The Impact of Amino Acid Mutations on Binding in HIV Proteins

CSE 307: Structural Bioinformatics
Joseph D'Onofrio

Introduction

A protein, much like many other biological molecules, begins its life as a string of nucleotides in an organism's DNA. This sequence of nucleotides is translated into codons, each consisting of three adjacent nucleotides. Each codon corresponds to either a signal to start/stop translation or to a specific amino acid. Amino acid chains, known as polypeptides, are created from the string of codons between the start and stop codon. The polypeptides are then released, folding into distinct structures and they either become a protein on their own or join with other polypeptides to create a protein. In this text, the polypeptide(s) which make up a protein will be referred to as a protein's amino acid sequence.

A mutation to the amino acid sequence of any given protein is the result of a change to a single nucleotide in the segment of the DNA sequence which defines that protein. This is known as a point mutation, and causes a codon to form which is different from the codon which would typically be generated from that section of DNA. When this codon defines a different amino acid than is standard, the protein is created with a swapped amino acid in its sequence.

It is important to study amino acid mutations as they can change the function of a protein, block or expose binding sites, etc. One specific area where mutations can have a large impact is the field of pharmaceutics, which is interested in the development of drugs with medicinal properties. A mutation to even a single amino acid can alter a binding site of a protein which is targeted by a drug, increasing/decreasing resistance to that drug. For a person with this mutation, this medicine may function differently, potentially requiring a different level of dosage, or simply not working at all. This poses a problem for pharmaceutical scientists attempting to develop medicine for various illnesses, as they must account for mutations in order for their drugs to have maximum effectiveness in the population. Upon consideration, this issue spawns a valuable question:

> *How can pharmaceutical scientists predict which mutations may resist their drug without resorting to costly and time intensive experimentation?*

This text seeks to provide initial insight into the solution of this problem through a computational analysis of amino acid mutations in the three main proteins targeted by HIV drugs and their associated drug resistance.

Hypothesis

Each unique amino acid has some inherent impact on a protein's ability to bind such that if a mutation to that amino acid appears in a protein then that protein is either more or less likely to bind to the inhibitors of drugs which target it. Furthermore, a factor of this inherent impact on binding is caused by the category (Polar, Hydrophobic, Charged-Positive, or Charged-Negative) of the amino acid mutated to. Finally, when an amino acid mutation results in a change of category, a factor of the inherent impact on binding stems from the specific change of category (Polar to Hydrophobic, etc.) between the original and mutated amino acids.

Data: Information

This computational analysis utilizes the mutation and drug resistance data provided by Stanford University's HIV Drug Resistance Database. A web tool is offered on the Stanford HIV Database website (*See Appendix A) which provides drug resistance data for amino acid mutations to the three main HIV-1 proteins targeted by HIV medicine. These proteins are: HIV-1 Protease, HIV-1 Integrase, and HIV-1 Reverse Transcriptase. For each protein, amino acid mutations have drug resistance scores for each of the HIV drugs which target that protein. This data set was selected for its ease of access to mutation data that is directly correlated to drug resistance data.

Data: Gathering and Manipulation

The Stanford HIV Database (HIVdb) allows for programmatic access to their web tool through the use of the "sierrapy" Python package. Three queries to HIVdb were generated, one for each HIV protein. Each query is given a list of amino acid mutations and a selection of desired attributes. The list of mutations used for each protein consists of all possible mutations at each position in that protein associated with drug resistance. Since data is returned on a by-drug basis, the attributes selected were drug name, mutations, and resistance to the drug by mutation. Queries return as a JSON response containing mutation resistance data for each drug.

