Use of MutFuncVis to explore patterns in mutations of the antimalarial drug-resistance protein Kelch13

Brian Gruessner

1 Background:

Proteins can be considered chains of amino acids which, depending on the nature of the amino acid chain, fold into a particular shape that confers a specific function to the protein. As a result, a change to the sequence of the amino acid chain, as can be the result of a DNA mutation, affects the final conformation and function of the protein. The study of the effect of mutations on protein and organism function is of great interest to many fields of medicine, including infectious disease, inheritable disease, and cancer. The introduction of mutations into proteins is also a common technique in molecular biology to dissect the underlying mechanisms of protein function. As such, effective tools to link protein mutation with effect on function would be useful to the life and medical sciences. Typically, the effect of mutations is dependent on their position and severity [14]. Certain structures within the greater protein can be critical to the function of the protein, and these structures will naturally be more susceptible to mutations. These structures can be defined as secondary structures, which are short, commonly seen motifs in proteins, and tertiary structures, which are large, major regions of the protein. Visualization of the locations of mutations with the effects of the mutations on function can reveal the most important structures in the protein. The other factor is the nature of the mutation. Mutations typically substitute one amino acid for another, and this substitution can be either to a similar or drastically different amino acid. The similarity between these substitutions determine how severely the mutation affects protein function. Tools to visualize the affects of both position and substitution dissimilarity suffer from flaws. The VERMONT tool is designed to output mutations that will substantially affect protein function [14]. However, it relies on visualization through a multiple sequence alignment, which makes the visualization messy. It does not lend itself to exploring and understanding the nature of the protein or the severe mutants. [9] PVViewer nicely predicts and visualizes the effects of mutations on secondary structures. However, it does not make judgements on final protein functon. Both [2] and [3] effectively visualize the locations and distributions of mutations by position, but without a clear definition of which mutations affect protein function. The need for a tool that allows for a visualization of the mutations by effect remains unmet. The Mutation Functionality Visualizer (MutFuncVis) was designed with this goal in mind. Given a set of mutations with known effect on protein function, these effects are related to the mutations' positions and similarity of amino acid substitution. This tool will leverage the ability to find patterns through visualization to discover the critical structures and amino acid properties necessary for protein function. Additionally, it will allow for a prediction of the effect of novel mutations. To demonstrate this tool, an analysis of the Kelch13 protein from the malaria-causing protozoan *Plasmodium falciparum* follows. Mutations in this protein have been found to be the driving force behind the emerging resistance to artemisinins [1,10], which make up the foundation of current falciparum malaria treatments [13]. Mutations in Kelch13 and their

worldwide spread are currently of great interest, and their monitoring is important to effective antimalarial resource allocation and containment of drug resistance [11, 13]. The importance of predictive tools became apparent upon the discovery of new Kelch13 mutants in Africa, which carries the majority of the worldwide malarial burden [13]. While it was eventually experimentally determined that these mutants were not drug-resistant, there was a delay between their discovery and understanding of their function [12]. This tool will assist in the understanding of antimalarial drug resistance and better prediction of the effects of new mutations.

2 Goals and Validation:

The goal of MutFuncVis is to leverage effective visualization techniques to allow biologists to discern patterns in mutations that affect protein function. To validate that the tool is effective for this function, MutFuncVis will be used to visualize Kelch13 mutations and draw conclusions about critical structures and properties of Kelch13. These conclusions will then be validated by literature search. The predictive power of the vis will also be examined. While the displayed dataset of mutations from [5] is extensive, it is not exhaustive. More mutations in Kelch13 were experimentally characterized in [3]. These mutations will be entered into the simulation function of this vis and it will be determined whether or not they are predicted to be resistance-conferring. Overall, it is predicted that MutFuncVis will lend itself to generating insights on proteins through examinations of the effects of their mutations. It will also be a useful tool to help predict the effects of uncharacterized mutations.

3 Methods:

An initial set of Kelch13 mutants was found in [5], while a second set used in downstream validation was taken from [3]. To characterize the dissimilarity between mutants, a BLOSUM62 matrix was found from [4], hydrophobicity values from [7], and pKa values from [6]. Kelch13 amino acid sequence and secondary structure data was taken from [8]. Kelch13 tertiary structural data was taken from [3]. A visualization tool was built in D3. A case study was performed to validate the effectiveness of the tool. Finally, the accuracy of the predictions was evaluated with a blinded evaluation of novel mutations found in a separate dataset.

