     1       5.591  26.002  20.256  1.00  0.00           C
HETATM  20  C   LIG     1       6.864  25.100  20.153  1.00  0.00           C
HETATM  21  C   LIG     1       6.038  27.141  21.164  1.00  0.00           C
HETATM  22  C   LIG     1       6.412  23.775  19.561  1.00  0.00           C
HETATM  23  C   LIG     1       7.328  25.050  21.612  1.00  0.00           C
HETATM  24  C   LIG     1       7.940  25.739  19.272  1.00  0.00           C
HETATM  25  C   LIG     1       5.920  24.057  18.134  1.00  0.00           C
HETATM  26  C   LIG     1       7.105  26.526  22.101  1.00  0.00           C
HETATM  27  C   LIG     1       8.854  24.874  21.659  1.00  0.00           C
HETATM  28  O   LIG     1       8.365  27.184  21.858  1.00  0.00           O
HETATM  29  C   LIG     1       9.378  26.219  22.195  1.00  0.00           C
HETATM  30  O   LIG     1       9.547  26.154  23.607  1.00  0.00           O
HETATM  31  C   LIG     1      10.693  26.577  21.509  1.00  0.00           C
HETATM  32  C   LIG     1      10.001  27.373  24.201  1.00  0.00           C
HETATM  33  C   LIG     1      11.358  27.756  23.605  1.00  0.00           C
HETATM  34  C   LIG     1      11.220  27.896  22.086  1.00  0.00           C
HETATM  35  C   LIG     1      11.821  29.087  24.201  1.00  0.00           C
HETATM  36  C   LIG     1       9.230  23.736  22.610  1.00  0.00           C
HETATM  37  O   LIG     1       1.837  26.058  13.320  1.00  0.00           O
HETATM  38  H   LIG     1       0.897  26.278  13.251  1.00  0.00           H
HETATM  39  O   LIG     1       5.569  20.526  16.689  1.00  0.00           O
HETATM  40  C   LIG     1       5.145  20.047  15.411  1.00  0.00           C
HETATM  41  C   LIG     1       6.168  20.455  14.347  1.00  0.00           C
HETATM  42  C   LIG     1       7.515  19.797  14.667  1.00  0.00           C
HETATM  43  C   LIG     1       7.324  18.278  14.740  1.00  0.00           C
HETATM  44  C   LIG     1       6.256  17.956  15.788  1.00  0.00           C
HETATM  45  O   LIG     1       5.039  18.623  15.447  1.00  0.00           O
HETATM  46  C   LIG     1       6.017  16.445  15.825  1.00  0.00           C
HETATM  47  O   LIG     1       5.719  20.023  13.061  1.00  0.00           O
HETATM  48  H   LIG     1       6.380  19.523  12.562  1.00  0.00           H
HETATM  49  O   LIG     1       8.458  20.112  13.641  1.00  0.00           O
HETATM  50  H   LIG     1       8.950  19.347  13.311  1.00  0.00           H
HETATM  51  O   LIG     1       8.558  17.659  15.108  1.00  0.00           O
HETATM  52  H   LIG     1       9.331  18.029  14.661  1.00  0.00           H
HETATM  53  O   LIG     1       4.553  19.083  18.943  1.00  0.00           O
HETATM  54  C   LIG     1       5.624  19.826  19.531  1.00  0.00           C
HETATM  55  C   LIG     1       5.767  19.436  21.004  1.00  0.00           C
HETATM  56  C   LIG     1       6.955  20.191  21.611  1.00  0.00           C
HETATM  57  C   LIG     1       8.215  19.874  20.800  1.00  0.00           C
HETATM  58  C   LIG     1       7.978  20.246  19.335  1.00  0.00           C
HETATM  59  O   LIG     1       6.841  19.533  18.841  1.00  0.00           O
HETATM  60  O   LIG     1       4.573  19.780  21.708  1.00  0.00           O
HETATM  61  H   LIG     1       4.090  19.019  22.059  1.00  0.00           H
HETATM  62  O   LIG     1       7.142  19.780  22.967  1.00  0.00           O
HETATM  63  H   LIG     1       7.811  20.291  23.443  1.00  0.00           H
HETATM  64  O   LIG     1       9.316  20.628  21.313  1.00  0.00           O
HETATM  65  H   LIG     1       9.896  20.990  20.629  1.00  0.00           H
HETATM  66  O   LIG     1       3.022  21.175  19.997  1.00  0.00           O
HETATM  67  H   LIG     1       3.986  21.211  20.052  1.00  0.00           H
```

29_apo_SANC00421.pdb

```
HETATM   1  C   LIG     1       8.893  19.999  20.810  1.00  0.00           C
HETATM   2  C   LIG     1       7.536  19.358  20.741  1.00  0.00           C
HETATM   3  C   LIG     1       6.956  18.992  19.395  1.00  0.00           C
```

```
HETATM    4  C   LIG     1      6.235  20.238  18.880  1.00  0.00           C
HETATM    5  C   LIG     1      5.997  17.828  19.579  1.00  0.00           C
HETATM    6  C   LIG     1      8.079  18.637  18.424  1.00  0.00           C
HETATM    7  C   LIG     1      5.444  17.223  18.303  1.00  0.00           C
HETATM    8  C   LIG     1      4.884  18.211  20.558  1.00  0.00           C
HETATM    9  C   LIG     1      6.526  17.113  17.252  1.00  0.00           C
HETATM   10  C   LIG     1      4.255  18.006  17.745  1.00  0.00           C
HETATM   11  C   LIG     1      4.958  15.788  18.608  1.00  0.00           C
HETATM   12  C   LIG     1      7.495  18.254  17.060  1.00  0.00           C
HETATM   13  O   LIG     1      6.612  16.117  16.569  1.00  0.00           O
HETATM   14  C   LIG     1      5.545  18.537  21.885  1.00  0.00           C
HETATM   15  O   LIG     1      4.974  18.304  22.933  1.00  0.00           O
HETATM   16  C   LIG     1      6.880  19.138  21.876  1.00  0.00           C
HETATM   17  C   LIG     1      7.533  19.511  23.184  1.00  0.00           C
HETATM   18  C   LIG     1      8.843  21.372  20.136  1.00  0.00           C
HETATM   19  C   LIG     1      9.726  22.353  20.911  1.00  0.00           C
HETATM   20  C   LIG     1     10.400  23.305  19.920  1.00  0.00           C
HETATM   21  C   LIG     1     11.200  24.403  20.653  1.00  0.00           C
HETATM   22  C   LIG     1     11.424  22.437  19.157  1.00  0.00           C
HETATM   23  C   LIG     1      9.373  23.910  18.960  1.00  0.00           C
HETATM   24  C   LIG     1     10.359  25.062  21.751  1.00  0.00           C
HETATM   25  O   LIG     1     11.471  25.441  19.711  1.00  0.00           O
HETATM   26  C   LIG     1      9.789  24.033  22.719  1.00  0.00           C
HETATM   27  C   LIG     1      8.866  23.102  21.922  1.00  0.00           C
HETATM   28  C   LIG     1      7.803  23.925  21.190  1.00  0.00           C
HETATM   29  C   LIG     1     12.266  25.167  18.579  1.00  0.00           C
HETATM   30  C   LIG     1     12.696  26.488  17.936  1.00  0.00           C
HETATM   31  C   LIG     1     13.510  24.392  19.015  1.00  0.00           C
HETATM   32  C   LIG     1     11.510  24.353  17.568  1.00  0.00           C
HETATM   33  O   LIG     1     10.902  24.887  16.672  1.00  0.00           O
HETATM   34  C   LIG     1     11.531  22.846  17.699  1.00  0.00           C
```

35_apo_SANC00422.pdb

```
HETATM    1  C   LIG     1      9.010  20.297  21.276  1.00  0.00           C
HETATM    2  C   LIG     1      7.593  19.882  21.001  1.00  0.00           C
HETATM    3  C   LIG     1      7.128  19.572  19.593  1.00  0.00           C
HETATM    4  C   LIG     1      6.388  20.818  19.100  1.00  0.00           C
HETATM    5  C   LIG     1      6.201  18.372  19.671  1.00  0.00           C
HETATM    6  C   LIG     1      8.305  19.292  18.670  1.00  0.00           C
HETATM    7  C   LIG     1      5.769  17.777  18.352  1.00  0.00           C
HETATM    8  C   LIG     1      5.054  18.684  20.596  1.00  0.00           C
HETATM    9  C   LIG     1      6.963  17.593  17.411  1.00  0.00           C
HETATM   10  C   LIG     1      4.679  18.595  17.662  1.00  0.00           C
HETATM   11  C   LIG     1      5.186  16.371  18.627  1.00  0.00           C
HETATM   12  C   LIG     1      7.789  18.868  17.292  1.00  0.00           C
HETATM   13  C   LIG     1      5.330  19.348  21.721  1.00  0.00           C
HETATM   14  C   LIG     1      6.709  19.781  21.994  1.00  0.00           C
HETATM   15  C   LIG     1      7.117  20.161  23.394  1.00  0.00           C
HETATM   16  O   LIG     1      6.484  17.216  16.115  1.00  0.00           O
HETATM   17  H   LIG     1      6.032  17.927  15.641  1.00  0.00           H
HETATM   18  C   LIG     1      9.449  21.333  20.240  1.00  0.00           C
HETATM   19  C   LIG     1     10.186  22.476  20.942  1.00  0.00           C
HETATM   20  C   LIG     1     10.774  23.413  19.884  1.00  0.00           C
HETATM   21  C   LIG     1     11.417  24.657  20.532  1.00  0.00           C
HETATM   22  C   LIG     1     11.916  22.611  19.222  1.00  0.00           C
HETATM   23  C   LIG     1      9.715  23.805  18.853  1.00  0.00           C
HETATM   24  C   LIG     1     10.472  25.303  21.548  1.00  0.00           C
HETATM   25  O   LIG     1     11.594  25.635  19.506  1.00  0.00           O
HETATM   26  C   LIG     1      9.995  24.301  22.593  1.00  0.00           C
HETATM   27  C   LIG     1      9.214  23.201  21.863  1.00  0.00           C
```

```
HETATM   28  C   LIG     1       8.086  23.824  21.039  1.00  0.00           C
HETATM   29  C   LIG     1      12.451  25.361  18.419  1.00  0.00           C
HETATM   30  C   LIG     1      12.744  26.663  17.671  1.00  0.00           C
HETATM   31  C   LIG     1      13.764  24.781  18.950  1.00  0.00           C
HETATM   32  C   LIG     1      11.827  24.378  17.470  1.00  0.00           C
HETATM   33  O   LIG     1      11.189  24.757  16.518  1.00  0.00           O
HETATM   34  C   LIG     1      12.020  22.902  17.736  1.00  0.00           C
```

2_apo_SANC00488.pdb