The initial results returned by each query must be manipulated to a format where it can first be analyzed by mutation and then by amino acid. This is accomplished by defining two Python objects which store data in the desired format. The Mutation object initially stores the name of a mutation, a list of its resistance scores to each drug, and an integer designating which protein the mutation occurs in. The MutationList object is initialized as a blank list and will store a list of Mutation objects and includes functions to manipulate Mutation objects as a whole for future calculations. For each drug in the JSON result, the mutations under that drug are parsed one at a time. If that mutation is not already in the MutationList, we create a new Mutation object and add the resistance score for the current drug to its list of scores. If a mutation is found that already exists in the MutationList, then its corresponding Mutation object is found and its resistance to the current drug is appended to the list of scores. After parsing the entire JSON result, MutationList contains a list of all mutations for the given protein and the drug resistance scores associated with them. Once the complete MutationList is compiled, the average resistance score is calculated for each mutation using the list of scores stored in each Mutation object. The averages are calculated since the value of drug resistance data to this analysis lies in a mutation's overall drug resistance rather than a mutation's resistance to each drug. One MutationList is created for each of the three HIV proteins.

The last step before binding impact computations can begin is to define an AminoAcid object which will initially store the one letter code of an amino acid and the average drug resistance of all mutations to that amino acid. These objects are created by, for each of the twenty amino acids, iterating through all three MutationLists for and calculating the average drug resistance across every mutation. The average resistance of mutations to each unique amino acid is calculated here since it will provide the basis for the computations on category resistances.

In order to preserve all of the data compiled thus far, a ResistanceData object is defined which initialized with all the Mutation objects (as a list of the three MutationLists) and a list of the twenty AminoAcid objects as parameters. All further computations take place upon the creation of the ResistanceData object. This allows for easy preservation and use of subsequent calculations by storing the values as object attributes.

Computation: Examination

Before moving on to the methods used for computation, it is important to note a few items which are ultimately products of the available data. The main objectives of this analysis all revolve around the notion of generating plus-minus scores which indicate the impact a mutation to a specific amino acid has on binding. The intuitive way to accomplish this given the data, is to identify a point of comparison for each computation that will expose a meaningful value when mutations are compared to it. In this analysis, each point of comparison will be the overall average resistance scores across all mutations in the sub-category being examined. This will allow for the calculation of a plus-minus score for each amino acid that represent the distance of that amino acid's resistance to the mean resistance of comparable mutations in the same sub-category.

Additionally, the assumption is made that higher drug resistance scores correspond to a more negative impact on binding, as resistance to a drug indicates that the drug's inhibitors are not binding to the target protein. Therefore, when a comparison is made to the sub-category mean, a negative score (lower resistance than the mean) is interpreted as a positive impact on binding, and a positive score (higher resistance than the mean) is viewed as a negative impact on binding.

Computation: Methods

The first clause of the hypothesis seeks to discover if there exists a significant difference in the resistance of mutations to a specific amino acid compared to the resistance of any amino acid mutation. The average resistance of all mutations is selected as the point of comparison to test this clause. This average is calculated from the list of all mutations stored in the ResistanceData object and stored as an attribute for the object. The set of results corresponding to the first clause of the hypothesis is computed as the difference between the average resistance of mutations to a given amino acid and the average resistance for all mutations. This value represents the impact a mutation to a specific amino acid has on binding over/under all amino acid mutations.

The second clause of the hypothesis generalizes the first, suggesting that some factor of an amino acids inherent impact on binding is a result of its category (Polar, etc). The first point of comparison selected to test this clause will again be the overall average resistance of all mutations. For each category, the average is computed for all mutations to amino acids in that category. The difference between the category average and the overall average represents the impact a mutation to a specific category of amino acids has on binding over/under all mutations. As an attempt to further expose an amino acid's inherent impact on binding separate from its category, the average resistance for mutations to each category is selected as a second point of comparison. The difference between the average resistance of all mutations to a given amino acid and the average resistance of mutations to its category represents the impact an amino acid has on binding over/under all amino acids in its category.

