BIF701 Lab 7 Protein Structure Prediction: Rational Drug Design
By Christopher Eeles

Introduction

Despite heavy research investment starting in the 1980s, human immunodeficiency virus (HIV) is still a global pandemic—affecting nearly 37 million people worldwide.[8] The majority of progress to date has been in decreasing morbidity and mortality via the development of anti-retroviral drugs such as AZT.[8] This class of drugs has drastically improved quality of life for people living with HIV: "turning what was once a uniformly fatal disease into a manageable chronic condition for many".[8] The end goal, however, has always been the development of a vaccine, or other long-term treatment, to halt the spread of the disease in populations and/or eliminate infection in individuals already afflicted.[8] Progress towards this goal has been difficult given the rate of evolutionary change endemic to viruses. Effective treatments can be rendered useless by viral adaptation to selective pressures a compound applies, or even by random mutations affecting the structure of the drug's target. As such research is risky, disincentivizing investment by private industry, therefore leaving the bulk of research funding to government agencies and non-profit research charities.[9] In this lab, we will be examining the affect of mutations in the HIV-1 Protease gene on the secondary structure of expressed protein using a variety of web-based protein analysis tools.[1] In doing so we will learn practical approaches to structural proteomics and demonstrate the valuable insights bioinformatics can provide for basic and applied HIV research. Hopefully such information helps elucidate new treatment targets and incentivises the increased funding necessary to meet the UNAIDS FAST-TRACK target for HIV elimination by 2030.[9]

HIV-1 Protease plays a key role in the virulence of HIV in infected hosts.[10] It has therefore been a target for antiretroviral development, leading to creation of the protease inhibitor (PI) class of HIV drugs: *i.e.,* amprenavir and other -navir suffixed pharmaeuticals.[11] The protease is necessary for completion of the HIV life cycle within infected cells, cleaving the long protein chains which characterize the immature virus thus enabling the final infectious version to spread into the surrounding tissue.[10] Without this protein, released HIV would be immature and unable to infect neighboring cells, effectively stopping the exponential growth characteristic of viral infections; this provides the host's innate immune system time to dispose of infected cells, thereby slowing—or even stopping—the spread of the virus.[10] While these drugs have been clinically effective in treating HIV, the specificity of the drug target makes them susceptible to development of resistance in viral populations; this is analogous to antibiotic resistance in bacteria.[11] For this reason these drugs must be used conservatively, and ongoing research is required to bring to market new variants which account for adaptive changes in the structure of HIV-1 protease.[11]

In this analysis we will be examining one such mutation by comparing primary structures between the mutated and wild-type protease. We will determine the characteristics of the amino acid substitutions and infer what affect it may have on the secondary structure of the HIV-1 Protease homodimer.[1] This information will be used to predict the effects this mutation may have on the clinical efficacy of HIV PIs. We will also use homology modelling to predict the secondary structure of the mutated HIV-1 protease from its amino acid sequence and compare it with the actual structure uncovered via X-ray crystallography.[1] If prediction is accurate it may be possible to avoid costly and time intensive crystallography techniques using computer modelling, thereby allowing rapid, low-cost study of HIV-1 protease structural variants. This information would be invaluable for rapid reaction to antiretroviral resistances and, as we accumulate data, may even allow detection of trends in mutation over time thus facilitating predictive development of drugs for mutations which are probable to occur. Additionally, time-series data may reveal conserved sequence areas which could be used to development a more general treatment, or even cure, for HIV. This demonstrates the immense value of bioinformatic analysis within

the field of proteomics and should enable more informed allocation of resources in biomedical research on HIV.

Method

RCSB PDB[3] - https://www.rcsb.org/

The Research Collaboratory for Structural Bioinformatics' (RBSB) protein data bank (PDB) was used to search '1KJF', the protein ID of the substrate bound HIV-1 Protease sequence. This database was originally constructed in the 1980s to serve as an archive for biological macromolecular crystal structures. It has since been updated to include structures determined via nuclear magnetic resonance imaging (NMR) and implemented as a web-based service freely accessible to both researchers in the field and the general public. From it, a wealth of information about proteins sequence, primary, secondary and tertiary structures can be gathered to facilitate comprehensive structural analysis for proteomics applications. The results from the search were used to get the PDB file for the sequence of native HIV-1 Protease with a substrate for input into other tools. A second search was conducted for native HIV-1 Protease '10DW' without substrate and a second PDB file was downloaded for comparison to the '1KJF' PDB using SWISS-MODEL. Additionally, the "Sequence Chain View" of '1KJF' was captured in Fig. 1 for comparison to the PSIPRED predicted secondary structure of mutated HIV-1 Protease in subsequent analyses.

