Identifying Structurally Important Regions in the HIV-1 Particle as Drug Targets

Edward Wang, Kevin Chen
Amador Valley High School
Grade 11

Contents

1 Background Research

Although many have heralded our triumph over a multitude of infections and viruses, one of the most prominent pathogens is the human immunodeficiency virus, or HIV-1. It has plagued our populations since the 1930s. The virus seeks out and destroys $CD4^+$ T-cells of the immune system, binding to the cells' surface receptors with its glycoproteins, entering the cell [4, 37], and using reverse transcriptase and other viral enzymes to transcribe its RNA genome into cDNA. [1] The virus then inserts this cDNA into the host cell's nucleus, forcing the cell to mass produce parts of the immature virus, known as the Gag polyprotein. These parts self assemble prior to exiting the cell in order to infect other cells. [12]

Over time, enough T-cells are killed that the body's immune system is severely compromised. The individual becomes highly vulnerable to many pathogens, including those that were previously harmless. At this point, the individual has the acquired immunodeficiency syndrome (AIDS). [1]

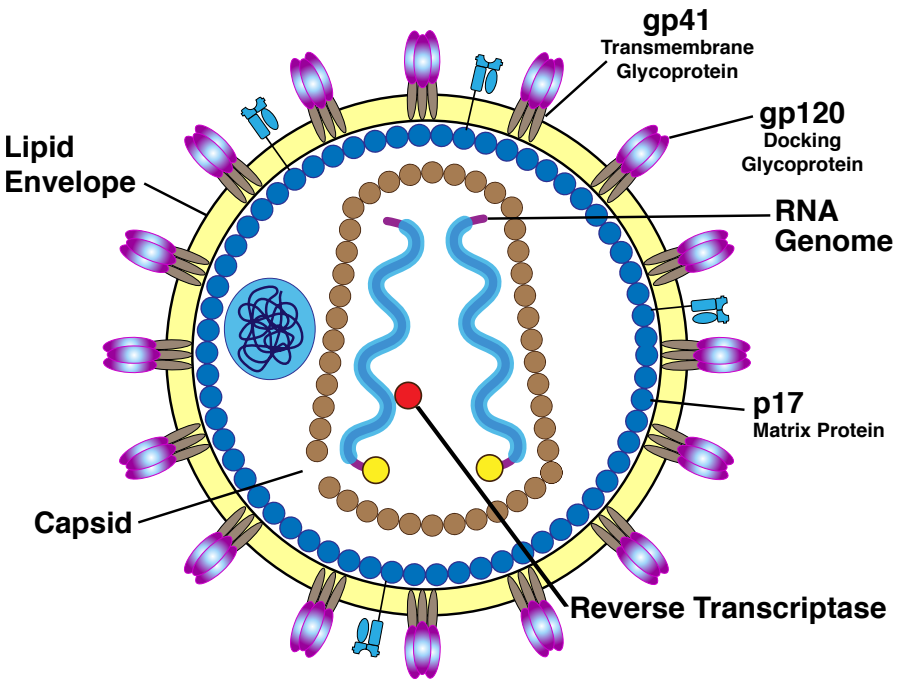


Figure 1: Structure of the HIV-1 particle.

Because of the high error rate during the transcription of RNA to cDNA, HIV-1 has a strikingly high mutation rate — perhaps 1 error per every 2,000 nucleotides. This is a source of difficulty in developing treatments: a mutation in the HIV-1 genome can potentially nullify a ligand if its binding site changes so that the ligand no longer has access to the site. Currently, there is no cure, but there are a variety of pharmaceuticals designed to inhibit the process of HIV-1 assembly and maturation, to prevent its entry into the cell, or to compromise its internal proteins such as reverse transcriptase and the capsid. However, because certain pharmaceuticals may be more or less effective on different viral strains, these antiretrovirals are administered simultaneously in a shotgun approach known as an antiretroviral "cocktail." And while some patients have successfully delayed the onset of AIDS with the cocktail method, no silver bullet that might lead to a universal pharmaceutical or vaccine has been discovered. [13]

Proteins often identified with or targeted in HIV-1 include gp120, gp41, the capsid protein, and reverse transcriptase. Therefore, these were the proteins that we decided to investigate in our project. The glycoproteins gp120 and gp41 are found on the surface of the virus and resemble mushrooms dotting its surface. gp41 can be considered the stalk of the mushroom, lodged between phospholipids in the viral lipid bilayer, while gp120 can be considered the cap. These are the proteins which bind to T-cell receptors. [4] The capsid protein encapsulates the viral RNA and reverse transcriptase, protecting the genetic information inside. [12] There is also evidence to suggest that the capsid is crucial and directly plays a role in the HIV-1 infection of nondividing cells. [38] The reverse transcriptase, again, binds to and transcribes viral RNA into cDNA, which can be read by the host cell.

All proteins have different structure levels: primary structure refers to the amino acid sequence, secondary structure deals with arrangements of amino acids forming beta-pleated-sheets and alpha helices, tertiary structure involves interactions between amino acids that fold the protein and give it a shape, and quaternary structure includes interactions with other proteins. Here, we are interested in removing interactions that maintain or determine tertiary structure, namely salt bridges and disulfide bonds. Salt bridges are amino acid interactions that include both hydrogen bonding and ionic interaction. Disulfide bonds are covalent interactions formed from the thiol groups of two cysteine residues. [23]

Site-directed mutagenesis studies of HIV-1 can turn up intriguing results, such as the identification of a salt bridge in a capsid that inactivates the virus

when the amino acids are changed through a mutation in the corresponding viral RNA. [36] Yet, short of testing in a laboratory, there is no sure-fire method to test the importance of certain amino acids in the function of HIV-1. However, some bioinformatics resources can be used in the prediction of important regions and are readily available online for anyone to access. This project relied heavily on the following online resources:

- The **Protein Data Bank (PDB)** from the Research Collaboratory for Structural Bioinformatics (RCSB) contains a wealth of information about proteins, including their 3D structures in the PDB format (essentially coordinate points of atoms) and their FASTA sequences (amino acid sequences). In this project, we utilized it to obtain the wild-type 3D structures of the four viral proteins listed previously. The works contained in the database are public domain and are accessible from `http://www.rcsb.org/pdb/`.

- **Visual Molecular Dynamics (VMD)** is a powerful program designed to analyze, visualize, and animate 3D protein structures and other molecules. One can use the software to perform a wide variety of calculations, including ones to locate amino acids involved in salt bridges or hydrogen bonds. We used the salt bridges extension in VMD on our four proteins, although one can identify nearly any interaction with the correct VMD commands. VMD is also integrated with molecular dynamics software such as NAMD and the Chemistry at Harvard Macromolecular Mechanics (CHARMM) software. VMD is available from `http://www.ks.uiuc.edu/Research/vmd/`.

- The **Center for Informational Biology in Ochanomizu University's disulfide bond calculator** locates disulfide bonds between cysteine residues of each chain. It is accessible from `http://cib.cf.ocha.ac.jp/bitool/SSBOND/`.