The final clause of the hypothesis presents the idea that some factor of a mutation's impact on binding stems from the specific change of category (Polar to Hydrophobic, etc.) caused by the mutation. To test this clause, the initial point of comparison is selected as the average resistance of all mutations. For each type of category-category mutation, the difference between the average resistance of mutations that fit that type and the average resistance of all mutations represents the impact mutations with that type of category change have on binding over/under all mutations. The average resistance for each specific type of category change is selected as another point of comparison to again attempt to further expose an amino acids inherent binding impact separate from its category change type. The difference between the average resistance of all mutations to a specific amino acid within the appropriate category change type and the average resistance for all mutations in that category change type (See Example) represents the impact a mutation to an amino acid has on binding over/under all mutations in its category change type. ***Example***: The average resistance of all mutations to Histidine (Polar) from a Hydrophobic amino acid – the average resistance of all Hydrophobic to Polar mutations = Histidine's impact on binding as compared to all other Hydrophobic to Polar mutations.

Results: Presentation

*Notes: The results presented here consist only of binding impact values. Run Python script to dump all data to console.
Negative values indicate positive impact on binding (+), Positive values indicate a negative impact on binding (-).

Hypothesis Clause 1: The impact on binding of mutations to a specific amino acid as compared to all mutations.

Overall Average Resistance: 11.139

Asparagine (N): -8.034 (+)
Threonine (T): -7.03 (+)
Tryptophan (W): -6.594 (+)
Glutamic Acid (E): -6.242 (+)
Aspartic Acid (D): -5.941 (+)
Phenylalanine(F): -5.029 (+)
Glycine (G): -3.64 (+)
Isoleucine (I): -2.68 (+)
Leucine (L): -1.322 (+)
Cysteine (C): -1.253 (+)
Valine (V): -0.501 (+)
Serine (S): -0.098 (+)
Lysine (K): 0.931 (-)
Arginine (R): 1.739 (-)
Methionine (M): 3.732 (-)
Tyrosine (Y): 4.997 (-)
Alanine (A): 5.166 (-)
Proline (P): 6.758 (-)
Glutamine (Q): 8.86 (-)
Histidine (H): 14.466 (-)

Hypothesis Clause 2a: The impact on binding of mutations to a specific category as compared to all mutations.

Overall Average Resistance: 11.139

Polar: 1.005 (-)
Hydrophobic: -0.178 (+)
Charged-Positive: 1.335 (-)
Charged-Negative: -6.091 (+)

Hypothesis Clause 2b: The impact on binding of mutations to a specific amino acid as compared to all mutations to that category.

Any to Polar:

Average Resistance: 12.145

Asparagine (N): -9.039 (+)
Threonine (T): -8.035 (+)
Tryptophan (W): -7.599 (+)
Cysteine (C): -2.259 (+)
Serine (S): -1.103 (+)
Methionine (M): 2.727 (-)
Tyrosine (Y): 3.991 (-)
Glutamine (Q): 7.855 (-)
Histidine (H): 13.461 (-)

Any to Hydrophobic:

Average Resistance: 10.962

Phenylalanine(F): -4.851 (+)
Glycine (G): -3.462 (+)
Isoleucine (I): -2.501 (+)
Leucine (L): -1.144 (+)
Valine (V): -0.322 (+)
Alanine (A): 5.344 (-)
Proline (P): 6.936 (-)

Any to Charged-Positive:

Average Resistance: 12.475

Lysine (K): -0.404 (+)
Arginine (R): 0.404 (-)

Any to Charged-Negative:

Average Resistance: 5.048

Glutamic Acid (E): -0.151 (+)
Aspartic Acid (D): 0.151 (-)

Hypothesis Clause 3a: The impact on binding of mutations with specific category change as compared to all mutations.

Overall Average Resistance: 11.139

Polar to Polar 8.1 (-)
Hydrophobic to Hydrophobic 0.33 (-)
Charged-Positive to Charged-Positive 5.072 (-)
Charged-Negative to Charged-Negative -9.776 (+)
Polar to Hydrophobic -0.672 (+)
Polar to Charged-Positive 1.777 (-)
Polar to Charged-Negative -5.387 (+)
Hydrophobic to Polar 1.482 (-)
Hydrophobic to Charged-Positive -4.473 (+)
Hydrophobic to Charged-Negative -5.912 (+)
Charged-Positive to Polar -6.674 (+)
Charged-Positive to Hydrophobic 10.678 (-)
Charged-Positive to Charged-Negative -5.912 (+)
Charged-Negative to Polar 1.417 (-)
Charged-Negative to Hydrophobic -5.786 (+)
Charged-Negative to Charged-Positive -7.049 (+)

Hypothesis Clause 3b: The impact on binding of mutations to a given amino acid within a specific category change type as compared to all mutations of that category change type.