```
HETATM    1  C   LIG     1      21.339  27.015  11.486  1.00  0.00           C
HETATM    2  C   LIG     1      20.192  27.394  12.426  1.00  0.00           C
HETATM    3  C   LIG     1      19.377  26.141  12.766  1.00  0.00           C
HETATM    4  C   LIG     1      18.465  25.877  11.559  1.00  0.00           C
HETATM    5  C   LIG     1      17.496  27.064  11.589  1.00  0.00           C
HETATM    6  O   LIG     1      17.928  28.323  12.093  1.00  0.00           O
HETATM    7  C   LIG     1      17.371  27.442  13.070  1.00  0.00           C
HETATM    8  C   LIG     1      16.017  27.725  13.711  1.00  0.00           C
HETATM    9  C   LIG     1      16.064  29.024  14.522  1.00  0.00           C
HETATM   10  C   LIG     1      14.688  29.272  15.144  1.00  0.00           C
HETATM   11  C   LIG     1      14.309  28.137  16.096  1.00  0.00           C
HETATM   12  C   LIG     1      12.950  28.448  16.724  1.00  0.00           C
HETATM   13  C   LIG     1      12.393  27.267  17.474  1.00  0.00           C
HETATM   14  N   LIG     1      11.439  27.481  18.369  1.00  0.00           N
HETATM   15  C   LIG     1      10.904  26.474  19.041  1.00  0.00           C
HETATM   16  C   LIG     1       9.843  26.777  20.065  1.00  0.00           C
HETATM   17  C   LIG     1       8.986  25.549  20.377  1.00  0.00           C
HETATM   18  C   LIG     1       8.157  25.186  19.143  1.00  0.00           C
HETATM   19  C   LIG     1       9.908  24.377  20.742  1.00  0.00           C
HETATM   20  C   LIG     1       9.104  23.126  21.113  1.00  0.00           C
HETATM   21  C   LIG     1       7.768  23.455  21.784  1.00  0.00           C
HETATM   22  C   LIG     1       7.929  24.767  22.540  1.00  0.00           C
HETATM   23  C   LIG     1       6.729  25.222  23.341  1.00  0.00           C
HETATM   24  C   LIG     1       5.596  24.615  23.477  1.00  0.00           C
HETATM   25  C   LIG     1       4.626  25.418  24.297  1.00  0.00           C
HETATM   26  O   LIG     1       3.570  25.857  23.423  1.00  0.00           O
HETATM   27  C   LIG     1       4.038  27.076  22.832  1.00  0.00           C
HETATM   28  O   LIG     1       5.038  26.848  21.843  1.00  0.00           O
HETATM   29  C   LIG     1       4.653  27.481  20.617  1.00  0.00           C
HETATM   30  C   LIG     1       5.192  28.911  20.554  1.00  0.00           C
HETATM   31  C   LIG     1       3.113  27.482  20.690  1.00  0.00           C
HETATM   32  C   LIG     1       2.914  27.880  22.170  1.00  0.00           C
HETATM   33  C   LIG     1       4.613  27.829  24.043  1.00  0.00           C
HETATM   34  C   LIG     1       3.499  28.379  24.935  1.00  0.00           C
HETATM   35  C   LIG     1       5.367  26.686  24.733  1.00  0.00           C
HETATM   36  C   LIG     1       6.742  26.532  24.101  1.00  0.00           C
HETATM   37  C   LIG     1       7.825  26.509  25.182  1.00  0.00           C
HETATM   38  C   LIG     1       7.050  27.620  23.068  1.00  0.00           C
HETATM   39  C   LIG     1       8.285  27.226  22.246  1.00  0.00           C
HETATM   40  C   LIG     1       8.051  25.907  21.524  1.00  0.00           C
HETATM   41  C   LIG     1      10.724  24.002  19.501  1.00  0.00           C
HETATM   42  C   LIG     1      11.332  25.186  18.795  1.00  0.00           C
HETATM   43  N   LIG     1      12.306  24.975  17.923  1.00  0.00           N
HETATM   44  C   LIG     1      12.853  25.984  17.263  1.00  0.00           C
HETATM   45  C   LIG     1      13.977  25.690  16.303  1.00  0.00           C
HETATM   46  C   LIG     1      14.179  26.830  15.305  1.00  0.00           C
HETATM   47  C   LIG     1      13.027  26.899  14.301  1.00  0.00           C
HETATM   48  C   LIG     1      15.514  26.633  14.618  1.00  0.00           C
HETATM   49  C   LIG     1      16.207  25.519  14.837  1.00  0.00           C
HETATM   50  C   LIG     1      17.510  25.303  14.201  1.00  0.00           C
HETATM   51  O   LIG     1      17.918  24.182  13.983  1.00  0.00           O
```

```
HETATM   52  C   LIG     1      18.330  26.510  13.822  1.00  0.00           C
HETATM   53  C   LIG     1      18.955  27.181  15.043  1.00  0.00           C
HETATM   54  O   LIG     1      19.732  28.316  14.641  1.00  0.00           O
HETATM   55  C   LIG     1      20.765  28.006  13.705  1.00  0.00           C
HETATM   56  O   LIG     1      21.731  27.122  14.297  1.00  0.00           O
HETATM   57  C   LIG     1      22.456  27.969  15.214  1.00  0.00           C
HETATM   58  C   LIG     1      21.619  28.249  16.463  1.00  0.00           C
HETATM   59  C   LIG     1      23.790  27.324  15.597  1.00  0.00           C
HETATM   60  C   LIG     1      22.689  29.264  14.420  1.00  0.00           C
HETATM   61  C   LIG     1      21.570  29.265  13.376  1.00  0.00           C
HETATM   62  O   LIG     1       1.633  27.466  22.648  1.00  0.00           O
HETATM   63  H   LIG     1       1.670  26.886  23.421  1.00  0.00           H
HETATM   64  C   LIG     1       5.138  26.670  19.414  1.00  0.00           C
HETATM   65  O   LIG     1       5.347  27.544  18.303  1.00  0.00           O
HETATM   66  H   LIG     1       5.550  28.456  18.553  1.00  0.00           H
HETATM   67  O   LIG     1       5.425  26.839  26.153  1.00  0.00           O
HETATM   68  H   LIG     1       6.298  26.668  26.531  1.00  0.00           H
HETATM   69  O   LIG     1       7.303  28.855  23.741  1.00  0.00           O
HETATM   70  H   LIG     1       7.771  29.507  23.202  1.00  0.00           H
HETATM   71  O   LIG     1      15.060  27.911  12.665  1.00  0.00           O
HETATM   72  H   LIG     1      15.450  28.145  11.812  1.00  0.00           H
HETATM   73  O   LIG     1      22.113  29.185  12.057  1.00  0.00           O
HETATM   74  H   LIG     1      22.476  30.019  11.730  1.00  0.00           H
```

10_apo_SANC00518.pdb

```
HETATM    1  C   LIG     1       6.271  25.305  17.265  1.00  0.00           C
HETATM    2  C   LIG     1       6.469  26.800  17.495  1.00  0.00           C
HETATM    3  C   LIG     1       5.815  27.161  18.838  1.00  0.00           C
HETATM    4  C   LIG     1       6.494  26.395  19.973  1.00  0.00           C
HETATM    5  C   LIG     1       7.918  26.952  20.127  1.00  0.00           C
HETATM    6  C   LIG     1       6.502  24.902  19.696  1.00  0.00           C
HETATM    7  C   LIG     1       7.204  24.142  20.816  1.00  0.00           C
HETATM    8  C   LIG     1       6.483  24.416  22.144  1.00  0.00           C
HETATM    9  C   LIG     1       6.478  25.918  22.432  1.00  0.00           C
HETATM   10  C   LIG     1       5.792  26.233  23.735  1.00  0.00           C
HETATM   11  C   LIG     1       6.315  25.623  25.030  1.00  0.00           C
HETATM   12  C   LIG     1       6.017  26.678  26.130  1.00  0.00           C
HETATM   13  C   LIG     1       5.233  27.814  25.416  1.00  0.00           C
HETATM   14  C   LIG     1       5.826  27.757  23.996  1.00  0.00           C
HETATM   15  C   LIG     1       7.246  28.329  24.025  1.00  0.00           C
HETATM   16  C   LIG     1       5.014  28.457  22.927  1.00  0.00           C
HETATM   17  C   LIG     1       5.645  28.153  21.560  1.00  0.00           C
HETATM   18  C   LIG     1       5.751  26.657  21.296  1.00  0.00           C
HETATM   19  C   LIG     1       4.317  26.125  21.175  1.00  0.00           C
HETATM   20  C   LIG     1       7.029  24.503  18.316  1.00  0.00           C
HETATM   21  C   LIG     1       8.539  24.664  18.226  1.00  0.00           C
HETATM   22  C   LIG     1       6.695  23.014  18.101  1.00  0.00           C
HETATM   23  C   LIG     1       7.131  22.592  16.701  1.00  0.00           C
HETATM   24  C   LIG     1       6.365  23.374  15.643  1.00  0.00           C
HETATM   25  C   LIG     1       6.489  24.887  15.830  1.00  0.00           C
HETATM   26  C   LIG     1       5.369  25.521  14.967  1.00  0.00           C
HETATM   27  C   LIG     1       7.819  25.377  15.261  1.00  0.00           C
HETATM   28  O   LIG     1       6.858  23.023  14.345  1.00  0.00           O
HETATM   29  C   LIG     1       6.105  22.178  13.611  1.00  0.00           C
HETATM   30  O   LIG     1       6.014  22.334  12.408  1.00  0.00           O
HETATM   31  C   LIG     1       5.419  21.108  14.241  1.00  0.00           C
HETATM   32  C   LIG     1       6.052  19.935  14.456  1.00  0.00           C
HETATM   33  C   LIG     1       5.342  18.826  15.110  1.00  0.00           C
HETATM   34  C   LIG     1       6.053  17.713  15.575  1.00  0.00           C
HETATM   35  C   LIG     1       5.380  16.677  16.190  1.00  0.00           C
```

```
HETATM   36  C   LIG     1       3.995  16.736  16.341  1.00  0.00           C
HETATM   37  C   LIG     1       3.289  17.838  15.877  1.00  0.00           C
HETATM   38  C   LIG     1       3.951  18.878  15.264  1.00  0.00           C
HETATM   39  O   LIG     1       6.068  15.596  16.647  1.00  0.00           O
HETATM   40  H   LIG     1       5.645  15.147  17.392  1.00  0.00           H
HETATM   41  O   LIG     1       3.335  15.713  16.945  1.00  0.00           O
HETATM   42  C   LIG     1       2.381  15.001  16.155  1.00  0.00           C
HETATM   43  C   LIG     1       7.782  25.307  24.995  1.00  0.00           C
HETATM   44  C   LIG     1       8.658  26.227  25.313  1.00  0.00           C
HETATM   45  C   LIG     1       8.247  23.930  24.595  1.00  0.00           C
```

8_apo_SANC00585.pdb

```
HETATM    1  C   LIG     1       7.962  23.426  16.641  1.00  0.00           C
HETATM    2  C   LIG     1       7.499  22.690  17.874  1.00  0.00           C
HETATM    3  C   LIG     1       6.843  21.499  17.743  1.00  0.00           C
HETATM    4  C   LIG     1       6.442  20.764  18.867  1.00  0.00           C
HETATM    5  C   LIG     1       5.939  19.390  18.766  1.00  0.00           C
HETATM    6  C   LIG     1       6.069  18.605  17.606  1.00  0.00           C
HETATM    7  C   LIG     1       6.940  19.082  16.467  1.00  0.00           C
HETATM    8  C   LIG     1       7.404  17.889  15.629  1.00  0.00           C
HETATM    9  C   LIG     1       8.188  18.390  14.412  1.00  0.00           C
HETATM   10  O   LIG     1       6.230  17.202  15.189  1.00  0.00           O
HETATM   11  C   LIG     1       5.474  16.587  16.229  1.00  0.00           C
HETATM   12  C   LIG     1       4.041  16.410  15.707  1.00  0.00           C
HETATM   13  C   LIG     1       5.446  17.391  17.500  1.00  0.00           C
HETATM   14  C   LIG     1       4.776  16.856  18.614  1.00  0.00           C
HETATM   15  C   LIG     1       4.711  17.566  19.813  1.00  0.00           C
HETATM   16  C   LIG     1       4.045  17.020  21.004  1.00  0.00           C
HETATM   17  O   LIG     1       3.702  15.857  21.058  1.00  0.00           O
HETATM   18  C   LIG     1       3.805  17.882  22.108  1.00  0.00           C
HETATM   19  C   LIG     1       2.967  17.395  23.262  1.00  0.00           C
HETATM   20  C   LIG     1       4.341  19.120  22.128  1.00  0.00           C
HETATM   21  C   LIG     1       5.214  19.581  21.116  1.00  0.00           C
HETATM   22  C   LIG     1       5.989  20.746  21.286  1.00  0.00           C
HETATM   23  C   LIG     1       6.247  21.280  22.580  1.00  0.00           C
HETATM   24  C   LIG     1       5.868  20.484  23.803  1.00  0.00           C
HETATM   25  C   LIG     1       6.777  20.872  24.974  1.00  0.00           C
HETATM   26  C   LIG     1       6.304  20.137  26.236  1.00  0.00           C
HETATM   27  O   LIG     1       6.651  22.280  25.176  1.00  0.00           O
HETATM   28  C   LIG     1       7.072  23.098  24.098  1.00  0.00           C
HETATM   29  C   LIG     1       6.270  24.409  24.142  1.00  0.00           C
HETATM   30  C   LIG     1       6.891  22.461  22.748  1.00  0.00           C
HETATM   31  C   LIG     1       7.468  23.124  21.623  1.00  0.00           C
HETATM   32  O   LIG     1       8.129  24.143  21.767  1.00  0.00           O
HETATM   33  C   LIG     1       7.267  22.578  20.280  1.00  0.00           C
HETATM   34  C   LIG     1       7.749  23.226  19.139  1.00  0.00           C
HETATM   35  C   LIG     1       6.566  21.358  20.145  1.00  0.00           C
HETATM   36  C   LIG     1       5.300  18.846  19.905  1.00  0.00           C
HETATM   37  O   LIG     1       4.197  15.637  18.534  1.00  0.00           O
HETATM   38  H   LIG     1       3.234  15.664  18.446  1.00  0.00           H
HETATM   39  O   LIG     1       8.479  24.366  19.248  1.00  0.00           O
HETATM   40  H   LIG     1       9.424  24.246  19.079  1.00  0.00           H
```

5_apo_SANC00594.pdb

```
HETATM    1  C   LIG     1       6.659  20.977  21.977  1.00  0.00           C
HETATM    2  C   LIG     1       6.569  20.886  23.323  1.00  0.00           C
HETATM    3  C   LIG     1       6.488  22.100  24.212  1.00  0.00           C
HETATM    4  C   LIG     1       6.555  19.499  23.861  1.00  0.00           C
HETATM    5  O   LIG     1       7.173  19.180  24.860  1.00  0.00           O
HETATM    6  C   LIG     1       5.728  18.528  23.087  1.00  0.00           C
```