- **RaptorX** is a protein-structure prediction program that, from an arbitrary FASTA sequence input, can produce a model of the tertiary structure of that protein by using structures in the PDB database as templates that it aligns against the original structure, and in fact can use multiple templates as guides to improve its accuracy. We used the server in our project to predict the tertiary structure of the four proteins above with our own modifications in their FASTA sequences. RaptorX

can also return predictions of secondary structure and disorder predictions for the generated model, as well as analyze the sequence to check for protein domains. [27, 24, 25, 26, 39, 40]

- **metaPocket 2.0** is a program for identifying locations for ligand binding called pockets in a protein structure input. It uses a total of eight predictors (LIGSITE, PASS, Q-SiteFinder, SURFNET, Fpocket, GHECOM, ConCavity, and POCASA) to predict pocket sites and then takes extra steps to ensure the accuracy of the predictors, obtain the top results, and rank the results based on a calculation of Z-scores for each pocket site from each predictor. Finally, it searches for amino acid residues around the identified pocket sites as potential ligand binding sites and returns them in its output. We used metaPocket to determine whether the amino acid interactions we identified could even be accessible to pharmaceuticals. [10]

- **Conserved Domain Database (CDD) Search** is a bioinformatics tool from the National Center for Biotechnology Information (NCBI) that identifies conserved protein domains within a given FASTA sequence using the CDD, a database containing domain sequences as well as annotations of their functions. Generally used to predict the function of unknown proteins based on amino acid sequences, we utilized the CDD Search in order to identify amino acids in each of the proteins that are highly conserved among multiple mutations. [21, 22, 20]

- **ProFunc** is a program from the European Bioinformatics Institute (EBI) that is pinpoints a potential function of a protein using its 3D structure and its amino acid sequence. In this project, we used ProFunc to find likely biochemical functions based on 3D structure and the results of its reverse template analysis feature, which uses enzymatic active sites, DNA-binding and ligand-binding sites as well as a template generated from the query protein to identify similar templates in the EBI databases. [15, 16]

- **Java-based Combinatorial Extension (jCE)** is a program designed to perform pairwise alignment of protein structures or perform an alignment of a structure against the PDB database and returns quantifiable data as to the similarity of the structures with one another in the

form of a Z-score. jCE analyzes the protein by dividing it into multiple fragments; the Z-score is a function of the fragments' distances from one another, the number of gaps between fragments, and the average standard deviations of both the distances and number of gaps (Fig. 2). We used the pairwise alignment function of jCE to align the modified protein structures produced from RaptorX with the original protein structures from PDB. The Java-based client is available from `http://pdbx.org/jfatcatserver/`. [29, 33]

$$\rho\left(O_j 1, -z\right) = \rho\left(D_i^{av}, D_i^{sd}, D^{obs}\right) \tag{1}$$

$$\rho\left(\mu, \sigma, x\right) = \begin{cases} 2\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2} dy & , \quad x < \mu \\ 1 & , \quad x \geq \mu \end{cases} \tag{2}$$

Figure 2: jCE finds the Z-score of an alignment by solving equation 1 for z using a normal distribution with an average value of 0 and a standard deviation of 1. [33]

Attempting to find potential drug targets is common, although these investigations are usually done *in vitro* in the laboratory. A program that has come close to computational drug target prediction, Similarity-based Inference of drug-TARgets (SITAR), is indeed a highly effective and well-tested program that predicts conserved ligand binding sites. However, SITAR is partly based on ligand similarity measures, so it requires information on a specific ligand before it can test for its drug targets. [28]

Therefore, it appears that computational prediction of conserved and structurally critical drug targets with no prior information given of a ligand has not been attempted nor published.

2 Introduction

Our method was to find more effective targets for drugs or ligands in HIV-1 through a process that considers regions of a protein potentially critical for the maintenance of tertiary structure ("structurally important regions"),

regions accessible to ligands (called *clefts* or *pockets*), and regions that show conservation of amino acids over time and over mutations.

This project is also our exploration into the effectiveness of publicly accessible bioinformatics tools for scientific research: we only used tools that were available to us freely, partly to probe the limits of modern bioinformatics tools given to the public and partly because we did not have the means nor the funds to run resource-intensive and expensive molecular-dynamics software on home computers (or with cloud-computing services). In addition, there would be clear safety concerns with handling the virus itself. We open this paper with the understanding that more accurate results can be obtained through *in vitro* experiments or through paid software, but that we did not have the privilege of accessing these resources. Instead, we used an assortment of free software to find the regions we were seeking, change them, and then reassemble them using structure-prediction software.

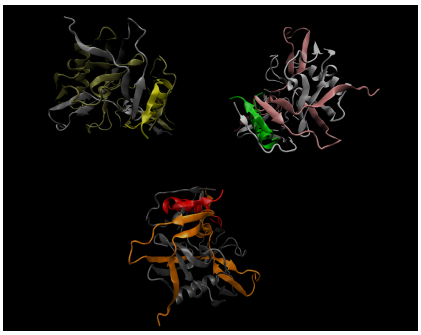
To summarize: by taking into consideration all of the above, we believed we could find amino acids that are required for HIV-1 to function, accessible to drugs, and conserved so that future antiretrovirals based on inactivating regions where these amino acids are located can be more versatile and effective than current pharmaceuticals used to treat the virus.

3 The Experiment

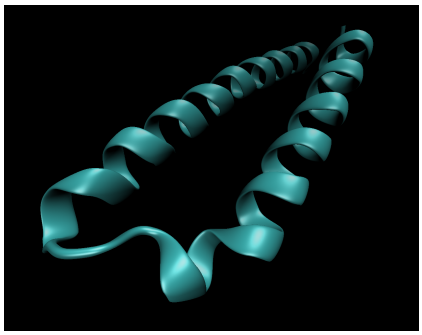
We can separate the experimental process into two parts: identification of the structurally important sites and analysis of their effects on the protein. Whenever possible, we used wild-type proteins from the Protein Data Bank: gp120 (PDB ID 3DNN) [17], gp41 (1I5Y) [18], capsid (3NTE) [7], and reverse transcriptase (2HMI) [6].

3.1 Phase 1: Identifying "Important" Sites

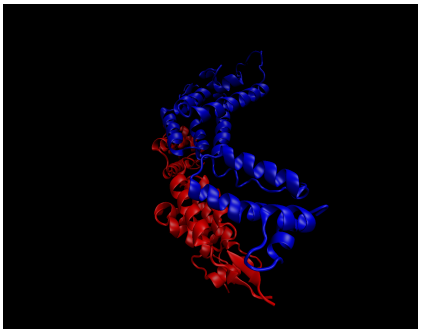
The first phase began with the identification of salt bridges and disulfide bonds. We found salt bridges in each protein chain by running VMD's salt bridge extension with the default settings. The extension searches for ionic interactions between two amino acids with an oxygen–nitrogen cutoff distance of 3.2 angstroms. We used the Center for Informational Biology in Ochanomizu University's disulfide bond calculator that identifies these interactions based on the following criteria: the distance cutoff between the two



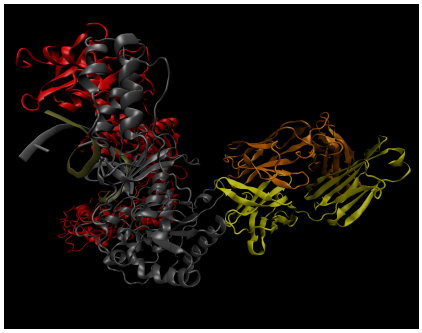
(a) gp120 (PDB ID 3DNN)



(b) gp41 (1I5Y)



(c) capsid (3NTE)



(d) reverse transcriptase (2HMI)

Figure 3: The four original, unmodified proteins we studied in this project.

sulfur atoms is 3.0 angstroms, and the angle of the dihedral C–S–S–C bond is between 60 and 90 degrees.