Polar to Polar:

Average Resistance: 19.240056818181817

Histidine (H): 22.427 (-)
Serine (S): -9.865 (+)
Tyrosine (Y): -11.134 (+)
Cysteine (C): -9.695 (+)
Methionine (M): 6.669 (-)

Hydrophobic to Hydrophobic:

Average Resistance: 11.46933040078201

Alanine (A): 6.571 (-)
Isoleucine (I): -2.823 (+)
Leucine (L): -0.087 (+)
Phenylalanine(F): -6.651 (+)
Valine (V): -0.83 (+)
Glycine (G): -5.106 (+)

Charged-Positive to Charged-Positive:

Average Resistance: 16.21212121212121

Lysine (K): 3.788 (-)
Arginine (R): -1.894 (+)

Charged-Negative to Charged-Negative:

Average Resistance: 1.3636363636363635

Glutamic Acid (E):0.0

Polar to Hydrophobic:

Average Resistance: 10.46750398724083

Alanine (A): 6.755 (-)
Isoleucine (I): -2.286 (+)
Leucine (L): -2.996 (+)
Phenylalanine(F): -2.74 (+)
Proline (P): 5.47 (-)
Glycine (G): 0.442 (-)

Polar to Charged-Positive:

Average Resistance: 12.916666666666666

Lysine (K): 0.972 (-)
Arginine (R): -2.917 (+)

Polar to Charged-Negative:

Average Resistance: 5.752840909090909

Aspartic Acid (D): 0.876 (-)
Glutamic Acid (E): -2.628 (+)

Hydrophobic to Polar:

Average Resistance: 12.622245179063363

Histidine (H): -4.44 (+)
Serine (S): -0.747 (+)
Tyrosine (Y): 11.544 (-)
Cysteine (C): -2.395 (+)
Methionine (M): -1.429 (+)
Tryptophan (W): -8.077 (+)

Hydrophobic to Charged-Positive:

Average Resistance: 6.666666666666667

Lysine (K): 0.0

Hydrophobic to Charged-Negative:

Average Resistance: 5.227272727272727

Aspartic Acid (D): -4.318 (+)
Glutamic acid (E):4.318 (-)

Charged-Positive to Polar:

Average Resistance: 4.465909090909091

Asparagine (N): -5.375 (+)
Histidine (H): 6.443 (-)
Threonine (T): -0.356 (+)

Charged-Positive to Hydrophobic:

Average Resistance: 21.818181818181817

Proline (P): 0.0

Charged-Positive to Charged-Negative:

Average Resistance: 5.227272727272727

Glutamic acid (E):0.0

Charged-Negative to Polar:

Average Resistance: 12.556818181818182

Glutamine (Q): 7.443 (-)
Asparagine (N): -7.443 (+)

Charged-Negative to Hydrophobic:

Average Resistance: 5.353535353535354

Alanine (A): 1.768 (-)
Glycine (G): -3.535 (+)

Charged-Negative to Charged-Positive:

Average Resistance: 4.090909090909091

Lysine (K): 0.0

Results: Analysis

Hypothesis Clause 1: Each unique amino acid has some inherent impact on a protein's ability to bind such that if a mutation to that amino acid appears in a protein then that protein is either more or less likely to bind to the inhibitors of drugs which target it.

The result set which investigates this clause seeks to show that there does exist a significant difference between the drug resistances of mutations to each amino acid. It displays the impact on binding of mutations to a specific amino acid as compared to all mutations. Looking at the data, it is clear that the binding impacts of amino acid mutations occupy a fairly wide range of value, with mutations to Histidine or Glutamine being highly resistant and mutations to Asparagine or Threonine being much less resistant. This appears to validate this part of the hypothesis; the range of resistances is too large to be explained by variation in the data, thus amino acids must have some attribute that differs between amino acids and impacts binding ability. The following clauses explore the idea that this attribute could be in part related to the category of amino acid mutated to.