```
HETATM    7  C   LIG     1       5.823  18.656  21.738  1.00  0.00           C
HETATM    8  C   LIG     1       4.940  17.801  20.879  1.00  0.00           C
HETATM    9  C   LIG     1       5.142  16.332  21.253  1.00  0.00           C
HETATM   10  C   LIG     1       3.479  18.186  21.120  1.00  0.00           C
HETATM   11  C   LIG     1       5.267  17.991  19.427  1.00  0.00           C
HETATM   12  O   LIG     1       5.147  17.066  18.660  1.00  0.00           O
HETATM   13  C   LIG     1       5.741  19.321  18.919  1.00  0.00           C
HETATM   14  C   LIG     1       6.902  19.842  19.740  1.00  0.00           C
HETATM   15  C   LIG     1       6.753  19.685  21.217  1.00  0.00           C
HETATM   16  C   LIG     1       8.115  19.142  21.651  1.00  0.00           C
HETATM   17  C   LIG     1       6.661  22.309  21.273  1.00  0.00           C
HETATM   18  C   LIG     1       7.883  22.401  20.357  1.00  0.00           C
HETATM   19  C   LIG     1       8.520  23.785  20.494  1.00  0.00           C
HETATM   20  C   LIG     1       8.568  24.165  21.969  1.00  0.00           C
HETATM   21  C   LIG     1       9.162  25.576  22.070  1.00  0.00           C
HETATM   22  C   LIG     1       9.452  23.175  22.731  1.00  0.00           C
HETATM   23  C   LIG     1      10.598  25.511  21.565  1.00  0.00           C
HETATM   24  C   LIG     1      10.666  25.083  20.097  1.00  0.00           C
HETATM   25  C   LIG     1       9.902  23.762  19.865  1.00  0.00           C
HETATM   26  O   LIG     1      12.046  24.837  19.825  1.00  0.00           O
HETATM   27  C   LIG     1      10.723  22.588  20.401  1.00  0.00           C
HETATM   28  C   LIG     1       9.656  23.633  18.348  1.00  0.00           C
HETATM   29  C   LIG     1      10.702  22.760  17.681  1.00  0.00           C
HETATM   30  C   LIG     1      12.075  23.057  18.242  1.00  0.00           C
HETATM   31  O   LIG     1      12.862  22.167  18.455  1.00  0.00           O
HETATM   32  C   LIG     1      12.436  24.487  18.526  1.00  0.00           C
HETATM   33  C   LIG     1      13.952  24.656  18.406  1.00  0.00           C
HETATM   34  C   LIG     1      11.750  25.391  17.499  1.00  0.00           C
HETATM   35  O   LIG     1       4.941  17.594  23.683  1.00  0.00           O
HETATM   36  H   LIG     1       4.749  16.827  23.124  1.00  0.00           H
```

12_apo_SANC00595.pdb

```
HETATM    1  C   LIG     1       4.881  22.100  13.421  1.00  0.00           C
HETATM    2  C   LIG     1       5.511  21.068  12.472  1.00  0.00           C
HETATM    3  C   LIG     1       6.111  19.882  13.232  1.00  0.00           C
HETATM    4  C   LIG     1       7.131  20.354  14.275  1.00  0.00           C
HETATM    5  C   LIG     1       7.714  19.137  14.997  1.00  0.00           C
HETATM    6  C   LIG     1       6.473  21.283  15.304  1.00  0.00           C
HETATM    7  C   LIG     1       7.509  21.795  16.305  1.00  0.00           C
HETATM    8  C   LIG     1       7.043  21.220  17.651  1.00  0.00           C
HETATM    9  C   LIG     1       6.279  19.964  17.197  1.00  0.00           C
HETATM   10  O   LIG     1       5.551  20.538  16.101  1.00  0.00           O
HETATM   11  C   LIG     1       5.731  22.448  14.662  1.00  0.00           C
HETATM   12  C   LIG     1       6.736  23.527  14.256  1.00  0.00           C
HETATM   13  C   LIG     1       4.719  23.028  15.658  1.00  0.00           C
HETATM   14  C   LIG     1       4.420  24.520  15.676  1.00  0.00           C
HETATM   15  C   LIG     1       4.342  25.026  14.252  1.00  0.00           C
HETATM   16  O   LIG     1       4.814  26.098  13.956  1.00  0.00           O
HETATM   17  C   LIG     1       3.669  24.176  13.208  1.00  0.00           C
HETATM   18  C   LIG     1       2.447  23.518  13.854  1.00  0.00           C
HETATM   19  C   LIG     1       3.166  25.103  12.099  1.00  0.00           C
HETATM   20  O   LIG     1       4.481  23.191  12.591  1.00  0.00           O
HETATM   21  C   LIG     1       5.343  19.453  18.287  1.00  0.00           C
HETATM   22  C   LIG     1       4.808  18.030  17.964  1.00  0.00           C
HETATM   23  C   LIG     1       3.459  18.274  17.378  1.00  0.00           C
HETATM   24  C   LIG     1       5.696  17.317  16.976  1.00  0.00           C
HETATM   25  C   LIG     1       4.721  17.264  19.287  1.00  0.00           C
HETATM   26  C   LIG     1       2.898  19.359  18.364  1.00  0.00           C
HETATM   27  O   LIG     1       2.472  17.282  17.209  1.00  0.00           O
```

HETATM 28 C LIG 1 4.087 20.360 18.362 1.00 0.00 C
HETATM 29 C LIG 1 2.808 15.944 16.791 1.00 0.00 C
HETATM 30 C LIG 1 2.719 14.988 17.982 1.00 0.00 C
HETATM 31 C LIG 1 1.701 15.501 15.806 1.00 0.00 C
HETATM 32 C LIG 1 4.085 15.734 16.091 1.00 0.00 C
HETATM 33 O LIG 1 4.077 15.445 14.912 1.00 0.00 O
HETATM 34 C LIG 1 5.426 15.877 16.825 1.00 0.00 C
HETATM 35 C LIG 1 6.056 19.437 19.640 1.00 0.00 C

42_apo_SANC00685.pdb

HETATM 1 C LIG 1 4.802 16.758 17.289 1.00 0.00 C
HETATM 2 C LIG 1 4.768 15.799 16.109 1.00 0.00 C
HETATM 3 C LIG 1 3.336 15.380 15.779 1.00 0.00 C
HETATM 4 C LIG 1 5.458 16.376 14.884 1.00 0.00 C
HETATM 5 O LIG 1 5.283 17.753 14.682 1.00 0.00 O
HETATM 6 C LIG 1 5.112 18.689 15.715 1.00 0.00 C
HETATM 7 O LIG 1 4.381 19.825 15.184 1.00 0.00 O
HETATM 8 C LIG 1 5.376 20.643 14.557 1.00 0.00 C
HETATM 9 C LIG 1 4.871 22.061 14.483 1.00 0.00 C
HETATM 10 O LIG 1 4.061 22.496 13.699 1.00 0.00 O
HETATM 11 C LIG 1 5.571 22.846 15.568 1.00 0.00 C
HETATM 12 C LIG 1 4.796 23.894 16.353 1.00 0.00 C
HETATM 13 C LIG 1 4.368 25.034 15.410 1.00 0.00 C
HETATM 14 C LIG 1 3.685 26.095 16.221 1.00 0.00 C
HETATM 15 C LIG 1 3.911 26.292 17.489 1.00 0.00 C
HETATM 16 C LIG 1 3.135 27.378 18.206 1.00 0.00 C
HETATM 17 C LIG 1 4.152 28.276 18.915 1.00 0.00 C
HETATM 18 C LIG 1 4.999 27.428 19.861 1.00 0.00 C
HETATM 19 C LIG 1 5.814 26.397 19.081 1.00 0.00 C
HETATM 20 C LIG 1 4.906 25.471 18.276 1.00 0.00 C
HETATM 21 C LIG 1 4.102 24.619 19.257 1.00 0.00 C
HETATM 22 C LIG 1 5.727 24.541 17.381 1.00 0.00 C
HETATM 23 C LIG 1 6.597 23.578 18.165 1.00 0.00 C
HETATM 24 C LIG 1 7.226 22.469 17.320 1.00 0.00 C
HETATM 25 C LIG 1 6.101 21.783 16.536 1.00 0.00 C
HETATM 26 C LIG 1 5.005 21.341 17.499 1.00 0.00 C
HETATM 27 C LIG 1 6.518 20.671 15.578 1.00 0.00 C
HETATM 28 C LIG 1 6.436 19.280 16.209 1.00 0.00 C
HETATM 29 C LIG 1 7.627 18.439 15.718 1.00 0.00 C
HETATM 30 C LIG 1 4.311 18.142 16.892 1.00 0.00 C
HETATM 31 O LIG 1 6.584 25.630 20.016 1.00 0.00 O
HETATM 32 C LIG 1 7.770 26.286 20.469 1.00 0.00 C
HETATM 33 C LIG 1 8.003 25.956 21.945 1.00 0.00 C
HETATM 34 C LIG 1 9.307 26.614 22.406 1.00 0.00 C
HETATM 35 C LIG 1 10.455 26.130 21.514 1.00 0.00 C
HETATM 36 C LIG 1 10.125 26.453 20.053 1.00 0.00 C
HETATM 37 O LIG 1 8.886 25.837 19.698 1.00 0.00 O
HETATM 38 O LIG 1 6.915 26.455 22.726 1.00 0.00 O
HETATM 39 C LIG 1 6.491 25.570 23.764 1.00 0.00 C
HETATM 40 C LIG 1 6.211 26.372 25.036 1.00 0.00 C
HETATM 41 C LIG 1 5.068 27.357 24.770 1.00 0.00 C
HETATM 42 C LIG 1 3.839 26.579 24.291 1.00 0.00 C
HETATM 43 C LIG 1 4.210 25.768 23.047 1.00 0.00 C
HETATM 44 O LIG 1 5.299 24.895 23.356 1.00 0.00 O
HETATM 45 C LIG 1 3.004 24.941 22.599 1.00 0.00 C
HETATM 46 O LIG 1 5.838 25.483 26.092 1.00 0.00 O
HETATM 47 H LIG 1 4.882 25.383 26.203 1.00 0.00 H
HETATM 48 O LIG 1 4.751 28.062 25.972 1.00 0.00 O
HETATM 49 H LIG 1 5.523 28.404 26.443 1.00 0.00 H
HETATM 50 O LIG 1 2.789 27.493 23.971 1.00 0.00 O

```
HETATM   51  H   LIG     1       2.164  27.158  23.313  1.00  0.00          H
HETATM   52  O   LIG     1       9.573  26.253  23.764  1.00  0.00          O
HETATM   53  H   LIG     1       9.152  25.427  24.040  1.00  0.00          H
HETATM   54  O   LIG     1      10.617  24.718  21.666  1.00  0.00          O
HETATM   55  H   LIG     1      10.227  24.197  20.950  1.00  0.00          H
HETATM   56  O   LIG     1       3.455  29.272  19.667  1.00  0.00          O
HETATM   57  H   LIG     1       2.578  28.996  19.967  1.00  0.00          H
HETATM   58  O   LIG     1       2.921  18.122  16.567  1.00  0.00          O
HETATM   59  H   LIG     1       2.358  17.765  17.267  1.00  0.00          H
HETATM   60  O   LIG     1       3.977  16.249  18.346  1.00  0.00          O
HETATM   61  C   LIG     1       3.606  17.228  19.319  1.00  0.00          C
HETATM   62  C   LIG     1       2.845  16.551  20.462  1.00  0.00          C
HETATM   63  C   LIG     1       2.504  17.597  21.527  1.00  0.00          C
HETATM   64  C   LIG     1       3.797  18.263  22.005  1.00  0.00          C
HETATM   65  C   LIG     1       4.521  18.875  20.803  1.00  0.00          C
HETATM   66  O   LIG     1       4.781  17.854  19.836  1.00  0.00          O
HETATM   67  C   LIG     1       5.844  19.491  21.264  1.00  0.00          C
HETATM   68  O   LIG     1       1.640  15.972  19.957  1.00  0.00          O
HETATM   69  H   LIG     1       1.497  15.057  20.234  1.00  0.00          H
HETATM   70  O   LIG     1       1.634  18.584  20.970  1.00  0.00          O
HETATM   71  H   LIG     1       1.418  18.439  20.039  1.00  0.00          H
HETATM   72  O   LIG     1       4.638  17.288  22.624  1.00  0.00          O
HETATM   73  H   LIG     1       4.956  16.603  22.020  1.00  0.00          H
```

4. Main scripts used

4.1. Data organization and filtering