FirstGlance in Jmol[4] - https://www.bioinformatics.org/firstglance/fgij/

Jmol is software develop in the java language for molecular modelling chemical structures in 3D.[12] It is available as an OS application, a JavaApplet for integration into other java based apps, and in a JavaScript only versions using HTML5 to run on computers without java.[12] FirstGlance in Jmol is a browser based implementation of the Jmol software which is specialized for quick, widely accessible 3D visualisations of proteins, DNA, RNA from their PDB identification code. The FirstGlance in Jmol web-app is integrated with the RCSB-PDB to directly fetch the PDB file of a given sequence with its protein ID. In this lab it was used to explore the secondary and tertiary structure of the mutated HIV-1 Protease, examine regions of polar vs non-polar amino acid residues, as well as mapping H2O interactions with moieties of the protein. Two angles of the polar/hydrophobic view were recorded in Fig 2. and Fig 3, as well as view similar to what is often seen in the literature for easy orientation.[14]

SWISS-MODEL[5] - https://swissmodel.expasy.org/

SWISS-MODEL is a homology modelling tool which uses a web-server to automate the modelling workflow for use by researchers lacking specific computational expertise. It allows input of a variety of sequence file types (FASTA, Clustal, PDB, etc.) to enable homology modelling of individual protein targets, protein targets and their templates, or target-template alignments for visualization and interpretation. As homology modelling is currently the most accurate method for creating 3D protein structure models, SWISS-MODEL's accessibility and ease of use allows delivery of complex computation modelling tools to whomever may required it. In this lab we used SWISS-MODEL to model the target mutated HIV-1 Protease based on the known structure of the native HIV-1 Protease. A FASTA file of the mutant HIV-1 Protease sequence was copied as the target and the PDB file of the native HIV-1 Protease from RCSB-PDB was uploaded as the template. The output for the model of mutated HIV-1 Protease was captured in Fig. 5 and the PDB file was downloaded for use in PDBeFold.

BIF701 Lab 7 Protein Structure Prediction: Rational Drug Design
By Christopher Eeles

PDBeFold[6] - http://www.ebi.ac.uk/msd-srv/ssm/

Protein Data Bank in Europe's Fold tool (PDBeFold) allows pairwise or multiple comparison and 3D alignments of protein structures. Unlike sequence alignment, the algorithm compares the geometric location of amino-acid residues between two sequences and therefore residue type is irrelevant. In this lab we used the PDB file from SWISS-MODEL as first source as a coordinate input and the protein ID ' 1KJF' for the second source as a PDB entry. The search yielded a single result, from which we selected 'view superposed' with the 'superpose whole entries option' selected. The resulting 3D model was colourized for easy comparison between the two structures. The wild-type was coloured red, the mutated protein blue, and the substrate white using the console of the Jmol viewer. The model was first rotated to match the 'literature view' captured in previous tools and recorded in Fig. 6, it was then rotated to highlight the three regions where relevant differences between the two sequence occur and recorded in Fig. 7.