From there, we used metaPocket to find potential ligand binding pockets, considering only the top ten ranked pockets. We compared the set of amino acids in the predicted pockets with the residues involved in the salt bridges and disulfide bonds. If the residues involved in the salt bridges and disulfide bonds matched up with residues inside a pocket, we then checked if the residues were part of a conserved domain using a CDD Search. (To automate the process of matching residues, we wrote a small command-line program called `matchacids`. Its source code and documentation are available online at `https://github.com/kevin1/matchacids`.)

We ran the CDD search (Fig. 4) under default settings on each protein, with the expect value threshold at 0.01, the low-complexity filter on, and the maximum number of amino acid hits set at 500. We were only concerned with the position of the amino acids that were conserved, as, of course, we already knew the protein domains; none of the results would be surprising. We recorded only the chain and position of amino acids that were consistent among multiple mutations in the database (CDD colored them red), ignoring the less consistent amino acids (blue). If there was more than one conserved domain, we combined the results and recorded the amino acids that were consistent within any of the domains as conserved. At this point, out of the remaining salt bridges and disulfide-bond interactions, only those with both of the amino acids involved in the interaction as conserved were denoted as potentially structurally important sites.

3.2 Phase 2: Effects of the "Important" Sites

The next phase of the experiment was to modify the FASTA sequences of the proteins, deleting the amino acids from the potential sites (salt bridges or disulfide bonds) identified from the first part of the experiment. We deleted only the amino acids that were accessible to ligands in a ligand pocket (according to metaPocket). We created a separate FASTA sequence for each potential site with one or both of the amino acids removed. We then ran these modified FASTA sequences through the RaptorX server and, from its output (Fig. 6), collected the top ten 3D structure predictions for a protein chain's tertiary structure. We would only use the structure with the greatest alignment score (the top scoring model) to run through ProFunc and jCE. Note that RaptorX processed each chain independently of the others.

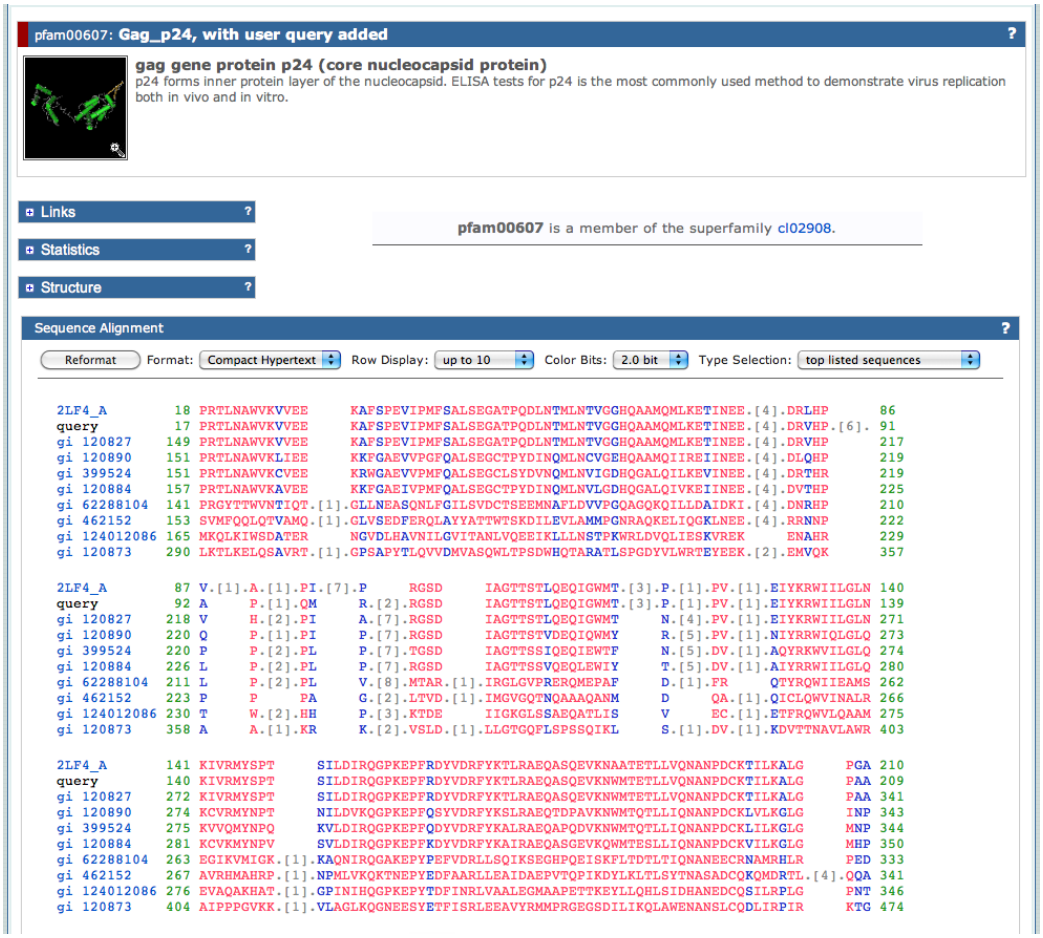


Figure 4: One FASTA sequence comparison result from running a CDD search with the capsid, PDB ID 3NTE.

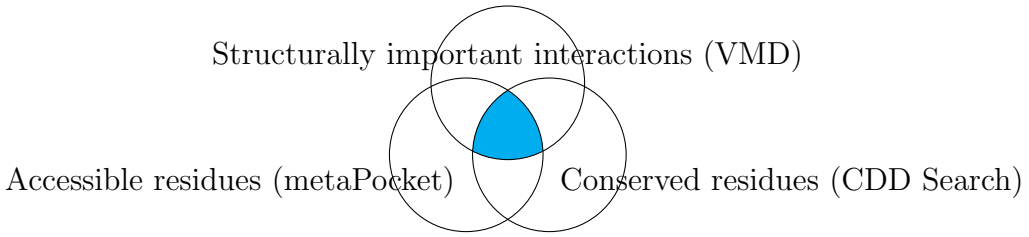


Figure 5: The intersection of the three circles represents the residues identified for potential drug targeting.