```
#!/usr/bin/python3

#The Patient class

import os, subprocess

from glob import glob


class Patient:

    def __init__(self):

        self.patientID = None

        self.yearTrtAaSeq = {}

        self.country = None

        self.alignment = []

        self.mutationReport = []

        self.reference = None

        self.seqduplicates = set()


filePath = 'datasets/PR.txt'
```

```
patients = {}
consensusProtease = [i for i in
'PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMNLPGRWKPKMIGGIGGFIKVRQYDQILIE
ICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF']
outPath = 'output/'


with open(filePath, 'r') as handle:
    handle.readline()
    patientSeen = []
    records = handle.readlines()
    for record in records:
        record = record.split('\t')
        reference = record[0]
        patientID =  record[1]
        treatment = record[6]
        year = record[4].replace(' ', '')
        aaSeq = record[7:-2]
        country = record[3]
        subtype = record[5]
        ntSeq = record[-1].rstrip()

        if country == 'South Africa' and subtype == 'C' and\
        '#' not in aaSeq and '.' not in aaSeq and len(''.join(aaSeq)) == 99:
            if patientID not in patientSeen:
                patients[patientID] = Patient()
                patients[patientID].patientID = patientID
                patients[patientID].yearTrtAaSeq[treatment] = [year, aaSeq, ntSeq]
                patients[patientID].country = country
                patients[patientID].reference = reference
                patientSeen.append(patientID)
```

```
            else:
                #Overwrite previous entry if same treatment is used
                patients[patientID].yearTrtAaSeq[treatment] = [year, aaSeq, ntSeq]

def filterSingletons():
    output = {}
    for patientID in patients:
        if len(patients[patientID].yearTrtAaSeq) > 1:
            output[patientID] = patients[patientID]
    return output

#Filter out patients with only one sequence/ treatment
patients = filterSingletons()

def displayMutations(seq1, seq2):
    #Shows the mutation locations with a '*' character
    output = [' ' for i in range(99)]
    for s1, s2, idx in zip(seq1, seq2, range(99)):
        if s1 != s2:
            output[idx] = '*'
    output = ''.join(output)
    return output

def convSeq(seq, consensus):
    #Replaces '-' by their respective residues from the consensus
    output = seq
    for seq_aa, cons_aa, idx in zip(seq, consensus, range(99)):
        if seq_aa == '-':
            output[idx] = cons_aa
```

```
    output = ''.join(output)
    return output

def blastp(query, output):
    #BLASTP of query and writes output
    subprocess.check_output(['blastp', '-query', query, '-db', 'blastdb/pdbaa',
        '-out', output, '-evalue', '0.00001', '-matrix', 'BLOSUM62', '-outfmt', '7'])

def toFasta(patientID, treatment, nt=False):
    #Creates a directories for every patient/ treatment
    #Write fasta alignment (aa seq by default) and the individual files for each patient.
    #Note: no alignment program was used as the seqs were of same lengths.
    #Visual inspection (and a mock muscle alignment) was done to verify the assumption
    year, aa_sequence, nt_sequence = patients[patientID].yearTrtAaSeq[treatment]

    #The fasta sequence header
    header = '>'+patientID+'_'+treatment+'_'+year+'\n'

    if nt:
        #Take the coding nt sequence and write the alignment file
        with open(outPath+patientID+'/'+patientID+'_aligned_nt.fas', 'a') as nt_handle:
            nt_handle.write(header+nt_sequence+'\n')

    aa_sequence = convSeq(aa_sequence, consensusProtease)
    aa_sequence += '\n'
    tmptreatment = treatment.replace('<', '').replace('=', '').replace('>', '')

    #Create directory if it doesn't exist already: patient directory
    if not os.path.exists(outPath+patientID):
```

```
        os.mkdir(outPath+patientID)

    #Treatment directory
    if not os.path.exists(outPath+patientID+'/'+patientID+'_'+tmptreatment+'/'):
        os.mkdir(outPath+patientID+'/'+patientID+'_'+tmptreatment+'/')

    #Directories for models of open and closed conformations
    if not os.path.exists(outPath+patientID+'/'+patientID+'_'+tmptreatment+'/open/'):
        os.mkdir(outPath+patientID+'/'+patientID+'_'+tmptreatment+'/open')
        os.mkdir(outPath+patientID+'/'+patientID+'_'+tmptreatment+'/closed')

    #Writing the individual file
    individualFile = outPath+patientID+'/'+ patientID+'_'+tmptreatment+'/'
+patientID+'_'+tmptreatment+'.fas'

    with open(individualFile, 'w') as tmphandle:
        tmphandle.write(header+aa_sequence)

    #Look for homologs (using protein sequence) and write output
    blastp(individualFile, individualFile[:-4]+'_blast.txt')

    #Writing the alignment file (for the drug-naive and LPV treatment)
    with open(outPath+patientID+'/'+patientID+'_aligned.fas', 'a') as handle:
        handle.write(header+aa_sequence)

def filterNoEffect():
    #Filter out sequences that don't change after treatment
    output = {}
    for patientID in patients.keys():
        naiveSeq = ''.join(patients[patientID].yearTrtAaSeq['None'][1])
```

```
        LPVSeq = ''.join(patients[patientID].yearTrtAaSeq['LPV'][1])
        if naiveSeq != LPVSeq:
            output[patientID] = patients[patientID]

    return output
patients = filterNoEffect()

def flagDuplicates():
    #Flag patient sequences that are same
    for patient1 in patients.keys():
        for trt1 in patients[patient1].yearTrtAaSeq.keys():
            sequence1 = patients[patient1].yearTrtAaSeq[trt1][1]
            for patient2 in patients.keys():
                for trt2 in patients[patient2].yearTrtAaSeq.keys():
                    sequence2 = patients[patient2].yearTrtAaSeq[trt2][1]
                    if sequence1 == sequence2 and patient1 != patient2:
                        patients[patient1].seqduplicates.add((patient2, trt2))
                        patients[patient2].seqduplicates.add((patient1, trt1))

flagDuplicates()

#Displaying the duplicate sequences
print('Duplicates:')
for patientID in patients:
    if len(patients[patientID].seqduplicates) > 0:
        duplicate = ' '.join(['_'.join(list(i)) for i in patients[patientID].seqduplicates])
        print(patientID, duplicate)
print()

#Clean previous alignment files
```

```
files = glob('output/*/*aligned*.fas')
for i in files:
    os.remove(i)


#summarize display
for patientID in patients:
    tmpSeq = []
    print('>'+patientID, 'refID:', patients[patientID].reference)
    for i in sorted(patients[patientID].yearTrtAaSeq.items(), reverse=True):
        sequence = i[1][1]
        sequence = convSeq(sequence, consensusProtease)
        year = i[1][0]
        treatment = i[0]
        tmpSeq.append(sequence)
        print(' {:<5}{:<7}{} {}'.format(treatment, year, sequence, len(sequence)))
        #Write alignment
        toFasta(patientID, treatment, nt=True)
    print('{}{}'.format(' '*13, displayMutations(tmpSeq[0], tmpSeq[1])))
    print()
```

4.2. Modeling (model-default.py)

```
#!/usr/bin/python3
# Comparative modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class
import os, sys
from glob import glob
from Bio.PDB import *
from Bio import SeqIO
```

```
'''
Usage: model-default.py [-open|closed]
'''
log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

def getTargetSequence(target):
    #Extracts the sequence string and id
    seq = SeqIO.read(target ,'fasta')
    return [seq.id, str(seq.seq)]

#Target sequence
for target in glob('output/*/*_*/*_*.fas'):
    path = '/'.join(target.split('/')[:-1])+'/'
    path_open = path + 'open/'
    path_closed = path + 'closed/'
    cur_target = target.split('/')[-1]

    if '-open' in sys.argv:
        template = '3BC4'
        os.chdir(path_open)
        # directories for input atom files
        env.io.atom_files_directory = ['../../../../templates/openconf/']
        seqid, seq = getTargetSequence('../'+cur_target)

    elif '-closed' in sys.argv:
        template = '3PWM'
        os.chdir(path_closed)
        env.io.atom_files_directory = ['../../../../templates/closedconf/']
        seqid, seq = getTargetSequence('../'+cur_target)
```

```
a = automodel(env,
              alnfile  = 'alignment.ali',     # alignment filename
              knowns   = template, # codes of the templates
              sequence = seqid)            # code of the target
a.starting_model= 41                 # index of the first model
a.ending_model  = 100                # index of the last model
                                # (determines how many models to calculate)
a.make()                        # do the actual comparative modeling
os.chdir('../../../../')
```

4.3. z-DOPE scoring (adapted from built-in MODELLER 9.14 example files: “assess_dope.py” & “assess_ga341.py”)

```
#!/usr/bin/python3
# Modified from MODELLER example file “model.assess_normalized_dope() ”
import os, sys
from glob import glob
from modeller import *
from modeller.scripts import complete_pdb

'''
Computes z-dope & ga341 for all pdb files in current directory
GA341 was dropped due to its inability to differentiate models (values)
'''

env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')

def assess_zdope(path):
    #Path is a directory containing pdb models
    #Delete previous entry if it exists
    if os.path.exists(path+'normalized_dope.txt'):
```

```
        os.remove(path+'normalized_dope.txt')

    with open(path+'normalized_dope.txt', 'a') as norm_dope_handle:

        for struct in glob(path+'*.pdb'):
            # Read a model previously generated by Modeller's automodel class
            mdl = complete_pdb(env, struct)

            zscore = mdl.assess_normalized_dope()
            zscore = str(zscore)
            norm_dope_handle.writelines(struct+'\t'+zscore+'\n')

def assess_ga341(path):
    #Path is a directory containing pdb models
    #Delete previous entry if it exists
    if os.path.exists(path+'ga341_out.txt'):
        os.remove(path+'ga341_out.txt')

    with open(path+'ga341_out.txt', 'a') as ga_handle:

        for struct in glob(path+'*.pdb'):
            # Read a model previously generated by Modeller's automodel class
            mdl = complete_pdb(env, struct)

            # Set template-model sequence identity. (Not needed in this case, since
            # this is written by Modeller into the .pdb file.)

            score = mdl.assess_ga341()
            print('.....',score)
            ga_score = str(score[0])
            ga_handle.writelines(struct+'\t'+ga_score+'\n')

if __name__ == '__main__':
    if '-curdir' in sys.argv:
```

```
    #Option to use any structure in current directory
    path = './'
  else:
    path = 'output/*/*_*/*/'
  for dirpath in glob(path):
    print(dirpath)
    assess_zdope(dirpath)
    assess_ga341(dirpath)
```

4.4. z-DOPE plots (zdope_ranking.R)

```
rm(list=ls())
setwd('/home/olivier/git/project/output/')
zdope_path = list.files(pattern='normalized_dope\\.txt', recursive=T)[-c(27,28)] #excluding duplicate (in open(27) & closed(28) conf)
ga341_path = list.files(pattern='ga341_out\\.txt', recursive=T)[-c(-c(27,28))]


minima = data.frame()
color <- c()
counter = 1
zdope_dat <- data.frame()
for (path in zdope_path){
  tmp <- read.table(path ,header=F, sep='\t')
  filename <- tmp[1,1]
  min_score_index <- which.min(tmp[,2])
  minima <- rbind(minima ,tmp[min_score_index,])

  #Build the factor for the current model
  fac <- strsplit(as.character(filename), split='\\.')[[1]][1]
  fac <- regmatches(x=fac, m=regexpr(pattern='([[:alpha:]]+/[[:digit:]]+_[[:alpha:]]+)_.[[:digit:]]+', text = fac, perl = T))
  state <- strsplit(fac, split='/')[[1]][1]
  color[counter] <- state
  new_fac <- rep(fac, 100)
  tmp <- cbind(new_fac, tmp)
```

```
  zdope_dat <- rbind(zdope_dat, tmp)
  counter = counter + 1
}

color <- factor(color, levels=c('open', 'closed'), labels=c(1,2))
#Show distributions of scores
zdope_dat[,1] <- factor(zdope_dat[,1])
boxplot(zdope_dat$V2 ~ zdope_dat$new_fac, pch=19,
        cex=0.3, cex.axis=0.8, col=color,
        main='Distribution of z-DOPE scores for modeled HIV-1 proteases',
        ylab='z-DOPE score', xlab='Modeled patient proteases', names=1:42)
legend('bottomleft', legend=c('open conformation', 'closed conformation'),
       col=c(1,2), pch=15, cex=0.7)
abline(h=-1, lty=3)

minima <- cbind(1:42, minima)
write.table(x=minima, file='bestmodels.txt', sep='\t', quote=F, row.names=F, col.names=F)
```

4.5. Protein centering (center_receptors.py)

```
#!/usr/bin/python
import subprocess
import os, time
from glob import glob

os.chdir('receptors/')
#reference is the pdb file from which all other structures will center onto
reference = '1.pdb'