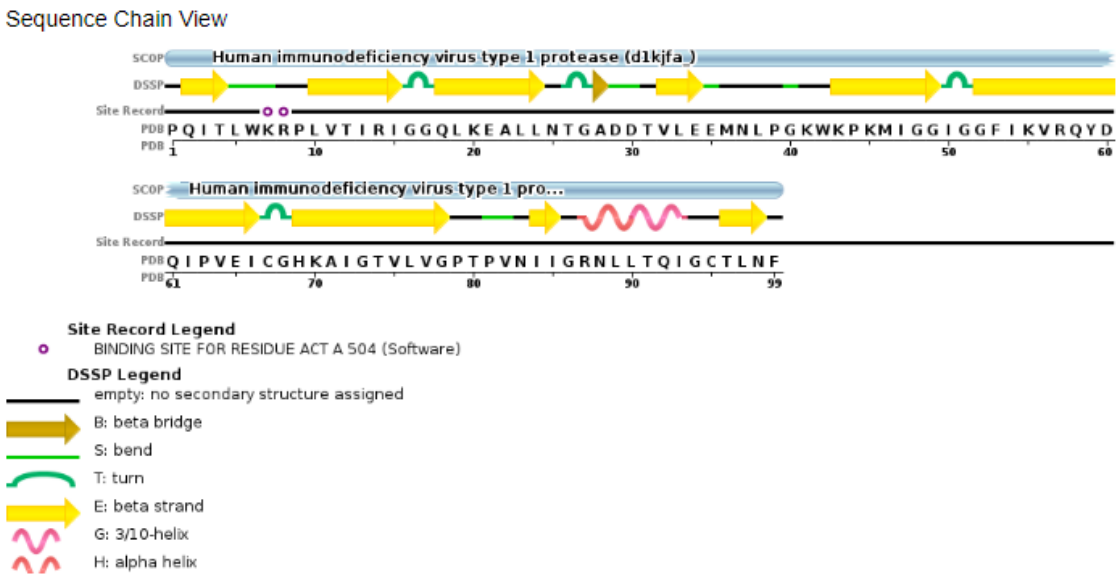
PSIPRED[7] - http://bioinf.cs.ucl.ac.uk/psipred/

PSI-blast based secondary structure PREDiction (PSIPRED) protein sequence analysis workbench is an online tool used to investigate protein secondary structures.[13] It is used for *ab initio* prediction of secondary structures form primary protein sequence data by applying different machine learning based algorithms. The at this time the web service is able to predict alpha helixes, beta sheets, and coils but will likely be improved over time to yield more complex predictions and better accuracy. In this lab we will be using the PSIPRED v3.3 model to predict the secondary structure of HIV-1 Protease '1KJF' for comparison to the known structure recorded in Fig. 1. Analysis will evaluate the accuracy of PSIPREDs prediction with hopes of drawing conclusions about the current state of structure prediction, areas for future improvement, and the implications these limitations have on its applications in field of biomedical research.

Results

RCSB PDB[3]

Figure 1 **– Experimentally Determined Secondary Structure of HIV-1 Protease 'IKJF'**



BIF701 Lab 7 Protein Structure Prediction: Rational Drug Design
By Christopher Eeles

FirstGlance in Jmol[4]

Figure 2 **– Channel to HIV-1 Protease Active Site (pink = polar; grey = hydrophobic; red = area of interest)**

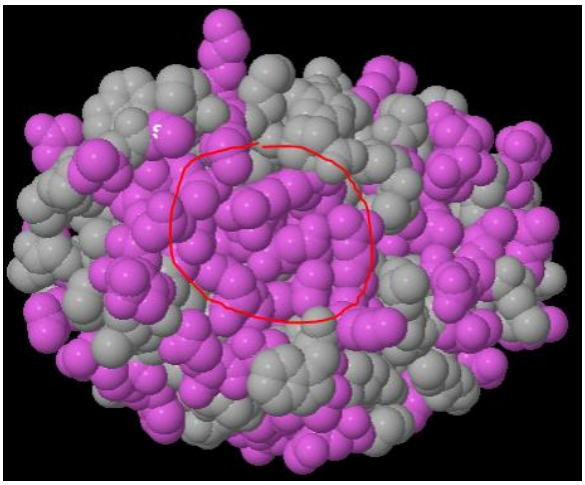


Figure 3 **- HIV-1 Protease Hydrophobic Region (pink = polar; grey = hydrophobic; red = area of interest)**