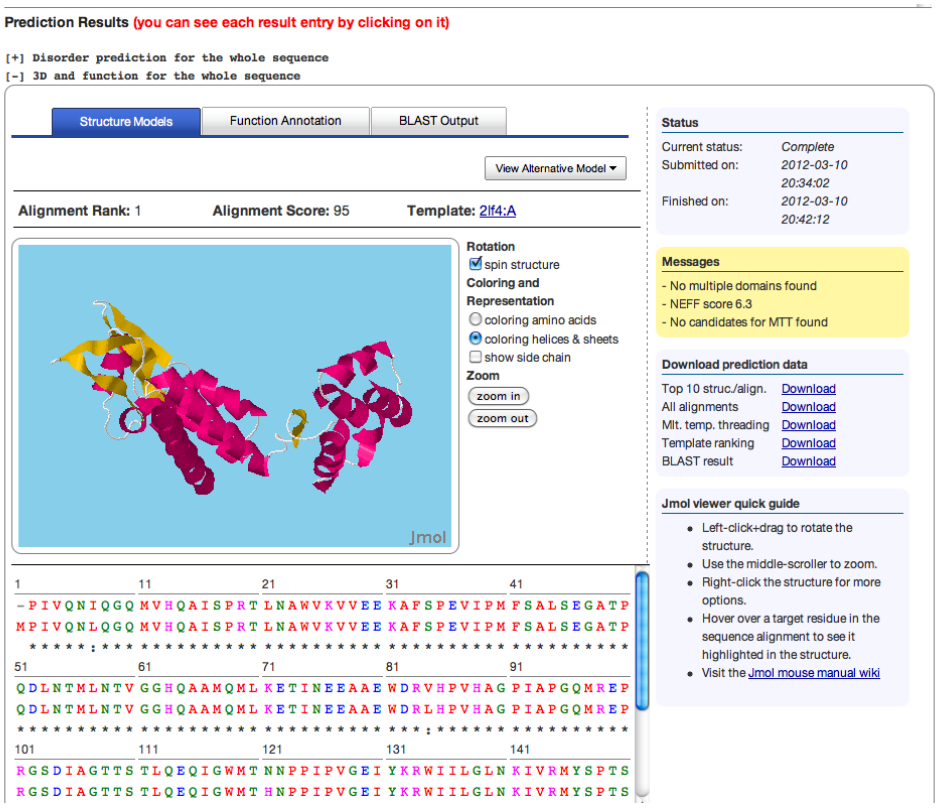


Figure 6: RaptorX result, showing a generated model of a modified capsid protein with amino acid deletions.

In ProFunc, we ran the modified protein chains and recorded the top five reverse template results as well as its prediction of the protein chain's function as a source of qualitative data to be analyzed. We obtained the more quantitative, statistical data through the use of the Java-based Combinatorial Extension (jCE) algorithm, recording the root mean square deviation (RMSD) and the Z-scores. We used the default settings for jCE, with a max gap size of 30 angstroms, no circular permutations, and an RMSD threshold of 99.0 angstroms. We compared the original, unmodified protein chains to themselves using jCE to establish a baseline Z-score. (This baseline is later used to obtain a percent difference so that Z-scores from different protein chains could be compared.)

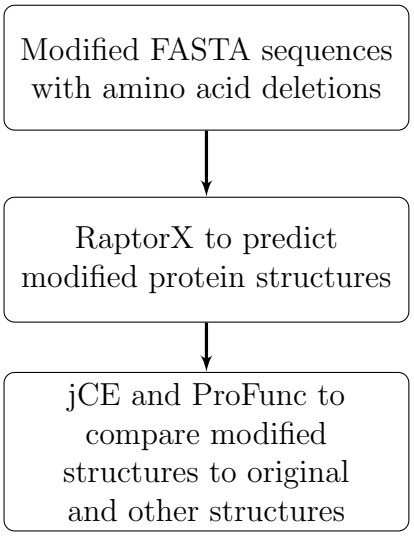


Figure 7: The flowchart is a summary of the steps in phase 2 in sequential order.

The function of a protein depends on its structure — if its structure is changed significantly, then it can no longer fulfill its purpose. Thus, we chose to find salt bridges and disulfide bonds, which are known to be key in strengthening or holding together a protein's shape, or tertiary structure. The rationale behind the ligand pockets is clear: interruption of these bonds wouldn't be possible unless the ligand was present. As for the conserved amino acids, if an amino acid is conserved among multiple mutations of the virus, then it may be critical in maintaining the shape and/or function. Therefore, mutations not containing that amino acid at that position may not able to reproduce.

4 Discussion

The "id numbers" in the graphs of Z-scores obtained from jCE correspond to a particular RaptorX model of the protein with specific amino acid deletions. Each modified protein is listed and explained below, with the exception of gp41, for which metaPocket did not return any ligand pocket. The data indicate that the best result out of the modified gp120 chains was gp120–id28; out of the modified capsid chains, capsid–id3; out of the modified reverse transcriptase chains, rt–id6. Note that in the raw Z-score comparison graph (Fig. 8), *lower* Z-scores indicate less structural similarity to the original chain, and of course, that is the case for *higher* values in the percent difference in Z-score comparison graph (Fig. 9). We decided to calculate the percent differences among the Z-scores because each original protein chain has its own Z-score; there is no universal baseline among the proteins. The percent difference allows us to compare Z-scores among different protein chains. In addition, jCE states that generally, Z-scores higher than 3.5 indicate good structural similarity. However, first, this threshold value varies from protein to protein, and second, it is difficult to conceptualize the degree of *dissimilarity* between the superimposed structures with just a Z-score. Note that the percent difference may not indicate the percentage difference between the structure of the original and the modified structure; we used it instead to more accurately compare the Z-scores among our results only, which are of course different types of protein chains. The jCE results state that all of the modified structures are still similar to the original sequence because all the Z-scores are above 3.5. However, the lowest Z-score in gp120–id28 of 4.58, somewhat close to the threshold value, and the high percent difference of Z-score of 21.8 percent indicate that there is still sizable structural dissimilarity.

It's not surprising that the structures are still significantly similar; we did not expect them to change dramatically after one or two deletions in the amino acid sequence. When RaptorX uses templates to aid its structure generation process, it returns an alignment score (from 0 to 100) of the generated model compared to the template structure. The alignment scores given to the generated model of the capsids using the original protein, 3NTE, as a template, were 91 in id3 and 92 in id11 — essentially a 9 percent and 8 percent difference from the original, very high considering that there was little change in the primary structure.