#for i in range(2,45):
for i in range(2,43):
    with open('tmp_run'+str(i)+'.pml', 'w') as handle:
        i = str(i)+'.pdb'
        handle.writelines('load '+reference+'\n')
        handle.writelines('load '+i+'\n')
```

```
        handle.writelines('align '+i[:-4]+', '+reference[:-4]+'\n')
#       if i[:-4] == '2':
#           handle.writelines('save '+reference[:-4]+'_centered.pdb, '+reference[:-4]+'\n')
        handle.writelines('save '+i[:-4]+'_centered.pdb, '+i[:-4]+'\n')
        pymolcmd = 'pymol -qc tmp_run'+i[:-4]+'.pml &'
        k = os.system(pymolcmd)

time.sleep(10)
for idx, centered_receptor in zip(range(2,43), glob('*_centered.pdb')):
    output = centered_receptor.replace('centered', 'apo')+'qt'
    os.system('prepare_receptor4.py -r '+centered_receptor+' -o '+output)
#   os.remove('tmp_run'+str(idx)+'.pml')
#   os.remove(centered_receptor)
```

4.6. Docking (submit_docking.py)

```
#!/usr/local/bin/python
import os, sys, random
from glob import glob
import manageQueuelib as mq
from time import sleep

'''
Creates job and the config files for docking by vina across various queues
Usage:
    submit_docking
'''

#Check that there are always 1000 jobs on my queue
#align proteins to one position in pymol before docking
os.chdir('receptors')
receptors = glob('*.pdbqt')
os.chdir('../')
base = '/home/olivier/project/docking/'
```

```
#Make a vina config file for every ligand
ligandPath = base+'SANCDB_cpds/'
vinaConfigPath = base+'vinaconf/'
receptorPath = base+'receptors/'
outputPath = base+'output/'
logPath = base+'log/'
jobPath = base+'job/'
center_x = str(13.633)
center_y = str(17.2725)
center_z = str(23.4855)
size_x = str(24)
size_y = str(24)
size_z = str(24)
energy_range = str(4)
exhaustiveness = str(4)
cpu = str(4)

for ligand in glob(ligandPath+'*.pdbqt'):
        ligand = ligand.split('/')[-1]
        ligand = ligand[:-6]

        for receptor in receptors:
                #server = random.choice(['priority', 'batch'])
                server = 'priority'

                #Create vina config file
                vinaConfigFilename = vinaConfigPath+receptor[:-6]+'_'+ligand+'.conf'
                vinaConf_handle = open(vinaConfigFilename, 'w')
                jobfilename = jobPath+receptor[:-6]+'_'+ligand+'.job'
                job_handle = open(jobfilename, 'w')

                vinaConf_handle.writelines(
                        'receptor = '+ receptorPath+receptor +'\n'+
                        'ligand = ' + ligandPath + ligand + '.pdbqt' + '\n'+
```

```
        'out = ' + outputPath +receptor[:-6] +'_'+ ligand+ '.pdbqt' +'\n'+
        'log = ' + logPath + receptor[:-6] +'_'+ ligand + '.log' +'\n'+
        'center_x = ' + center_x+'\n'+
        'center_y = ' + center_y+'\n'+
        'center_z = ' + center_z+'\n'+
        'size_x = ' + size_x+'\n'+
        'size_y = ' + size_y+'\n'+
        'size_z = ' + size_z+'\n'+
        'energy_range = ' + energy_range + '\n'+
        'exhaustiveness = ' + exhaustiveness + '\n'+
        'cpu = ' + cpu+'\n'
        )
#print(vinaConfigFilename + ' written.')
job_handle.writelines(
        '#!/bin/csh\n' +
        '#PBS -N ' + 'o'+ligand+'\n' +
        '#PBS -l nodes=1:ppn=4,walltime=99:00:00\n' +
        '#PBS -q ' + server + '\n' +
        '#PBS -m abe\n' +
        '/autodock/vina --config '+vinaConfigFilename+'\n\n'
        )
job_handle.close()
vinaConf_handle.close()

while not mq.queue_ok():
        #Wait for intervals of 3 secs to send next job
        sleep(0.5)

else:
        os.system('qsub '+ jobfilename)
        #print(jobfilename +' submitted.')
```

4.7. Network edge list building (buildEdgelistFromPlip.py)

```
#!/usr/bin/python3
```

```
from glob import glob
import sys, re

'''
Extracts edgelists from PLIP reports
Usage:   buildEdgelistFromPlip.py <-fda|-sancdb> <-hydrogen|-saltbridge>
'''

opt = sys.argv[1]
opt2 = sys.argv[2]

if opt == '-fda':
   report_pattern = 'receptorLigandInteractionsFromPlip_FDA/*.txt'

elif opt == '-sancdb':
   report_pattern = 'receptorLigandInteractionsFromPlip_SANCDB/*.txt'

complex_dict = {}

def extractHydrogenFeatures():
   #Extracts ligand atom / residue hydrogen interactions from PLIP report
   counter = 0
   while not lines[idx+counter].startswith('\n'):
      newidx = idx+counter

      #Only use the lines that contain H bonding information (from PLIP)
      if lines[newidx].startswith('|') and 'RESNR' not in lines[newidx]:
         contents = lines[newidx].split('|')
         contents = map(lambda x: x.strip(), contents)
         contents = list(contents)

         residue_no = contents[1]
         residue = contents[2]
         ligand_coords = contents[-3]
```

```
        ligand_atom_pattern = '{}{}'.format('HETATM\s+(\d+)\s+(\w+)\s+LIG.+',
ligand_coords.replace(", ", '\s+'))
        protein_coords = contents[-2]

        #Extracting protein/ ligand names from filenames
        if opt == '-sancdb':
            protein_ligand_names = filepath.split('/')[-1].split('_interactions_')[0]
            path_to_complex = 'ligand_receptor_complexes_{}/{}.pdb'.format(opt[1:].upper(),
protein_ligand_names)

        elif opt == '-fda':
            protein_ligand_names = filepath.split('/')[-1].split('.all_interactions_')[0]
            path_to_complex = 'ligand_receptor_complexes_{}/{}.all.pdb'.format(opt[1:].upper(),
protein_ligand_names)

        protein = protein_ligand_names.split('_')[0]
        ligand = protein_ligand_names.split('_')[2]

        #Open the complex reference pdb file to obtain atom identities
        with open(path_to_complex, 'r') as filehandle:
            complex_text = filehandle.read()
            no_protein_atoms = len(re.findall('ATOM', complex_text)) + 1 #Add 1 due to TER
            match = re.findall(ligand_atom_pattern, complex_text)

            if match:
                ligand_atom_idx = match[0][0]
                ligand_atom_idx = int(ligand_atom_idx) - no_protein_atoms
                ligand_atom_idx = str(ligand_atom_idx)
                ligand_atom = match[0][1]

                #print(ligand_atom_idx, ligand_atom)
                print('{}{}, {}:{}'.format(residue, residue_no, ligand.replace('00',''),
':'.join([ligand_atom_idx, ligand_atom])))

    counter += 1
```

```
def extractSaltBridgeFeatures():
    #Extracts ligand atom / residue salt bridges
    counter = 0
    while not lines[idx+counter].startswith('\n'):
        newidx = idx+counter

        #Only use the lines that contain information
        if lines[newidx].startswith('|') and 'RESNR' not in lines[newidx]:
            contents = lines[newidx].split('|')
            contents = map(lambda x: x.strip(), contents)
            contents = list(contents)

            residue_no = contents[1]
            residue = contents[2]
            ligand_coords = contents[-3]
            ligand_atom_pattern = '{}{}'.format('HETATM\s+(\d+)\s+(\w+)\s+LIG.+',
ligand_coords.replace(", ", '\s+'))
            protein_coords = contents[-2]

            #Extracting protein/ ligand names from filenames
            if opt == '-sancdb':
                protein_ligand_names = filepath.split('/')[-1].split('_interactions_')[0]
                path_to_complex = 'ligand_receptor_complexes_{}/{}.pdb'.format(opt[1:].upper(),
protein_ligand_names)

            elif opt == '-fda':
                protein_ligand_names = filepath.split('/')[-1].split('.all_interactions_')[0]
                path_to_complex = 'ligand_receptor_complexes_{}/{}.all.pdb'.format(opt[1:].upper(),
protein_ligand_names)

            protein = protein_ligand_names.split('_')[0]
            ligand = protein_ligand_names.split('_')[2]

            #Open the complex reference pdb file to obtain atom identities
```

```
        with open(path_to_complex, 'r') as filehandle:
            complex_text = filehandle.read()
            no_protein_atoms = len(re.findall('ATOM', complex_text)) + 1 #Add 1 due to TER
            match = re.findall(ligand_atom_pattern, complex_text)
            if match:
                #print('{}{}, {}:{}'.format(residue, residue_no, ligand.replace('00',''),
':'.join(match[0])))
                ligand_atom_idx = match[0][0]
                ligand_atom_idx = int(ligand_atom_idx) - no_protein_atoms
                ligand_atom_idx = str(ligand_atom_idx)
                ligand_atom = match[0][1]


                #print(ligand_atom_idx, ligand_atom)
                print('{}{}, {}:{}'.format(residue, residue_no, ligand.replace('00',''),
':'.join([ligand_atom_idx, ligand_atom])))


        counter += 1


for filepath in glob(report_pattern):


    with open(filepath, 'r') as handle:
        lines = handle.readlines()


        for idx, line in enumerate(lines):
            if line.startswith('**'):


                #Tackle hydrogens
                if "Hydrogen Bonds" in line and opt2 == '-hydrogen':
                    extractHydrogenFeatures()


                #Tackle salt bridges
                elif "Salt Bridges" in line and opt2 == '-saltbridge':
                    extractSaltBridgeFeatures()
```

4.8. Network plotting (draw_receptor_ligand_networks.R)

```
setwd('~/git/project/docking/')
library('igraph')

#Function to check if argument is an amino acid
isAA = function(x){
  x = as.character(x)
  x = strsplit(x, split='')
  x = x[[1]]
  return(length(x) < 7)
}

#Function for plot network from edgelist
plotNetwork = function(dat){
  g = graph.edgelist(as.matrix(dat), directed=F)

  adjMat = as_adjacency_matrix(g)
  adjMatGraph = graph.adjacency(adjMat, weighted=T, mode ='undirected')
  edge_widths = E(adjMatGraph)$weight
  nodes = V(adjMatGraph)

  #Checks whether node is aa residue
  sorted_nodes = sapply(names(nodes), isAA)

  #Coloring nodes red (aa residue) and green (ligand atom)
  node_color = ifelse(test=sorted_nodes, yes='red', no='green')

  tkplot(adjMatGraph,
       vertex.size=degree(adjMatGraph),
       vertex.color=node_color,
       vertex.label.cex=1.5,
       edge.width=E(adjMatGraph)$weight)
}
```

4.9. Energy profiling (calcDockingEnergy.py)

#!/usr/bin/python3

import sys, math

import numpy as np

'''

Calculates free energy scores between a docked ligand and a receptor, using the AutoDock 4's potentials

Adapted from (Morris et al., 2007)
"http://autodock.scripps.edu/resources/parameters/AD4_parameters.dat/view"

Usage: calcDockingEnergy <ligand.pdbqt> <receptor.pdbqt> [--score]

- Rii = sum of vdW radii of two like atoms (in Angstrom)

- epsii = vdW well depth (in Kcal/mol)

- vol = atomic solvation volume (in Angstrom^3)

- solpar = atomic solvation parameter

- Rij_hb = H-bond radius of the heteroatom in contact with a hydrogen (in Angstrom)

- epsij_hb = well depth of H-bond (in Kcal/mol)

- hbond = integer indicating type of H-bonding atom (0=no H-bond)

- rec_index = initialised to -1, but later on holds count of how many of this atom type are in receptor

- map_index = initialised to -1, but later on holds the index of the AutoGrid map

- bond_index = used in AutoDock to detect bonds; see "mdist.h", enum {C,N,O,H,XX,P,S}

- To obtain the Rij value for non H-bonding atoms, calculate the

arithmetic mean of the Rii values for the two atom types.

Rij = (Rii + Rjj) / 2

- To obtain the epsij value for non H-bonding atoms, calculate the

geometric mean of the epsii values for the two atom types.

*epsij = sqrt(epsii * epsjj)*

- Note that the Rij_hb value is non-zero for heteroatoms only, and zero for H atoms;

```
    to obtain the length of an H-bond, look up Rij_hb for the heteroatom only;
    this is combined with the Rii value for H in the receptor, in AutoGrid.
    For example, the Rij_hb for OA-HD H-bonds will be (1.9 + 1.0) Angstrom,
    and the weighted epsij_hb will be 5.0 kcal/mol * FE_coeff_hbond.
"""