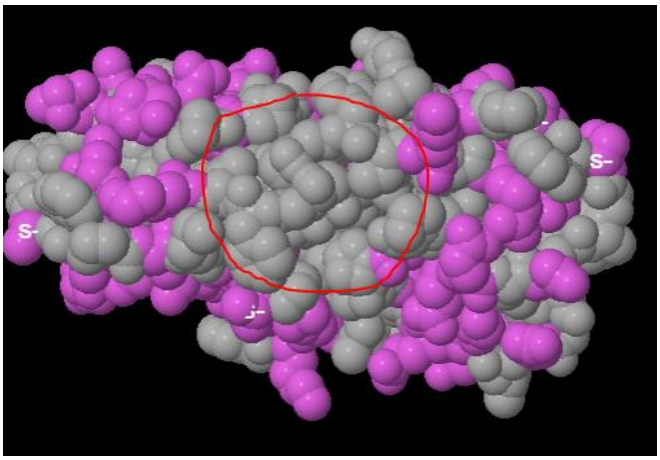
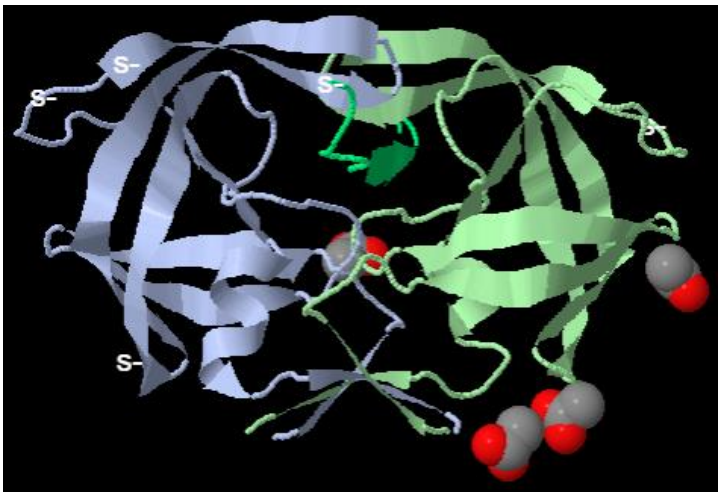


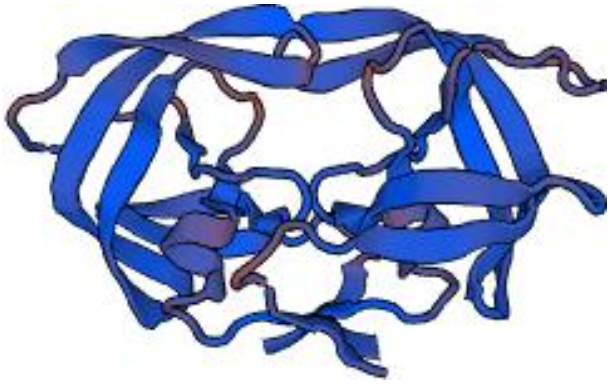
Figure 4 **- HIV-1 Protease '1KJF' Homodimer Oriented for Protein Channel View**



BIF701 Lab 7 Protein Structure Prediction: Rational Drug Design
By Christopher Eeles

SWISS-MODEL[5]

Figure 5 - Predicted 3D Structure of Mutated HIV-1 Protease



PDBeFold[6]

Figure 6 **- Mutated HIV-1 Protease (blue) vs Native HIV-1 Protease + Substrate (red)**

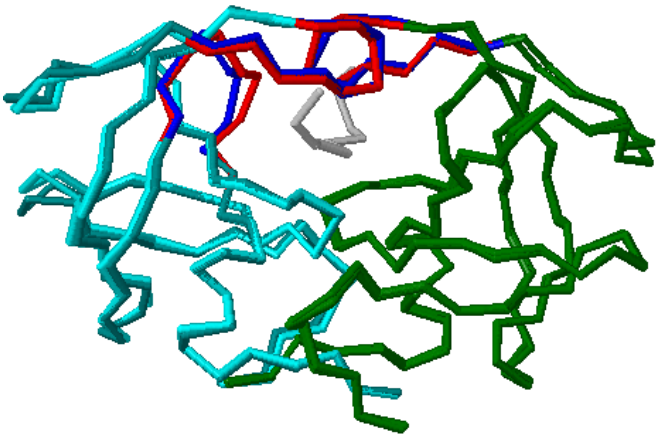


Figure 7 **- Mutated vs Native HIV-1 Protease Modified Regions (A = top, B = left, C = right)**

BIF701 Lab 7 Protein Structure Prediction: Rational Drug Design
By Christopher Eeles

PSIPRED[7]