However significant the quantitative data may be, we must look at the re-

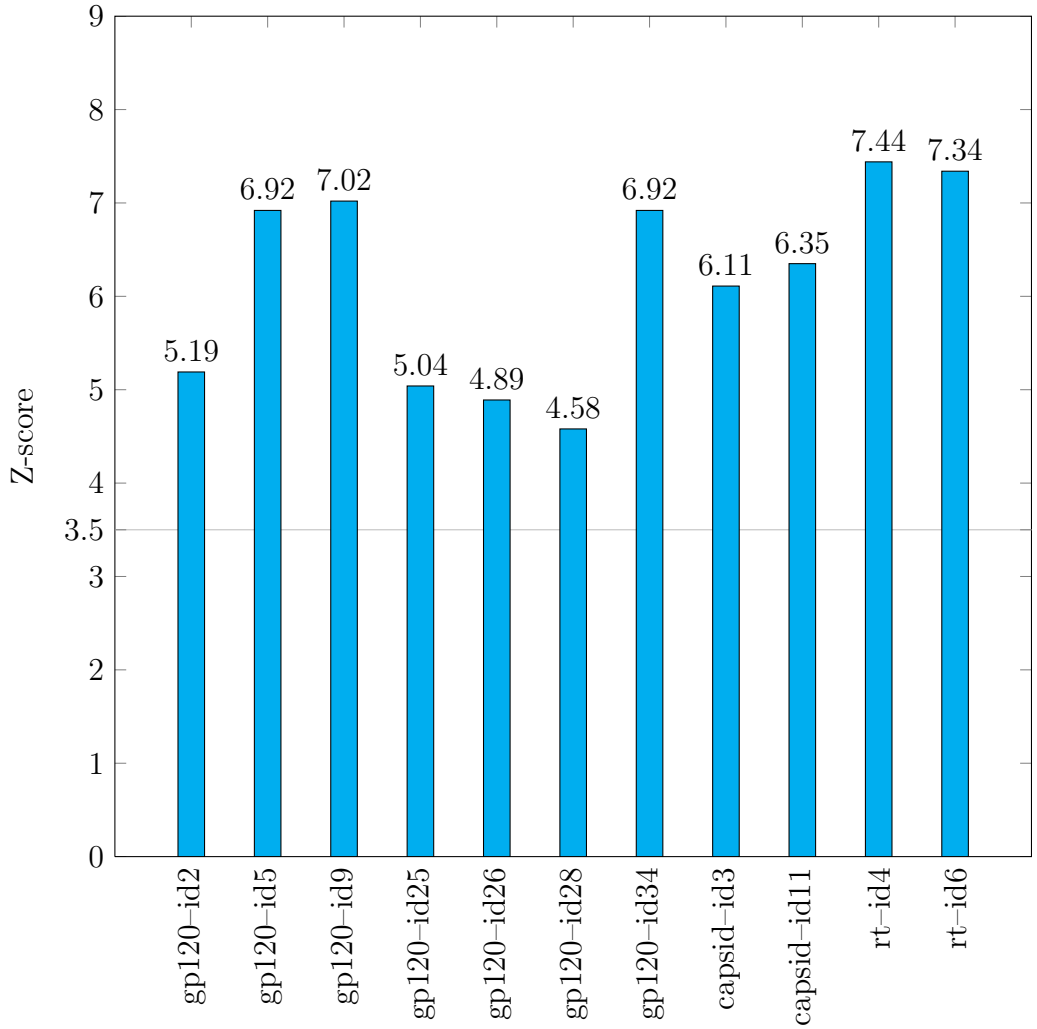


Figure 8: Raw Z-scores for the RaptorX models. Each id number corresponds to a model with a certain amino acid deletion. (The horizontal line indicates that alignments with scores below 3.5 show insignificant structural similarity.)

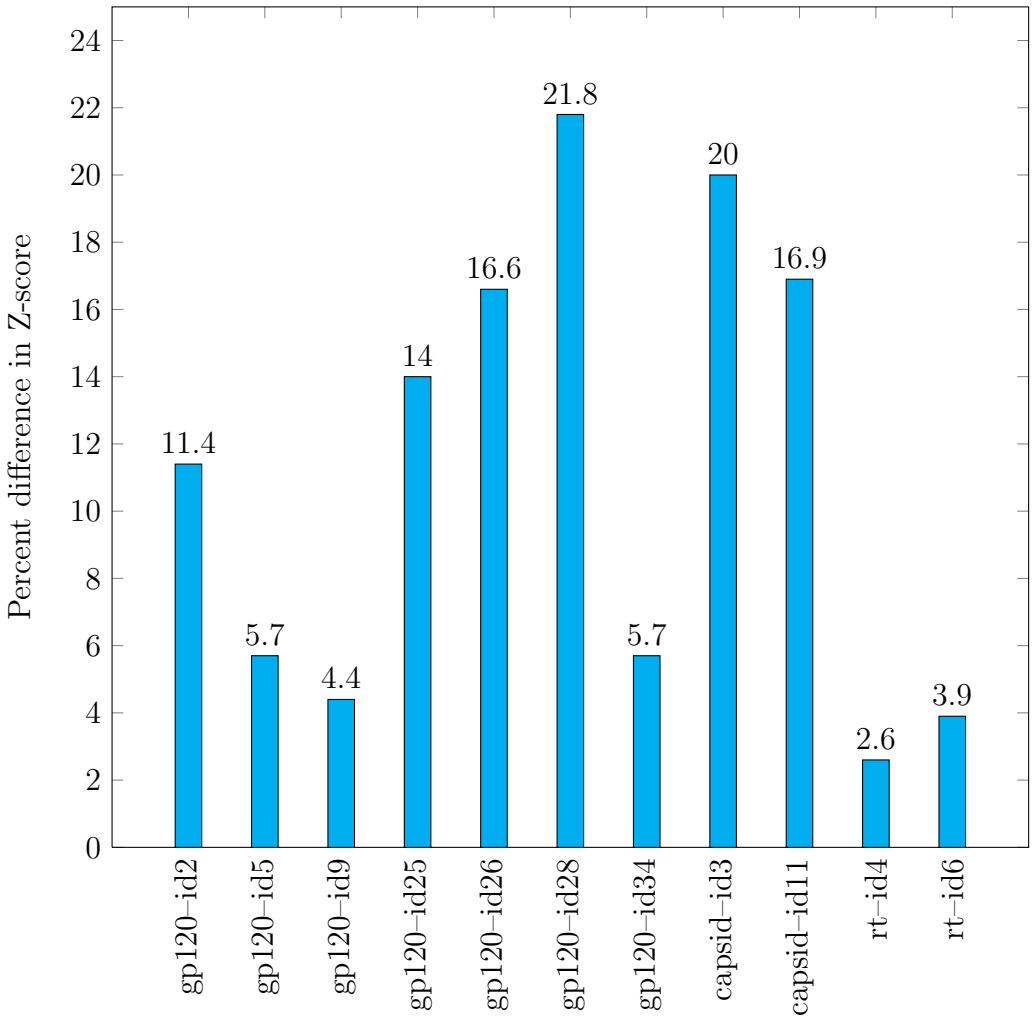


Figure 9: Percent difference in Z-scores for the RaptorX models.

sults in conjunction with the qualitative results and background information to get a fuller picture and assess the viability of each of the sites. Below, we identify each of the id numbers and discuss noteworthy past research on these amino acids or qualitative data in the form of reverse template results from ProFunc or RaptorX templates. Note that we deleted certain amino acids in the interaction based on whether they were accessible in a ligand pocket according to metaPocket; all of the amino acids involved in the interactions listed below are conserved.