#Free energy coefficients for AD4 potentials
FE_coeff_vdw = 0.1560
FE_coeff_hbond = 0.0974
FE_coeff_estat = 0.1465
FE_coeff_desolv = 0.1159
FE_coeff_tors = 0.2744

#Desolvation constant (unsure)
sigma = 3.5

#Electrostatics constants
epsilon_0 = 78.4
A = -8.5525
B = epsilon_0 - A
k = 7.7839
lambd = 0.003627

#Dictionary of atom type objects
atom_types = {}

class Atom:

    def __init__(self):
        self.Rii = None
        self.Rij_hb = None
        self.rec_index = None
        self.epsii = None
        self.solpar = None
```

```
        self.epsij_hb = None
        self.map_index = None
        self.vol = None
        self.hbond = None
        self.bond_index = None

#Create the objects
with open('AD4_parameters.dat', 'r') as handle:
    lines = handle.readlines()

    for line in lines:
        if line.startswith('atom_par'):
            tmp_values = line.split()[1:]
            atom_type = tmp_values[0]
            values = tmp_values[1:]

            #Fill in the object
            atom_types[atom_type] = Atom()
            atom_types[atom_type].Rii = float(values[0])
            atom_types[atom_type].epsii = float(values[1])
            atom_types[atom_type].vol = float(values[2])
            atom_types[atom_type].solpar = float(values[3])
            atom_types[atom_type].Rij_hb = float(values[4])
            atom_types[atom_type].epsij_hb = float(values[5])
            atom_types[atom_type].hbond = float(values[6])
            atom_types[atom_type].rec_index = float(values[8])
            atom_types[atom_type].map_index = float(values[9])
            atom_types[atom_type].bond_index = float(values[9])

def interatomicDistance(ligand_coord, protein_coord):
    #Calculates distance between ligand coordinate and protein coordinate
    squared_distance = 0.0

    for i, j in zip(ligand_coord, protein_coord):
```

```
        squared_difference = (i-j)**2
        squared_distance += squared_difference

    return squared_distance**.5

def vdw(ligand_atom, receptor_atom, rij, vdw_cutoff=9):
    #Calculates LJ (12,6) potential for vdw interactions, given cutoff. else vdw=0
    output = 0
    if rij < vdw_cutoff:

        #Calculate Rij and epsij
        Rij = (atom_types[ligand_atom].Rii + atom_types[receptor_atom].Rii)/2
        product = atom_types[ligand_atom].epsii * atom_types[receptor_atom].epsii
        epsij = np.sqrt(product)

        #requil is the equilibrium distance between the atom types
        requil = Rij

        #Using equation from (Morris et al., 1996)
        Aij = epsij * (requil**12)
        Bij = 2 * epsij * (requil**6)
        repulsion = Aij/(rij**12)
        attraction = Bij/(rij**6)

        output = FE_coeff_vdw * (repulsion - attraction)

    return output


def hbonding(ligand_atom, receptor_atom, rij, hbond_cutoff=2.8):
    #Calculates LJ (12,10) potential for H bonding. Simple model not taking into account H-
acceptor and H-donor angles
    #cutoff is the distance between H and H acceptor
    output = 0
```

```
    if  rij < hbond_cutoff:
        #Note the difference from before: Rij_h vs Rii and epsij_h vs epsii
        #One has to be HD or HS while the other one has to be a hydrogen acceptor
        if ligand_atom in ['HD','HS'] and receptor_atom in ['NA','NS','OA','OS','SA']:
            Rij_hb = atom_types[receptor_atom].Rij_hb
            epsij_hb = atom_types[receptor_atom].epsij_hb

        elif receptor_atom in ['HD','HS'] and ligand_atom in ['NA','NS','OA','OS','SA']:
            Rij_hb = atom_types[ligand_atom].Rij_hb
            epsij_hb = atom_types[ligand_atom].epsij_hb

        else:
            Rij_hb = 0
            epsij_hb = 0

        #requil is the equilibrium H bond distance between the atom types
        requil_hb = Rij_hb

        #Using equation from (Morris et al., 1996)
        Cij = 5 * epsij_hb * (requil_hb**12)
        Dij = 6 * epsij_hb * (requil_hb**10)
        repulsion = Cij/(rij**12)
        attraction = Dij/(rij**10)

        output = FE_coeff_hbond * (repulsion - attraction)

    return output


#Electrostatics potential function
def electrostatics(ligand_atom, receptor_atom, rij, ligand_atom_charge, receptor_atom_charge):
    #Calculates electrostatic potential
    charge_product = ligand_atom_charge * receptor_atom_charge
```

```
    #Distance-dependent dielectric calculations
    distance_dependent_dielectric = A + B/(1 + k*(math.exp(-lambd*B*rij)))

    output = FE_coeff_estat * charge_product/(distance_dependent_dielectric * rij)

    return output


#Desolvation potential function
def desolvation(ligand_atom, receptor_atom, rij):

    Si = atom_types[ligand_atom].solpar
    Sj = atom_types[receptor_atom].solpar
    Vi = atom_types[ligand_atom].vol
    Vj = atom_types[receptor_atom].vol

    exponent = -(rij**2)/(2 * (sigma**2))
    exp_function = math.exp(exponent)
    output = FE_coeff_desolv * (Si*Vj + Sj*Vi) * exp_function

    return output


#Parsing the ligand and receptorfiles
ligand_file = sys.argv[1]
receptor_file = sys.argv[2]

#The ligand file
handle1 = open(ligand_file, 'r')
ligand_lines = handle1.readlines()

#The receptor file
handle2 = open(receptor_file, 'r')
```

```
receptor_lines = handle2.readlines()

total_energy = 0
total_free_energy = 0

#Evaluating the potential functions with input files
for ligand_line_idx, ligand_line in enumerate(ligand_lines):

    if ligand_line.startswith('HETATM'):

        #Extract ligand atom type, charge and coordinates
        ligand_atom_type = ligand_line[77:].strip()
        ligand_atom_charge = float(ligand_line[70:77].strip())
        ligand_coord_x = float(ligand_line[32:39].strip())
        ligand_coord_y = float(ligand_line[40:47].strip())
        ligand_coord_z = float(ligand_line[48:55].strip())
        ligand_coord = [ligand_coord_x, ligand_coord_y, ligand_coord_z]

        for receptor_line_idx, receptor_line in enumerate(receptor_lines):

            if receptor_line.startswith('ATOM'):
                #Extract receptor atom type and charge
                receptor_residue = receptor_line[17:21].strip()
                receptor_residue_no = receptor_line[22:27].strip()
                receptor_atom_type = receptor_line[77:].strip()
                receptor_atom_charge = float(receptor_line[70:77].strip())
                receptor_coord_x = float(receptor_line[32:39].strip())
                receptor_coord_y = float(receptor_line[40:47].strip())
                receptor_coord_z = float(receptor_line[48:55].strip())
                receptor_coord = [receptor_coord_x, receptor_coord_y, receptor_coord_z]



                #Calculate interatomic distance
                rij = interatomicDistance(ligand_coord, receptor_coord)
```

```
            #Calculate potentials
            t1 = vdw(ligand_atom_type, receptor_atom_type, rij)
            t2 = hbonding(ligand_atom_type, receptor_atom_type, rij)
            t3 = electrostatics(ligand_atom_type, receptor_atom_type, rij, ligand_atom_charge,
receptor_atom_charge)
            #t4 = desolvation(ligand_atom_type, receptor_atom_type, rij)
            atomi_atomj_energy = t1+t2+t3#+t4
            total_energy += atomi_atomj_energy
            total_free_energy += atomi_atomj_energy

            current_residue = receptor_residue+receptor_residue_no
            next_residue = receptor_lines[receptor_line_idx+1]
            next_residue = next_residue[17:21].strip() + next_residue[22:27].strip()

            if current_residue != next_residue:
               #Reset energies to zero
               print(ligand_atom_type, receptor_residue+receptor_residue_no, total_energy)
               total_energy = 0

            #Print the energy for the last residue in any chain
            elif 'TER' in receptor_lines[receptor_line_idx+1]:
               print(ligand_atom_type, receptor_residue+receptor_residue_no, total_energy)
               total_energy = 0
if '--score' in sys.argv:
   print(total_free_energy)
```

4.10. Molecular Dynamics – running (MD_main.sh)

```
#!/bin/bash
#Usage: MD_main.sh <receptor_protonated.pdb> <ligand_GMX.gro> <prepare|mdrun>
#*_protonated.pdb is created using "protonate_receptor.py receptor.pdb >
receptor_protonated.pdb"

mdpPath='/home/olivier/project/mdpfiles/'
```

```
box_type='triclinic'
distance_from_box="1.5"
force_field='amber03'
water_model='spc'
protein_basename=`basename $1 'protonated.pdb'` #old version of basename
protein_output="${protein_basename}processed.gro"
ligand_basename=`basename $2 '_GMX.gro'`
curdir=`pwd`

#Function to call submit mdruns to chpc cluster
function submit_chpc {

   #args = <queue> <nvt|npt|md>
   queue=$1
   output="${2}.pbs"
   > ${output}
   echo '#!/bin/bash' >> ${output}
   echo '#PBS -l select=10:ncpus=8:mpiprocs=8:jobtype=nehalem,place=excl' >> ${output}
   echo '#PBS -l select=10:ncpus=8:mpiprocs=8' >> ${output}
   echo '#PBS -l walltime=100:00:00' >> ${output}
   echo "#PBS -N oli_${2}" >> ${output}
   echo "#PBS -q ${queue}" >> ${output}
   echo "#PBS -M oliserand@gmail.com" >> ${output}
   echo "#PBS -m be" >> ${output}
   echo "#PBS -V" >> ${output}
   echo "#PBS -e ${curdir}/jobfile.err" >> ${output}
   echo "#PBS -o ${curdir}/jobfile.out" >> ${output}
   #Module specs
   echo "MODULEPATH=/opt/gridware/bioinformatics/modules:$MODULEPATH" >> ${output}
   echo "source /etc/profile.d/modules.sh" >> ${output}
   echo "cd ${curdir}" >> ${output}
   echo "#######module add " >> ${output}
   echo "module add gromacs/4.5.7" >> ${output}
   echo "module add openmpi/openmpi-1.6.5-gnu" >> ${output}
```

```
    echo "OMP_NUM_THREADS=1" >> ${output}
    echo "NP=`cat ${PBS_NODEFILE} | wc -l`" >> ${output}
    echo "cd ${curdir}" >> ${output}

    echo "EXE=\"mdrun_mpi\"" >> ${output}
    echo "ARGS=\"-deffnm md_0_1\"" >> ${output}
    echo "mpirun -np ${NP} -machinefile ${PBS_NODEFILE} ${EXE} ${ARGS}" >> ${output}

    qsub ${output}
    echo "Submitted oli_${2}"
}

function submit_perkin {

    #Submits equilibration steps to chemistry dept (example queues: perkin, priority, tf)
    #args = <queue> <nvt|npt|md>
    queue=$1
    output="${2}.pbs"

    > ${output}
    echo '#!/bin/bash' >> ${output}
    echo '#PBS -l nodes=1:ppn=8,walltime=10:00:00' >> ${output}
    echo "#PBS -N oli_${2}" >> ${output}
    echo "#PBS -q ${queue}" >> ${output}
    echo "#PBS -d `pwd`" >> ${output}
    #if [ $2 == 'md' ]
    #    then echo "g_mdrun -nt 8 -deffnm md_0_1" >> ${output}
    #else
    echo "g_mdrun -deffnm `pwd`/${2}" >> ${output}
    #fi
    qsub ${output}
    echo 'Submitted oli_'$2 'job'
    while qstat | grep "oli_${2}"
    do sleep 60
```

```
    done

}

#Run at chemistry dept
if [ $3 == 'prepare' ];then


    g_pdb2gmx -f $1 -ff ${force_field} -o ${protein_output} -water ${water_model}
    echo '=> Generated topology file and "_processed.gro" file'

    #Build complex and update topology
    MD_build_complex.py $1 $2
    echo "=> Created complex"

    #Update topology with ligand topology (itp) and molecule name
    sed -i "/forcefield.itp/a\ \n; Include ligand topology\n#include \"$
{ligand_basename}_GMX.itp\"" topol.top
    sed -i "/^Protein2/a\\${ligand_basename}      1" topol.top
    echo "=> Updated topol.top"

    #Set up simulation box
    g_editconf -f conf.gro -o newbox.gro -bt ${box_type} -d ${distance_from_box}
    read

    #Solvate system
    g_genbox -cp newbox.gro -cs spc216.gro -p topol.top -o solv.gro
    echo "=> System solvated"
    read

    #Prepare tpr file to add ions to system
    g_grompp -f ${mdpPath}em.mdp -c solv.gro -p topol.top -o ions.tpr
    echo '=> Generated "ions.tpr"'
    read
```