Hypothesis Clause 2: a factor of this inherent impact on binding is caused by the category (Polar, Hydrophobic, Charged-Positive, or Charged-Negative) of the amino acid mutated to.

The result set corresponding to clause 2a was compiled in an attempt to explicitly verify this statement by showing there is a significant difference between the resistance data of mutations to different categories. It displays the impact on binding of mutations to a specific category as compared to all mutations. For mutations to Polar, Hydrophobic, and Charged-Positive amino acids, there is little deviation from the mean resistance. For mutations to Charged-Negative amino acids there is a significant drop in resistance. The two Charged-Negative amino acids, Glutamic Acid and Aspartic Acid, were both among the least resistant amino acids in the first set of data, so it is not a surprise for their category average resistance to be well below the mean. To check this was not an error, all

mutations to Charged-Negative amino acids were examined and all had low resistance scores. However, the data for the other categories seems to be inconclusive on this clause. A larger data set may unveil differences between the other categories, but for this data it seems the only category with significant impact on binding is Charged-Negative.

The result set corresponding to clause 2b was compiled to attempt to isolate the differences between each amino acid within a category in the case that the average category resistance was not a significant indicator of a category-based impact on binding. It displays the impact on binding of mutations to a specific amino acid as compared to all mutations to that category. The data for mutations to Polar and Hydrophobic amino acids is very similar to the data from result set 1, since the Polar and Hydrophobic average resistances are close to the mean. The binding impact values for amino acids in these two categories support the analysis of set 1 in that some attribute differentiates amino acids in terms of drug resistance, though now it can be said for Polar and Hydrophobic amino acids, this attribute is likely not category. For both Charged Positive and Charged Negative, the individual amino acids within each category are very close to the mean. This could indicate that the charge of an amino acid mutation is a factor in drug resistance. Comparing the two uncharged categories to the charged categories yields some interesting observations. Mutations to charged amino acids appear to have resistances which closely surround a mean resistance dependent on the charge, with positive charge indicating high resistance and negative charge indicating low resistance. Mutations to uncharged amino acid result in a resistance corresponding to the specific amino acid rather than the category. Factoring in these observations, there is a somewhat strong argument for the verification of clause 2 in that mutations to charged amino acids appear to have a category-based impact on binding stemming from the charge of the specific amino acid, while mutations to uncharged amino acid categories do not seem to have any category-based impact on binding.

Hypothesis Clause 3: When an amino acid mutation results in a change of category, a factor of the inherent impact on binding stems from the specific change of category (Polar to Hydrophobic, etc.) between the original and mutated amino acids.

The first result set which examines the validity of this clause attempts to show that there exists a significant difference between the average resistances of each category change type and the average resistance for all mutations. It displays the impact on binding of all mutations with a specific category change type as compared to all mutations. It can be viewed as an expansion of result set 2b, accounting for the category of the original amino acid as well as the mutated amino acid. The expanded data successfully shows that within the average resistances of all mutations to a certain category, there exists a wider spread of resistances that may be the result of the category of the original amino acid. Building off the previous charged-uncharged observations, it should be noted that the drug resistance values for category change types that correspond to mutations between uncharged and charged amino acids appear to support the earlier observations. For changes from uncharged to charged amino acids, the resistance values approach the mean for the given charged category. For mutations from Polar to Charged-Positive, the average resistance value (1.777 above the overall mean of 11.139 = 12.916) of closely mirrors the mean resistance of all mutations to Charged-Positive amino acids (12.475). For mutations from Polar to Charged-Negative, the average resistance value (5.387 below the overall mean of 11.139 = 5.752) of closely mirrors the mean resistance of all mutations to Charged-Negative amino acids (5.048). This seems to support the prior observation that mutations to charged amino acids tend to approach the category mean based on charge. It should be noted that, while the resistance value for Hydrophobic to Charged-Negative also supports the observation, the value for Hydrophobic to Charged-Positive does not support it. For mutations in the other direction, from charged to uncharged, all of the resistance values seem to support the observation that mutations to uncharged amino acids do not congregate around any category-based means. Based on these extended observations, it appears as if another strong argument could be made for the validity of clause 3 as mutations which swap charged and uncharged follow the previously stated category-based trends.