Figure 8 **– Predicted Secondary Structure for Mutated HIV-1 Protease**

Discussion/Conclusion

Visualisation of HIV-1 Protease '1KJF' with FirstGlance in Jmol demonstrated three major protein areas. Firstly, the channel leading to the active site was lined with polar amino-acids as shown in Fig. 1. Considering this channel needs to allow entry of aqueous protein substrates into the enzyme's catalytic core, the polar nature is as expected. The second region, displayed in Fig.2, is a region of concentrated hydrophobic residues in the "flap" portion of each dimer protein (the strands above the substrate in Fig. 4).[14] These "flaps" play a role in enzyme selectivity by changing conformation to open the channel in response to the proximity of substrate; their hydrophobic residues are therefore required to ensure that only the desired ligands enter the channel and bind the active site.[14] The remaining exterior of the protein was a combination of polar and non-polar amino acids which likely optimize the relative solubility of the complex in the cytosol at its preferred operating temperature (likely 37 C). There were charged amino acids spread throughout the sequence, mostly a positive charge aligned with a negative charge in the tertiary structure. These features likely play a role in stabilizing the 3D configuration and, along with disulfide and hydrogen bonds, ensure the desired conformation is achieved during protein folding. The

active site of the protein is characterized by an arginine-threonine-glycine triad starting at position 25. This sequence was modified to facilitate crystallization and in the functional protein will be the charged amino acid aspartic acid, the polar amino acid threonine, and the non-polar amino acid glycine which are characteristic of the catalytic triad found in various aspartic proteases.[15] Interestingly, the active site appears unchanged between the mutated and non-mutated versions and therefore is insufficient to directly account for the difference in PI effectives between the proteins.

Homology modelling of the mutated protein complex in SWISS-MODEL produced the structure captured Fig. 5. Comparing it to the substrate bound native HIV-1 Protease in Fig. 4, it is apparent that these two molecules are nearly identical. Closer examination using PDBeFold to overlay the two 3D models in Fig. 6 supports this conclusion, highlighting three regions bordering the substrate channel for closer examination. In Fig. 7, we have rotated the model to more clearly view these differences, the wild-type being displayed in red and the mutated version in blue. While quantitative analysis was not conducted, the structural variation between the mutated and wild-type complexes appear to be minor—the mutated channel possibly being slightly wider. Given these observations it seems the drug resistant HIV-1 protease has adaptively opened its channel to match or exceed that of the substrate bound wild-type. Considering that -navir class PIs act as reversible, competitive inhibitors of substrate binding at the active site, such changes could conceivably result in decreased enzyme specificity.[14] Taking a kinetic point of view, a wider channel may allow more random collisions per unit time and thus increase the probability of -navir drugs being displaced from the active site. A less specific channel could also increase the pool of molecular competitors for binding, further augmenting the probability of collisions which displace the drug. If these changes also affect the thermodynamic stability of binding to favour substrate over -navir compounds, this may also contribute to the decreased clinical efficacy observed from these structural changes. While confirming the mechanism for this adaptation will require further study, a significant amount of information was gained without ever entering a wet-lab.

Contrasting the known secondary structure of HIV-1 Protease '1KJF' in Fig.1 with the PSIPRED predicted one in Fig. 2, we immediately observe a major misprediction. Starting at position 20, PSIPRED has forecast an alpha-helix where there is in fact a beta-sheet. An error this significant likely invalidates most inferences which can be drawn about tertiary structure—and therefore about function. Based on these results it seems, at least with current state of the technology, protein structure prediction is not a suitable replacement for NMR or crystallography in structural proteomics. However, the remainder of the prediction is accurate to within a few residues for each respective structure. Therefore, it could serve to direct further lines of inquiry on novel protein sequences, allowing more efficient allocation of scare grant funds and improving the value for money of research investments. With time, and accumulation of data, the algorithms used in PSIPRED and other structure prediction tools will be improved; hopefully one day enabling *ab initio* structure predictions to approach the accuracy of more conventional techniques. Until then, there is much work to be done by bioinformaticians, computer scientists and statisticians alike: the future for the field of structural proteomics looks bright and advances in each field will undoubtedly enable innovate new techniques to be developed in the others.