```
#Add ions to system
echo 15 | g_genion -s ions.tpr -o solv_ions.gro -p topol.top -pname NA -nname CL -conc 0.15 -neutral
echo "=> Added ions/ updated topol.top"
read

#Run energy minimization and generate potential energy plot
g_grompp -f ${mdpPath}em_real.mdp -c solv_ions.gro -p topol.top -o em.tpr
echo '=> Generated "em.tpr" for minimization '
read

g_mdrun -deffnm em
echo 11 | g_energy -f em.edr -o em_pot_energy.xvg
echo "=> System minimized. Generated "em_pot_energy.xvg" for system potential energy"
read

#Add ligand restraints and update topology
echo 2 | g_genrestr -f $2 -o posre_lig.itp -fc 1000 1000 1000
echo "Generated restraints for ligand"

sed -i '/_GMX.itp"/a\\n; Ligand position restraints\n#ifdef POSRES\n#include "posre_lig.itp"\n#endif' topol.top
echo 'Updated "topol.top" with ligand restraints'

#Create index file for nvt equilibration
echo "1 | 13" && echo 'q' | g_make_ndx -f em.gro -o index.ndx
echo 'Generated "index.ndx" file for NVT equilibration'
read

 #NVT equilibration
g_grompp -f ${mdpPath}nvt.mdp -c em.gro -p topol.top -n index.ndx -o nvt.tpr
echo 'Generated "nvt.tpr" for NVT equilibration preparation'
submitjob perkin nvt
```

```
    #NPT equilibration
    g_grompp -f ${mdpPath}npt.mdp -c nvt.gro -t nvt.cpt -p topol.top -n index.ndx -o npt.tpr
    echo 'Generated "npt.tpr" for NPT equilibration preparation'
    submitjob perkin npt

    #Production md
    g_grompp -f ${mdpPath}md.mdp -c npt.gro -t npt.cpt -p topol.top -n index.ndx -o md_0_1.tpr
    echo 'Generated "md_0_1.tpr" for production MD'
    submitjob perkin md

#Run at chpc
elif [ $3 == 'mdrun' ];then
    submit_chpc workq md
fi
```

4.11. Molecular Dynamics – plotting (MDPlot.sh)

```
#!/bin/bash
#To run inside results directory

curdir=`pwd`
em_edrfile="../em.edr"
nvt_edrfile="../nvt.edr"
npt_edrfile="../npt.edr"
md_edrfile="../md_0_1.edr"
md_tprfile="../md_0_1.tpr"
md_xtcfile="../md_0_1.xtc"

if ! [ -d plots ];then
    mkdir plots
fi

#Switch to plots directory
cd ${curdir}/plots
```

#Energy minimization

echo "11" | g_energy -f ${em_edrfile} -o em_pot.xvg

#NVT equilibration

echo 13 17 | g_energy -f ${nvt_edrfile} -o nvt_pot-temp.xvg

#NPT equilibration

echo "13 18 16" | g_energy -f ${npt_edrfile} -o npt_pot-pres-temp.xvg

#Production MD

echo "12 17 15" | g_energy -f ${md_edrfile} -o md_pot-pres-temp.xvg

echo "1 13" | g_hbond -s ${md_tprfile} -f ${md_xtcfile} -num md_hbond_prot-lig.xvg

#Correct for periodicity from PBC, followed by rmsd

echo "0" | trjconv -s ${md_tprfile} -f ${md_xtcfile} -o md_0_1_noPBC.xtc -pbc mol -ur compact

echo "4 3" | g_rms -s ${md_tprfile} -f md_0_1_noPBC.xtc -o md_rms_calpha.xvg -tu ns

#Align/ fit frames

trjconv -s ${md_0_1.tpr} -f md_0_1_noPBC.xtc -o md_0_1_noPBC_fit.xtc -fit progressive

#Gyration radius

echo "1" | g_gyrate -f md_0_1_noPBC.xtc -s ${md_tprfile} -o md_gyrate.xvg

4.12. Molecular Dynamics – plotting (MDPlot.R)

setwd('~/git/project//molecularDynamics/final/')

path11 = '11_apo_SANC00290.acpype/plots/'

path32 = '32_apo_SANC00381.acpype/plots/'

path33 = '33_apo_SANC00347.acpype/plots/'

path39 = '39_apo_SANC00585.acpype/plots/'

############################## Energy minimization ##############################

datem11 = read.table(paste(path11, 'em_pot.xvg', sep=''), header=F, skip=9 , comment.char='@', col.names=c('time','potential_energy'))

```
datem32 = read.table(paste(path32, 'em_pot.xvg', sep=''), header=F, skip=9 , comment.char='@', col.names=c('time','potential_energy'))

datem33 = read.table(paste(path33, 'em_pot.xvg', sep=''), header=F, skip=9 , comment.char='@', col.names=c('time','potential_energy'))

datem39 = read.table(paste(path39, 'em_pot.xvg', sep=''), header=F, skip=9 , comment.char='@', col.names=c('time','potential_energy'))


#EM Potential

par(mfrow=c(1,1))

xmin = min(c(nrow(datem11), nrow(datem32), nrow(datem33), nrow(datem39)))

xmax = max(c(nrow(datem11), nrow(datem32), nrow(datem33), nrow(datem39)))

ymin = min(c(datem11$potential_energy, datem32$potential_energy, datem33$potential_energy, datem39$potential_energy))

ymax = max(c(datem11$potential_energy, datem32$potential_energy, datem33$potential_energy, datem39$potential_energy))


plot(1:nrow(datem11), datem11$potential_energy,
    type='n',
    ylim=c(ymin, ymax),
    xlim=c(1, xmax),
    xlab='Steps',
    ylab='Potential Energy (kJ/mol)',
    main='Change in potential energy during energy minimization')
lines(1:nrow(datem11), datem11$potential_energy, col=1)
lines(1:nrow(datem32), datem32$potential_energy, col=2)
lines(1:nrow(datem33), datem33$potential_energy, col=3)
lines(1:nrow(datem39), datem39$potential_energy, col=4)


legend(600,-450000, legend = c('11_apo:SANC00290 complex',
                    '32_apo:SANC00381 complex',
                    '33_apo:SANC00347 complex',
                    '39_apo:SANC00585 complex'),
    lty=1,
    col=1:4,
    cex=0.9,
```

```
    bty='n')
##############################################################################

############################## NVT equilibration ##############################
datnvt11 = read.table(paste(path11, 'nvt_pot-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature'))
datnvt32 = read.table(paste(path32, 'nvt_pot-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature'))
datnvt33 = read.table(paste(path33, 'nvt_pot-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature'))
datnvt39 = read.table(paste(path39, 'nvt_pot-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature'))



#NVT Potential
par(mfrow=c(2,2))
ymin = min(c(datnvt11$potential_energy, datnvt32$potential_energy, datnvt33$potential_energy,
datnvt39$potential_energy))
ymax = max(c(datnvt11$potential_energy, datnvt32$potential_energy, datnvt33$potential_energy,
datnvt39$potential_energy))
plot(datnvt11$time, datnvt11$potential_energy,
    type='l',
    col=1,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (kJ/mol)',
    main='Change in potential energy during NVT equilibration:\n11_apo/ SANC00290 complex')
plot(datnvt32$time, datnvt32$potential_energy,
    type='l',
    col=2,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (kJ/mol)',
    main='Change in potential energy during NVT equilibration:\n32_apo/ SANC00381 complex')
plot(datnvt33$time, datnvt33$potential_energy,
    type='l',
```

```
    col=3,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (kJ/mol)',
    main='Change in potential energy during NVT equilibration:\n33_apo/ SANC00347 complex')
plot(datnvt39$time, datnvt39$potential_energy,
    type='l',
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (kJ/mol)',
    main='Change in potential energy during NVT equilibration:\n39_apo/ SANC00585 complex')


#NVT Temperature
par(mfrow=c(2,2))
ymin = min(c(datnvt11$temperature, datnvt32$temperature, datnvt33$temperature,
datnvt39$temperature))
ymax = max(c(datnvt11$temperature, datnvt32$temperature, datnvt33$temperature,
datnvt39$temperature))
plot(datnvt11$time, datnvt11$temperature,
    type='l',
    col=1,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NVT equilibration:\n11_apo/ SANC00290 complex')
plot(datnvt32$time, datnvt32$temperature,
    type='l',
    col=2,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature energy during NVT equilibration:\n32_apo/ SANC00381
complex')
```

```
plot(datnvt33$time, datnvt33$temperature,
    type='l',
    col=3,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NVT equilibration:\n33_apo/ SANC00347 complex')
plot(datnvt39$time, datnvt39$temperature,
    type='l',
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NVT equilibration:\n39_apo/ SANC00585 complex')
################################################################################

############################## NPT equilibration ##############################
datnpt11 = read.table(paste(path11, 'npt_pot-pres-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature','pressure'))
datnpt32 = read.table(paste(path32, 'npt_pot-pres-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature','pressure'))
datnpt33 = read.table(paste(path33, 'npt_pot-pres-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature','pressure'))
datnpt39 = read.table(paste(path39, 'npt_pot-pres-temp.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','potential_energy','temperature','pressure'))

#NPT Potential
par(mfrow=c(2,2))
ymin = min(c(datnpt11$potential_energy, datnpt32$potential_energy, datnpt33$potential_energy,
datnpt39$potential_energy))
ymax = max(c(datnpt11$potential_energy, datnpt32$potential_energy, datnpt33$potential_energy,
datnpt39$potential_energy))
plot(datnpt11$time, datnpt11$potential_energy,
    type='l', col=1,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
```

```
    ylab='Potential Energy (KJ/mol)',
    main='Change in potential energy during NPT equilibration:\n11_apo/ SANC00290 complex')

plot(datnpt32$time, datnpt32$potential_energy,
    type='l',
    col=2,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (KJ/mol)',
    main='Change in potential energy during NPT equilibration:\n32_apo/ SANC00381 complex')
plot(datnpt33$time, datnpt33$potential_energy,
    type='l',
    col=3,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (KJ/mol)',
    main='Change in potential energy during NPT equilibration:\n33_apo/ SANC00347 complex')
plot(datnpt39$time, datnpt39$potential_energy,
    type='l',
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Potential Energy (KJ/mol)',
    main='Change in potential energy during NPT equilibration:\n39_apo/ SANC00585 complex')

#NPT Temperature
par(mfrow=c(2,2))
ymin = min(c(datnpt11$temperature, datnpt32$temperature, datnpt33$temperature,
datnpt39$temperature))
ymax = max(c(datnpt11$temperature, datnpt32$temperature, datnpt33$temperature,
datnpt39$temperature))
plot(datnpt11$time, datnpt11$temperature,
    type='l',
    col=1,
```

```
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NPT equilibration:\n11_apo/ SANC00290 complex')
plot(datnpt32$time, datnpt32$temperature,
    type='l',
    col=2,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NPT equilibration:\n32_apo/ SANC00381 complex')
plot(datnpt33$time, datnpt33$temperature,
    type='l',
    col=3,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NPT equilibration:\n33_apo/ SANC00347 complex')
plot(datnpt39$time, datnpt39$temperature,
    type='l',
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    main='Change in temperature during NPT equilibration:\n39_apo/ SANC00585 complex')

#Pressure
par(mfrow=c(2,2))

plotRunningAvg = function(pressure_dat){
  stepsize = 30
  running_average = c()
  maxlim = length(pressure_dat$pressure) - stepsize
  for (i in 1:maxlim){
```

```
    running_average = append(running_average, mean(pressure_dat$pressure[i:(i+stepsize)]))
  }
  lines(pressure_dat$time[1:length(running_average)], running_average,
        col='red',
        lwd=3)
}


ymin = min(c(datnpt11$pressure, datnpt32$pressure, datnpt33$pressure, datnpt39$pressure))
ymax = max(c(datnpt11$pressure, datnpt32$pressure, datnpt33$pressure, datnpt39$pressure))
lwd=0.6
plot(datnpt11$time, datnpt11$pressure,
     type='l',
     col=1,
     ylim=c(ymin, ymax),
     xlab='Time (ps)',
     ylab='Pressure (bar)',
     lwd=lwd,
     main='Change in pressure during NPT equilibration:\n11_apo/ SANC00290 complex')
plotRunningAvg(datnpt11)

plot(datnpt32$time, datnpt32$pressure,
     type='l',
     col=2,
     ylim=c(ymin, ymax),
     xlab='Time (ps)',
     ylab='Pressure (bar)',
     lwd=lwd,
     main='Change in pressure during NPT equilibration:\n32_apo/ SANC00381 complex')
plotRunningAvg(datnpt32)

plot(datnpt33$time, datnpt33$pressure,
     type='l',
     col=3,
```

```
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Pressure (bar)',
    lwd=lwd,
    main='Change in pressure during NPT equilibration:\n33_apo/ SANC00347 complex')
plotRunningAvg(datnpt33)

plot(datnpt39$time, datnpt39$pressure,
    type='l',
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Pressure (bar)',
    lwd=lwd,
    main='Change in pressure during NPT equilibration:\n39_apo/ SANC00585 complex')
plotRunningAvg(datnpt39)