Result set 3b seeks to break down each category change type by mutation to amino acid within that type to expose the differences between mutations to different amino acids within the same category change type. Unfortunately, this data doesn't seem to offer any new insights into either the charged-uncharged relationship or the cause differences between resistances for mutations to uncharged. Part of this may be because as a result of a limited data

set, breaking down mutations into categories this specific yields very small sample sizes, with a few categories that have only one mutation to pull data from (the categories with amino acids that have binding impacts of 0). Another part of the absence of new insight could be simply because this result set is an expansion of data sets which have already had all useful information gleaned from them. Without a concrete idea of what may cause the resistance variation in uncharged amino acids.

Conclusions

After thorough analysis of the results, a few conclusions can be drawn. One conclusion is that the first hypothesis clause is correct, each amino acid has some inherent impact on binding that differs from other amino acids. This is shown by the variation in the compiled drug resistance averages for mutations to each amino acid. Another conclusion that can be drawn is that mutations to charged amino acids have some level of category-based impact on binding, which partially validates both the second and third hypothesis clauses. This conclusion fails to fully validate the clauses because the clauses hold up only for mutations to charged amino acids. It is incorrect to say that some factor of an amino acids inherent impact is caused by its category, since the results indicate that mutations to uncharged amino acids have no determinable category-based impact on binding. The final conclusion that can be made is that each uncharged amino acid has some attribute which causes a significant difference in drug resistance across all amino acids. This conclusion can only be made confidently for uncharged amino acids since mutations to charged amino acids appear have low resistance variation and some level of category-based drug resistance, thus their resistance scores can't be reliably attributed to some non-category-based attribute. In order to attempt to determine this resistance attribute further analysis would be required which consider new factors such as physical structure, etc.

Impact

The conclusions drawn from this analysis can be used to generate some preliminary insight into the solution of the initial question posed in this text:

> *How can pharmaceutical scientists predict which mutations may resist their drug without resorting to costly and time intensive experimentation?*

While these results are far from becoming any practical application in pharmaceutics, the results of this analysis can be interpreted as a validation that a solution to this problem does exist. Factors which impact drug resistance in amino acid mutations can be isolated computationally, and once enough of the factors are uncovered, they could form the basis for reliable drug resistance predictions for amino acid mutations. Knowledge of how a certain mutation would resist a drug combined with a study of the frequency mutation in a protein associated with disease would allow pharmaceutical scientists to quickly create effective medicines for very serious diseases.

Appendix A: Tools

Stanford HIV Database*(https://hivdb.stanford.edu/)*:

> Main page of the Stanford HIV Database. Offers many different tools and datasets related to HIV drugs, resistances, mutations, etc.

Stanford HIVdb Program*(https://hivdb.stanford.edu/hivdb/by-mutations/)*:

> Web tool offered by Stanford HIV Database which provides the data used in this analysis. Can be accessed on the web using the above link or programmatically using the "sierrapy" Python package.

HIVdb Mutations Analysis Example(*https://hivdb.stanford.edu/page/graphiql/?example=mutationsAnalysis*):

> Tool offered by Stanford HIVdb showing an example of how programmatically querying their HIVdb Program works. Allows you to edit the query to produce more example results.

Appendix B: Sources

Amino Acid Mutation types:
https://www.ncbi.nlm.nih.gov/books/NBK21578/

Amino Acid Reference:
http://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html

DNA-Translation:
https://www.khanacademy.org/science/biology/gene-expression-central-dogma/translation-polypeptides/a/the-stages-of-translation