4.1 gp120

Nearly all of the ProFunc reverse template results of the modified gp120 structures, as well as the RaptorX template alignment (PDB IDs 3DNL, 3JWD, 3JWO, 2NY0 to 2NY6, 2NXY, 2NYZ, 1YYM, 2B4C), point to the modified structures being more similar to a complexed glycoprotein — that is, more similar to a liganded glycoprotein, bound to Fab b12, antibody 17b, or antibody x5, effectively neutralizing the virus. [41, 14] When considering that these models are more similar to neutralized glycoproteins than the original, all of the sites seem viable.

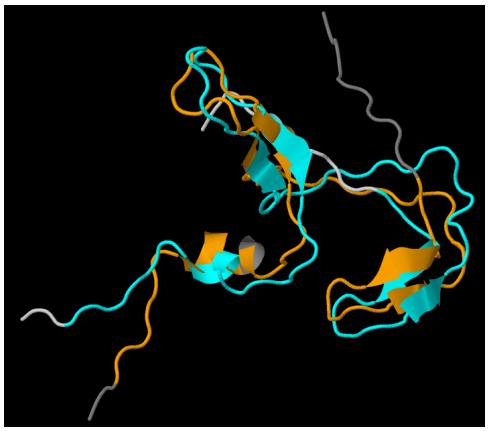


Figure 10: A visualization in jCE of gp120 aligned against gp120–id28, the best site that results in the most structural dissimilarity according to Z-score and percent difference in Z-score values. (The teal model is the original, and the orange model is the modified one.)

- **gp120–id2** (ASP-457 and ARG-469 salt bridge on chain I)
  We removed both amino acids; metaPocket indicated that both amino acids are inside a ligand pocket. ASP-457 is a highly conserved amino acid that directly contacts T-cell CD4 receptors, so it is located in a critical position and is conserved because of its function. [2] Thus, we

believe a ligand focused on binding to this site has great potential for becoming a highly effective antiretroviral.

- **gp120–id5** (GLU-268 and LYS-231 salt bridge on chain E)
  We removed both amino acids.

- **gp120–id9** (ASP-279 and LYS-282 salt bridge on chain H)
  We removed the lysine residue. ASP-279 is another residue that directly contacts CD4 receptors within the C2 region (273-288), so it may be an effective drug target, like the deleted residues in gp120–id2. [35]

- **gp120–id25** (ASP-477 and ARG-480 salt bridge on chain F)
  We removed the arginine residue.

- **gp120–id26** (GLU-466 and ARG-456 salt bridge on chain I)
  We removed the arginine residue.

- **gp120–id28** (Fig. 10) (ASP-477 and ARG-480 salt bridge on chain I)
  This was the best, most dissimilar model according to the statistical results. We deleted both amino acids in this site. Although this site is the best statistical result, we are not aware of any past work providing information on the distinctiveness of these amino acids.

- **gp120–id34** (CYS-228 and CYS-239 disulfide bond on chain E)
  We removed cysteine-228. The MRC/UCT Research Group for Receptor Biology in the University of Cape Town actually conducted a study that substituted a cysteine in for an arginine residue at position 239 for a recombinant variety of a gp120 glycoprotein. The glycoprotein, previously unable to bind to the CD4 receptors, was able to bind to the receptor after the substitution, supporting the prediction that this disulfide bond would make an effective drug target. That this study was conducted *in vitro* also provides some support for the method we have constructed. [9]

4.2 capsid

In the generation of both of these models, RaptorX returns a higher alignment score (95) for the template of PDB ID 2LF4 than for the original protein 3NTE, suggesting that the modified capsids are more similar to 2LF4, which is a monomeric mutant of the capsid protein that is non-infective. [32]

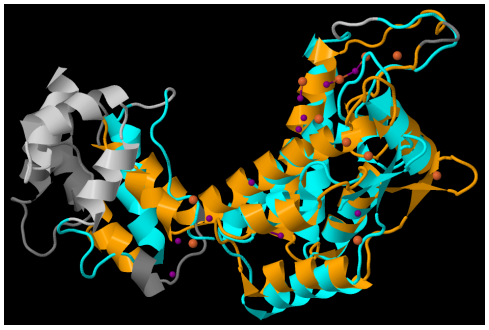


Figure 11: A visualization of capsid aligned against capsid–id3 in jCE. This modification deleted amino acids from the capsid's major homology region and yielded the second best statistical result. (Again, the teal model is the original, and the orange model is the modified one.)

- **capsid–id3** (Fig. 11) (ASP-152 and ARG-154 salt bridge on chain A) This is the second most dissimilar site and the best site among the capsid proteins. We deleted both amino acids. In addition, this deleted interaction is made up of two amino acids, both of which are part of the capsid's major homology region, a conserved functional region of the capsid which is key for its proper assembly. Note that mutations to the major homology region can also affect the assembly of the immature virus polyprotein, Gag. The presence of these amino acids in the major homology region makes this site even more promising. [19, 5]

- **capsid–id11** (GLU-28 and LYS-25 salt bridge on chain B)
  We deleted both amino acids. It appears that in the wild type capsid, GLU-28 is a part of a network of amino acids close to the terminals that interact with each other to connect the capsid subunits together into the capsid as a whole surrounding the viral RNA. However, it also appears that interactions of these terminals can still occur without GLU-28 becoming directly involved, and thus the caspid–id3 site is the better potential drug target. [3]

4.3 reverse transcriptase

The models of the modified reverse transcriptase are actually quite similar to the original, according to the statistical data. As such, it would appear that, as they are the lowest ranking chains in Z-score percent difference and are also rather close to each other in those scores, that these reverse transcriptase sites are less likely to be effective drug targets. With the returned ProFunc and RaptorX results, such as 1N5Y, however, we soon realized that the chain D structures we had generated were not actually reverse transcriptase at all,

but modifications of *Mus musculus* antibodies complexed with the reverse transcriptase. [31] Thus, we could not consider their data.

- **rt–id4** (ASP-91 and ARG-68 salt bridge on chain D)
  We deleted the arginine residue.

- **rt–id6** (GLU-48 and ARG-40 salt bridge on chain D)
  We removed both amino acids.

Once we consider past work and qualitative data along with quantitative data, we see that within gp120, there is evidence to suggest that the sites in gp120–id2, gp120–id9 (direct contact with CD4 receptor), and gp120–id34 (necessary disulfide bond for interaction with CD4 receptor) join the site in gp120–id28 as good candidates for drug targets. However, the percent difference in Z-score of gp120–id34 and gp120–id9 are quite low (5.7 percent and 4.4 percent, respectively), and gp120–id26, which has no past work suggesting that its residues are important, has a better percent difference in Z-score than gp120–id2 (16.6 percent vs. 11.4 percent). This may lead one to suspect whether there is actually a correlation between statistical data of this method and actual functional performance. Again, the basis of our reasoning behind obtaining data on structural similarity is that in proteins, structure is intimately related to function. We still realize, however, that small changes in structure can still dramatically impact function, as in the case of the amino acids deleted from gp120–id34, where the crucial disulfide bond facilitated CD4 binding.