Considering these results, while the currently available technologies for protein structure prediction may not enable anticipatory pharmaceutical development for forecast mutations, they will certainly prove increasingly valuable in the response to drug resistances developed in HIV and other viral pathogens. Even marginal increases in the rate of drug discovery and development for HIV could have significant impacts of the quality and quantity of life for those affected by this pandemic. The value of improved quality of

life should not be understated, and any gains in life expectancy increase the chances that a cure will be found in the lifetime of a patient. In the mean time, any impediment to the spread of the virus will decrease global prevalence and serve as a step towards eradication. Bioinformatics will be a powerful tool in the journey towards an HIV free world, and even though not every humble researcher can receive acclaim for their work, they are undoubtedly the foundation on which future breakthroughs will be built. Whether by epidemiological elimination through vaccines and other population level interventions, or by discovering a cure or indefinite treatment, the prospects for HIV elimination in our lifetimes are good. Though it is impossible to predict if the global community is able to meet the aggressive targets of UNAIDS for an HIV free world by 2030 will be met, it is most definitely a step in the right direction.

References

1. School of Biological Science and Applied Chemistry. (2018). Lab 7 Protein Structure Prediction: Rational Drug Design. Seneca College: Toronto, ON.
2. School of Biological Science and Applied Chemistry. (2018). Protein Structure Prediction. Seneca College: Toronto, ON.
3. Berman, H.M *et al.* (2000). The Protein Data Bank. In *Nucleic Acids Research*, Volume 28(1) pp. 235 242. https://doi.org/10.1093/nar/28.1.235.
4. Eric Martz. (2016). FirstGlance in Jmol. In *Bioinformatics.org* [Website]. Retrieved from https://www.bioinformatics.org/firstglance/fgij/ .
5. Waterhouse *et al*. (2018). SWISS-MODEL: homology modelling of protein structures and complexes In *Nucleic Acids Res. 46*(1), pp. 296-303. Retrieved from https://swissmodel.expasy.org/.
6. Krissnel, E., Henrick, K. (2018). PDBeFold at European Bioinformatics Institute. In *Protein Data Bank in Europe* [Website]. Retrieved from http://www.ebi.ac.uk/pdbe/docs/Tutorials/workshop_tutorials/PDBefold.pdf
7. UCL Department of Computer Science: Bioinformatics Group. (2018). The PSIPRED Protein Sequence Analysis Workbench. *University of Cambridge London* [Webite]. Retrieved from http://bioinf.cs.ucl.ac.uk/psipred/.
8. National Institute of Allergy and Infectious Disease (NIAID). (2018). HIV/AIDS. *National Institute of Health* [Website]. Retrieved from https://www.niaid.nih.gov/diseases-conditions/hivaids.
9. UCSF Medical Center. (2018). HIV InSite [Website]. University of California San Franscisco (OCSF): Sanfransico, CA, USA. Retrieved from http://hivinsite.ucsf.edu/InSite?page=li-03-02#S2X.
10. U.S. Department of Health and Human Services (HHS). (2018). AIDSinfo. *Natrional Institute of Health* [Website]. Retrieved from https://aidsinfo.nih.gov/understanding-hiv-aids/fact-sheets/19/73/the-hiv-life-cycle.
11. Wikipedia Community. (2018). Protease inhibitor (pharmacology). In *Wikipedia* [Website]. Retrieved from https://en.wikipedia.org/wiki/Protease_inhibitor_(pharmacology).
12. Wikipedia Community. (2018). Jmol. In *Wikpiedia* [Website]. Retrieved from https://en.wikipedia.org/wiki/Jmol.
13. Wikipedia Community. (2018). PSIPRED. In *Wikipedia* [Website]. Retrieved from https://en.wikipedia.org/wiki/PSIPRED.
14. McEvoy, G.K. (2003). American Hospital Formulary Service – Drug Information 2003 (p. 630). American Society of Health-System Pharmacists, Inc: Bethesda, MD.
15. Wikipedia Community. (2018). HIV-1 Protease. In In *Wikpiedia* [Website]. Retrieved from https://en.wikipedia.org/wiki/HIV-1_protease.

BIF701 Lab 7 Protein Structure Prediction: Rational Drug Design
By Christopher Eeles