#NPT Temperature
par(mfrow=c(2,2))

plotRunningAvg = function(pressure_dat){
  stepsize = 30
  running_average = c()
  maxlim = length(pressure_dat$temperature) - stepsize
  for (i in 1:maxlim){
    running_average = append(running_average, mean(pressure_dat$temperature[i:(i+stepsize)]))
  }
  lines(pressure_dat$time[1:length(running_average)], running_average,
        col='red',
        lwd=3)
}

ymin = min(c(datnpt11$temperature, datnpt32$temperature, datnpt33$temperature,
datnpt39$temperature))
```

```
ymax = max(c(datnpt11$temperature, datnpt32$temperature, datnpt33$temperature,
datnpt39$temperature))
plot(datnpt11$time, datnpt11$temperature,
    type='l',
    col=1,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    lwd=lwd,
    main='Change in temperature during NPT equilibration:\n11_apo/ SANC00290 complex')
plotRunningAvg(datnpt11)

plot(datnpt32$time, datnpt32$temperature,
    type='l',
    col=2,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    lwd=lwd,
    main='Change in temperature during NPT equilibration:\n32_apo/ SANC00381 complex')
plotRunningAvg(datnpt32)

plot(datnpt33$time, datnpt33$temperature,
    type='l',
    col=3,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    lwd=lwd,
    main='Change in temperature during NPT equilibration:\n33_apo/ SANC00347 complex')
plotRunningAvg(datnpt33)

plot(datnpt39$time, datnpt39$temperature,
    type='l',
```

```
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ps)',
    ylab='Temperature (K)',
    lwd=lwd,
    main='Change in temperature during NPT equilibration:\n39_apo/ SANC00585 complex')
plotRunningAvg(datnpt39)
################################################################################

############################## MD gyration ##############################
datmd11_gyr = read.table(paste(path11, 'md_gyrate.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','Rg','Rgx','Rgy','Rgz'))
datmd32_gyr = read.table(paste(path32, 'md_gyrate.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','Rg','Rgx','Rgy','Rgz'))
datmd33_gyr = read.table(paste(path33, 'md_gyrate.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','Rg','Rgx','Rgy','Rgz'))
datmd39_gyr = read.table(paste(path39, 'md_gyrate.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','Rg','Rgx','Rgy','Rgz'))

par(mfrow=c(2,2))
ymax = max(c(datmd11_gyr$Rg, datmd32_gyr$Rg, datmd33_gyr$Rg, datmd39_gyr$Rg))
ymin = min(c(datmd11_gyr$Rg, datmd32_gyr$Rg, datmd33_gyr$Rg, datmd39_gyr$Rg))
plot(datmd11_gyr$time/1000, datmd11_gyr$Rg,
    col=1,
    ylim=c(ymin, ymax),
    type='l',
    xlab='Time (ns)',
    ylab='Gyration radius (nm)',
    main='Change in gyration radius during production MD:\n11_apo/ SANC00290 complex')
plot(datmd32_gyr$time/1000, datmd32_gyr$Rg,
    col=2,
    lwd=0.3,
    ylim=c(ymin, ymax),
    type='l',
    xlab='Time (ns)',
```

```
    ylab='Gyration radius (nm)',
    main='Change in gyration radius during production MD:\n32_apo/ SANC00381 complex')
plot(datmd33_gyr$time/1000, datmd33_gyr$Rg,
    col=3,
    lwd=0.3,
    ylim=c(ymin, ymax),
    type='l',
    xlab='Time (ns)',
    ylab='Gyration radius (nm)',
    main='Change in gyration radius during production MD:\n33_apo/ SANC00347 complex')
plot(datmd39_gyr$time/1000, datmd39_gyr$Rg,
    col=4,
    lwd=0.3,
    ylim=c(ymin, ymax),
    type='l',
    xlab='Time (ns)',
    ylab='Gyration radius (nm)',
    main='Change in gyration radius during production MD:\n39_apo/ SANC00585 complex')

##############################################################################################

################################ MD RMSD ################################
datmd11 = read.table(paste(path11, 'md_rms_calpha.xvg', sep=''), header=F, skip=9 ,
comment.char='@' )
datmd32 = read.table(paste(path32, 'md_rms_calpha.xvg', sep=''), header=F, skip=9 ,
comment.char='@' )
datmd33 = read.table(paste(path33, 'md_rms_calpha.xvg', sep=''), header=F, skip=9 ,
comment.char='@' )
datmd39 = read.table(paste(path39, 'md_rms_calpha.xvg', sep=''), header=F, skip=9 ,
comment.char='@' )
colnames(datmd32) = colnames(datmd39) = colnames(datmd11) = colnames(datmd33) =
c('Time','RMSD')

ymin = min(c(datmd11$RMSD, datmd32$RMSD, datmd33$RMSD, datmd39$RMSD))
ymax = max(c(datmd11$RMSD, datmd32$RMSD, datmd33$RMSD, datmd39$RMSD))
```

```
par(mfrow=c(2,2))

plot(datmd11$Time, datmd11$RMSD,
    type='l',
    ylim=c(ymin, ymax),
    col=1,
    lwd=0.2,
    xlab='Time (ns)',
    ylab='RMSD (nm)',
    main='C-alpha after lsq fit to Backbone:\n11_apo/ SANC00290 complex')
plot(datmd32$Time, datmd32$RMSD,
    type='l',
    ylim=c(ymin, ymax),
    col=2,
    lwd=0.2,
    xlab='Time (ns)',
    ylab='RMSD (nm)',
    main='C-alpha after lsq fit to Backbone:\n32_apo/ SANC00381 complex')
plot(datmd33$Time, datmd33$RMSD,
    type='l',
    ylim=c(ymin, ymax),
    col=3,
    lwd=0.2,
    xlab='Time (ns)',
    ylab='RMSD (nm)',
    main='C-alpha after lsq fit to Backbone:\n33_apo/ SANC00347 complex')
plot(datmd39$Time, datmd39$RMSD,
    type='l',
    ylim=c(ymin, ymax),
    col=4,
    lwd=0.2,
    xlab='Time (ns)',
    ylab='RMSD (nm)',
    main='C-alpha after lsq fit to Backbone:\n39_apo/ SANC00585 complex')
```

```
################################################################################
############################## MD hydrogen bonds ##############################
datmd11_H = read.table(paste(path11, 'md_hbond_prot-lig.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','hydrogen_bonds','pairs_within_dist'))
datmd32_H = read.table(paste(path32, 'md_hbond_prot-lig.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','hydrogen_bonds','pairs_within_dist'))
datmd33_H = read.table(paste(path33, 'md_hbond_prot-lig.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','hydrogen_bonds','pairs_within_dist'))
datmd39_H = read.table(paste(path39, 'md_hbond_prot-lig.xvg', sep=''), header=F, skip=9 ,
comment.char='@', col.names=c('time','hydrogen_bonds','pairs_within_dist'))


par(mfrow=c(2,2))
ymin = min(c(datmd11_H$hydrogen_bonds, datmd32_H$hydrogen_bonds,
datmd33_H$hydrogen_bonds, datmd39_H$hydrogen_bonds))
ymax = max(c(datmd11_H$hydrogen_bonds, datmd32_H$hydrogen_bonds,
datmd33_H$hydrogen_bonds, datmd39_H$hydrogen_bonds))
plot(datmd11_H$time/1000, datmd11_H$hydrogen_bonds,
    type='l',
    lwd=0.5,
    col=1,
    ylim=c(ymin, ymax),
    xlab='Time (ns)',
    ylab='Number of hydrogen bonds',
    main='Number of hydrogen bonds:\n11_apo/ SANC00290 complex')
plot(datmd32_H$time/1000, datmd32_H$hydrogen_bonds,
    type='l',
    lwd=0.5,
    col=2,
    ylim=c(ymin, ymax),
    xlab='Time (ns)',
    ylab='Number of hydrogen bonds',
    main='Number of hydrogen bonds:\n32_apo/ SANC00381 complex')
plot(datmd33_H$time/1000, datmd33_H$hydrogen_bonds,
    type='l',
    lwd=0.5,
```

```
    col=3,
    ylim=c(ymin, ymax),
    xlab='Time (ns)',
    ylab='Number of hydrogen bonds',
    main='Number of hydrogen bonds:\n33_apo/ SANC00347 complex')
plot(datmd39_H$time/1000, datmd39_H$hydrogen_bonds,
    type='l',
    lwd=0.5,
    col=4,
    ylim=c(ymin, ymax),
    xlab='Time (ns)',
    ylab='Number of hydrogen bonds',
    main='Number of hydrogen bonds:\n39_apo/ SANC00585 complex')
################################################################################
#################### Receptor-ligand  distance plots ###################################
par(mfrow=c(2,2))
xaxis = seq(0, 22, length.out=10002)
dat11_dist = read.table(paste(path11, 'ligand_distance_ASP25(OD2)-LIG(H19).agr', sep=''),
skip=3, comment.char='&', col.names=c('time','distance'))


ymin = min(c(dat11_dist$distance, dat32_dist$distance, dat33_dist$distance, dat39_dist$distance))
ymax = max(c(dat11_dist$distance, dat32_dist$distance, dat33_dist$distance,
dat39_dist$distance))
plot(xaxis, dat11_dist$distance,
    type='l',
    lwd=0.5,
    col=1,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between 11_apo:ASP25:OD2 and SANC00290:H19')


dat32_dist = read.table(paste(path32, 'ligand_distance_ASP25(OD1)-LIG(H19).agr', sep=''),
skip=3, comment.char='&', col.names=c('time', 'distance'))
plot(xaxis, dat32_dist$distance,
```

```
    type='l',
    lwd=0.5,
    col=2,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between 32_apo:ASP25:OD1 and SANC00381:H19')


dat33_dist = read.table(paste(path33, 'ligand_distance_ASP25(OD1)-LIG199(H12).agr', sep=''), skip=3, comment.char='&', col.names=c('time', 'distance'))
plot(xaxis, dat33_dist$distance,
    type='l',
    lwd=0.5,
    col=3,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between 33_apo:ASP25:OD1 and SANC00347:H12')


dat39_dist = read.table(paste(path39, 'ligand_distance_ASP25(OD2)-LIG199(H5).agr', sep=''), skip=3,  comment.char='&', col.names=c('time', 'distance'))
plot(xaxis, dat39_dist$distance,
    type='l',
    lwd=0.5,
    col=4,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between 39_apo:ASP124:OD2 and SANC00585:H5')
################################################################################################
############################## Distance between flaps ##############################
par(mfrow=c(2,2))
xaxis = seq(0, 20, length.out=10002)
dat11_dist = read.table(paste(path11, 'flapdistance-GLY51(CA)-GLY150(CA).agr', sep=''), skip=3, comment.char='&', col.names=c('time','distance'))
```

```
dat32_dist = read.table(paste(path32, 'flapdistance-GLY51(CA)-GLY150(CA).agr', sep=''), skip=3, comment.char='&', col.names=c('time', 'distance'))

dat33_dist = read.table(paste(path33, 'flap_distance_GLY51(CA)-GLY150(CA).agr', sep=''), skip=3, comment.char='&', col.names=c('time', 'distance'))

dat39_dist = read.table(paste(path39, 'flapdistance-GLY51(CA)-GLY150(CA).agr', sep=''), skip=3, comment.char='&', col.names=c('time', 'distance'))

ymin = min(c(dat11_dist$distance, dat32_dist$distance, dat33_dist$distance, dat39_dist$distance))

ymax = max(c(dat11_dist$distance, dat32_dist$distance, dat33_dist$distance, dat39_dist$distance))


plot(xaxis, dat11_dist$distance,
    type='l',
    lwd=0.5,
    col=1,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between flaps: 11_apo:GLY51:CA-GLY150:CA')


plot(xaxis, dat32_dist$distance,
    type='l',
    lwd=0.5,
    col=2,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between flaps: 32_apo:GLY51:CA-GLY150:CA')


plot(xaxis, dat33_dist$distance,
    type='l',
    lwd=0.5,
    col=3,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
```

```
    main='Distance between flaps: 33_apo:GLY51:CA-GLY150:CA')


plot(xaxis, dat39_dist$distance,
    type='l',
    lwd=0.5,
    col=4,
    ylim=c(ymin, ymax),
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between flaps: 39_apo:GLY51:CA-GLY150:CA')
################################################################################
plot(dat32_dist$time, dat32_dist$distance,
    type='l',
    lwd=0.5,
    col=2,
    xlab='time (ns)',
    ylab='Distance (nm)',
    main='Distance between flaps: 32_apo:GLY51:CA-GLY150:CA')
```