Yet we also see that a deletion of amino acids that we know are in the capsid major homology region (capsid–id3) results in the second highest percent difference by Z-score of 20.0 percent; considerable structural change most likely greatly impacts function. Thus, we reason that gp120–id26 and gp120–id28 may be great candidates for drug targeting but simply have not been investigated. The same may hold true for the LYS-25 residue in capsid–id11, which also has a very high percent difference in Z-score (16.9 percent).

Once again, our goal in this project is to find the best potential sites for drug targets using the method we have constructed and to examine how our method performs (whether it can be used to gather accurate statistical data). Since our method encompasses only freely available, limited resources and is an experiment conducted *in silico*, we anticipated and recognize potential sources of error.

For example, we realize we cannot accurately simulate the effects of a ligand on a protein using our method of deleting specific amino acids. Yet we had little other choice, since we did not have specific ligands to simulate. At first, we considered substituting the amino acid in an interaction for another instead of deleting it, but this might introduce new, unwanted interactions between the new residues and other neighboring residues, whereas the primary goal was simply to eliminate the interaction. These new interactions caused by the substitutions would most likely obscure our data. With the resources available to us, we felt that deletion was the best route to take.

We realize that a great concern in our experimentation method was the use of the RaptorX server to predict the 3D tertiary structure of a modified FASTA sequence. The reasons are twofold. First, the nature of such predictive servers involves error in these types of experiments because they rely heavily on templates from a database to create their models. Second, RaptorX does not take into consideration the multichain aspect of proteins such as glycoprotein gp120; a RaptorX job containing all of the chains in gp120 will be interpreted as a batch job, in which each of the chains predicted separately. We did have data on salt bridges and disulfide bonds across chains obtained through VMD and the Protein Interactions Calculator (PIC) [34], but we were not able to simulate their removal. Processing multichain proteins requires servers such as I-TASSER, which requires its users to have a university email address and therefore is not available to the us and not available to the public. [30] However, in order to more accurately assess the results of inhibiting these potential drug targets, we would ideally use I-TASSER in the next round of experimentation to analyze the effect of quaternary structure.

Despite these potential sources of error, we still believe that our patchwork methodology has identified good candidates for drug targeting. Some of our identified sites included amino acids that have been shown to be crucial to the proteins' function, and one site was even tested and shown to halt glycoprotein bonding to the CD4 receptor. Our qualitative data from ProFunc and RaptorX also indicate that these changes have produced different structures more similar to neutralized or non-infective versions of the protein.

The next step is to apply what we have learned here and develop ligands that target these sites and test their effectiveness *in vitro* and *in vivo* in cultured cells to truly determine whether they are effective drug targets. However, to be certain before laboratory testing and to re-assess these sites, we should run the modified proteins through simulation or molecular dynamics software to predict more accurately the results and effects on the *function*

of a protein, not just its structure. In the lab, site-directed mutagenesis can also be performed to alter the genome of HIV-1 and substitute the amino acids in the identified sites. Nonetheless, we believe this research has revealed certain amino acid interactions that show great potential in becoming targets for future antiretrovirals.

5 Conclusion

Our computational method identified three sites as potential drug targets in gp120 that include amino acids that, as past research indicates, are crucial to glycoprotein binding with CD4 receptors on T-cells: the aspartic acid-457 and arginine-469 salt bridge on chain I, the aspartic acid-279 and lysine-282 salt bridge on chain H, and the cysteine-228 and cysteine-239 disulfide bond on chain E.

These results lend support to the validity of other sites previously unconsidered in past research including the salt bridge between glutamate-28 and lysine-25 on chain B of the capsid. We particularly emphasize the unconsidered sites that are also the two best results according to statistical data suggesting sizable structural dissimilarity: the salt bridge between aspartate-477 and arginine-480 on chain I of gp120 (21.8 percent difference in Z-scores) and the salt bridge between aspartate-152 and arginine-154 on chain A of the capsid (20.0 percent), which is conveniently part of the major homology region. Antiretrovirals targeting these sites may be more effective and versatile than the drugs available today; these antiretrovirals may become more affordable options that are easier to administer and distribute than a multi-drug "cocktail."

In addition to attempting to find good candidates for drug targets, the project was also designed as an exploration into free and public bioinformatics resources, and we feel we can say that our project is a testament to the advancement and spread of widely available bioinformatics tools today. We open the door to comments and criticism and are fully aware that the computational method we used is rough and imperfect. Yet, at the same time, it relies on well-tested servers and programs. The consistency of some of the results with past research also supports the method.

Indeed, the computational method requires refinement and further testing of its accuracy using more accurate predictors of function, such as molecular dynamics tools. The viability of the predicted sites should also be tested in

the laboratory, possibly with site-directed mutagenesis.

Otherwise, the next step is the development and testing of antiretrovirals for these sites that may lead to promising pharmaceuticals.

6 Acknowledgements

We would like to thank several individuals who, with their seemingly infinite patience, have provided us with guidance and advice throughout our project. Without them, our project would not be as successful as it is.

We would like to thank Eric Thiel, our dedicated AP Biology teacher and mentor in this project, for his consistent support and regular discussions during the course of our research. We appreciate his quick grasp of our research ideas and his broad knowledge of life sciences, as well as his reassurance and encouragement.

We extend our thanks to Heather Pereira, a chemistry teacher who reviewed our research paper and was supportive and encouraging during the course of the project. We are grateful for her allowing us to use her classroom as a place for us to think and work.

Thanks also goes to Robert Collins, a Research Associate at the Biomechanics Lab in UC Berkeley, for his extensive help in molecular analysis tools, including VMD and its salt bridges plugin.

And finally, we thank our parents for their unwavering support of our scientific endeavors, and for their tolerance of the time we spent working on this project.

References

[1] *AIDS Pathology.* (2011). Retrieved from `http://library.med.utah.edu/WebPath/TUTORIAL/AIDS/AIDS.html`

[2] Berger, E.A. (1998). And the Best Picture is — the HIV gp120 envelope, please!, *Nature Structural Biology, 5*, 671–674.

[3] Berthet-Colominas, C., Monaco, S., Novelli, A., Siba, G., Mallet, F., Cusack, S. (1999 Mar 1). Head-to-tail dimers and interdomain flexibility revealed by the crystal structure of HIV-1 capsid protein (p24) complexed with a monoclonal antibody Fab. *EMBO J., 18*(5), 1124–1136.

[4] Chan, D.C., Kim, P.S. (1998). HIV Entry and Its Inhibition. *Cell, 93*(5), 681–684. doi: 10.1016/S0092-8674(00)81430-0

[5] Chang, Y.F., Wang, S.M., Huang, K.J., Wang, C.T. (2007 Jul 13). Mutations in capsid major homology region affect assembly and membrane affinity of HIV-1 Gag. *J Mol Biol., 370*(3), 585–597.

[6] PDB ID: 2HMI Ding, J., Das, K., Hsiou, Y., Sarafianos, S.G., Clark Jr., A.D., Jacobo-Molina, A., Tantillo, C., Hughes, S.H., Arnold, E. (1998). Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 A resolution. *J.Mol.Biol., 284*, 1095–1111.

[7] PDB ID: 3NTE Du, S., Betts, L., Yang, R., Shi, H., Ahn, J., Concel, J., Zhang, P., Aiken, C., Yeh, J.I. Insights into Inhibition of HIV-1 Capsid Assembly from the Crystal Structure of the Wild-type Full-Length Capsid Protein.

[8] Frishman, D., Argos, P. (1995). Knowledge-based secondary structure assignment. *Proteins: structure, function and genetics, 23*, 566–579.

[9] Fromme, B.J., Coetsee, M., Van Der Watt, P., Chan, M.C., Sperling, K.M., Katz, A.A., Flanagan, C.A. (2008 Dec). High-affinity binding of southern African HIV type 1 subtype C envelope protein, gp120, to the CCR5 coreceptor. *AIDS Res Hum Retroviruses, 24*(12), 1527–1536.

[10] Huang, B. (2009). metaPocket: a meta approach to improve protein ligand binding site prediction. *Omics, 13*(4), 325–330.

[11] Humphrey, W., Dalke, A. and Schulten, K. (1996). VMD - Visual Molecular Dynamics. *J. Molec. Graphics, 14*, 33–38.

[12] Kaiser, G.E. (2011). *Animal Virus Life Cycles: The Life Cycle of HIV.* Retrieved from `http://faculty.ccbcmd.edu/courses/bio141/lecguide/unit3/viruses/hivlc.html`

[13] Klatt, E.C. (2011). *Pathology of AIDS: Version 22.* Retrieved from `http://library.med.utah.edu/WebPath/AIDS2011.PDF`

[14] Kwong, P.D., Wyatt, R., Majeed, S., Robinson, J., Sweet, R.W., Sodroski, J., Hendrickson, W.A. (2000 Dec 15). Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure, 8*(12), 1329–1339.

[15] Laskowski, R.A., Watson, J.D., Thornton, J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res., 33*, 89–93.

[16] Laskowski, R.A., Watson, J.D., Thornton, J.M. (2005). Protein function prediction using local 3D templates. *J. Mol. Biol., 351*, 614–626.

[17] PDB ID: 3DNN Liu, J., Bartesaghi, A., Borgnia, M.J., Sapiro, G., Subramaniam, S. (2008). Molecular architecture of native HIV-1 gp120 trimers. *Nature, 455*, 109–113.

[18] PDB ID: 1I5Y Liu, J., Lu, M. HIV-1 gp41 core.

[19] Mammano, F., Ohagen, A., Hglund, S., Gttlinger, H.G. (1994 Aug). Role of the major homology region of human immunodeficiency virus type 1 in virion morphogenesis. *J Virol., 68*(8), 4927–4936.

[20] Marchler-Bauer A, Bryant SH (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res., 32*, 327–331.

[21] Marchler-Bauer A. et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins., *Nucleic Acids Res., 39*, 225–229.

[22] Marchler-Bauer A. et al. (2009). CDD: specific functional annotation with the Conserved Domain Database., *Nucleic Acids Res, 37*, 205–210.

[23] Ophardt, C. (2003). *Tertiary Protein.* Retrieved from `http://www.elmhurst.edu/~chm/vchembook/567tertprotein.html`

[24] Peng, J., Xu, J. (2011). A multiple-template approach to protein threading. *PROTEINS*.

[25] Peng, J., Xu, J. (2009). Boosting protein threading accuracy. In the Proceedings of the 13th International Conference on Research in Computational Molecular Biology (RECOMB), Lecture Notes in Computer Science. *Springer, 5541*, 31–45.

[26] Peng, J., Xu, J. (2010). Low-homology protein threading. *Bioinformatics (Proceedings of ISMB 2010).*

[27] Peng, J., Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. *PROTEINS.*

[28] Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., Sharan, R. (2011 Feb). Combining drug and gene similarity measures for drug-target elucidation. *J Comput Biol., 18*(2), 133–145.

[29] Prlic, A., Bliven, S., Rose, P.W., Bluhm, W.F., Bizon, C., Godzik, A., Bourne, P.E. (2010). Precalculated Protein Structure Alignments at the RCSB PDB website. *Bioinformatics.* doi: 10.1093/bioinformatics/btq572

[30] Roy, A., Kucukural, A., Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc., 5*(4), 725–738.

[31] Sarafianos, S.G., Clark Jr., A.D., Das, K., Tuske, S., Birktoft, J.J., Ilankumaran, P., Ramesha, A.R., Sayer, J.M., Jerina, D.M., Boyer, P.L., Hughes, S.H., Arnold, E. (2002 Dec 2). Structures of HIV-1 reverse transcriptase with pre- and post-translocation AZTMP-terminated DNA. *EMBO J., 21*(23), 6614–6624.

[32] Shin, R., Tzou, Y.M., Krishna, N.R. (2011 Nov 8). Structure of a monomeric mutant of the HIV-1 capsid protein. *Biochemistry, 50*(44), 9457–9467.

[33] Shindyalov, I.N., Bourne, P.E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering, 11*(9), 739–747.

[34] Tina, K.G., Bhadra, R., Srinivasan, N. (2007 Jul). PIC: Protein Interactions Calculator. *Nucleic Acids Res., 35*, 473–476.

[35] Veljkovic, N., Branch, D.R., Metlas, R., Prljic, J., Vlahovicek, K., Pongor, S., Veljkovic, V. (2003 Oct). Design of peptide mimetics of HIV-1 gp120 for prevention and therapy of HIV disease. *J. Pept Res., 62*(4), 158–166.

[36] von Schwedler, U.K., Stemmler, T.L., Klishko, V.Y., Li, S., Albertine, K.H., Davis, D.R., Sundquist, W.I. (1998 Mar 16). Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO J., 17*(6), 1555–1568.

[37] Wyatt, R., Sodroski, J. (1998). The HIV-1 Envelope Glycoproteins: Fusogens, Antigens, and Immunogens. *Science, 280*(5371), 1884–1888. doi: 10.1126/science.280.5371.1884

[38] Yamashita, M., Perez, O., Hope, T.J., Emerman, M. (2007 Oct 26). Evidence for direct involvement of the capsid protein in HIV infection of nondividing cells. *PLoS Pathog., 3*(10), 1502–1510.

[39] Zhao, F., Peng, J., Xu, J. (2010). Fragment-free approach to protein folding using conditional neural fields. *Bioinformatics (Proceedings of ISMB 2010).*

[40] Zhao, F., Peng, J., DeBartolo, J., Freed, K.F., Sosnick, T.R., Xu, J. (2010). A probabilistic and continuous model of protein conformational space for template-free modeling. *J. Comp. Biol.*

[41] Zhou, T., Xu, L., Dey, B., Hessell, A.J., Van Ryk, D., Xiang, S.H., Yang, X., Zhang, M.Y., Zwick, M.B., Arthos, J., Burton, D.R., Dimitrov, D.S., Sodroski, J., Wyatt, R., Nabel, G.J., Kwong, P.D. (2007 Feb 15). Structural definition of a conserved neutralization epitope on HIV-1 gp120. *Nature, 445*(7129), 732–737.