4 Results:

4.1 Case Study: Exploration of Kelch13 properties

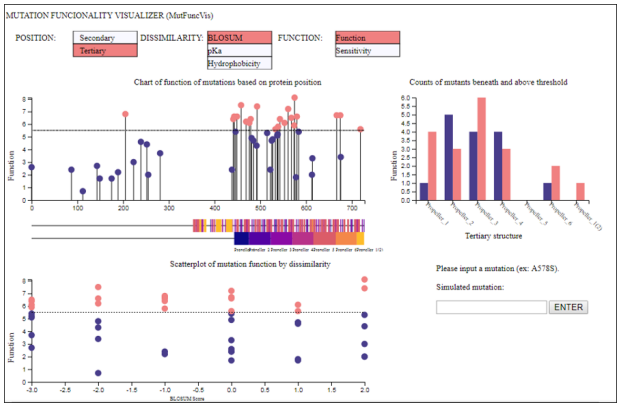

Figure 1: Figure 1: Overview of MutFuncVis

Figure 1 shows the general work area for MutFuncVis, set up to display the Kelch13 protein. At the top are buttons that allow the user to analyze data by different positional, amino acid dissimilarity, and functional output metrics. The upper left chart is a lollipop chart that displays the relative function (in this case, parasite survival half-life in patients undergoing artemisinin-based therapy) by position in the protein. The lollipops are divided into "blue" and "red" colors, representing whether they are below or above the cutoff threshold. It is initially set to 5.5, the cutoff point for "drug resistant malaria" recommended by [5]. The position of this threshold can be altered by clicking in the chart area, which may be useful to adjust the data for cutoffs recommended by other organizations citetilley2016artemisinin. At the bottom of the chart, secondary and tertiary data about the protein are displayed. This region can be

highlighted and brushed to zoom in on relevant regions. A zoom-in on the Kelch13

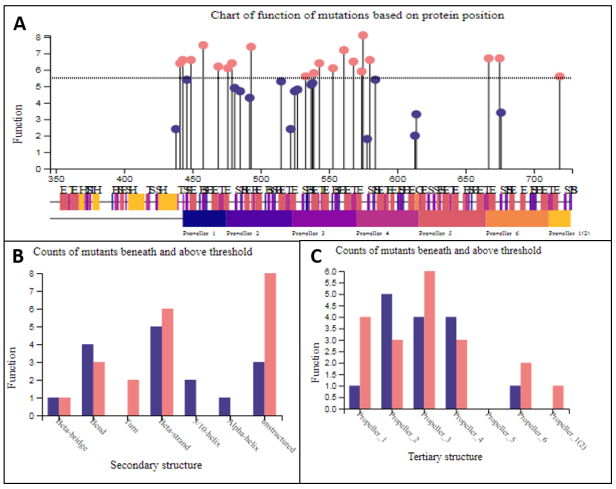

Figure 2: Figure 2: Effect of structural features on mutation effect

propeller domains shows that drug-resistant mutations are present in every propeller domain but Propeller 5 and concentrated in Propellers 1 and 3 (Figure 2A). An early paper that modeled the K13 protein after a human analog predicted disruptions from mutations in Propellers 1 through 4, roughly in line with the vis [1]. In the upper right corner of MutFuncVis, a grouped bar chart displays the total number of mutations above and below the threshold for a given structure. This chart shows that Propellers 1, 3, and 6 have more resistant-conferring mutations than non-resistant mutations, suggesting that they may be sensitive to mutations (Figure 2B). When secondary structures are considered, it appears that turns, beta-strands, and unstructured regions are the most sensitive (Figure 2C). While it is counterintuitive that unstructured regions could be susceptible to disruptive mutations, this was the conclusion of [16] in his analysis of secondary structure and disruptive mutations. It is possible that, due to the small number of observations for some structures, disproportionate numbers of severe observed mutations could skew the data. To account for this, a sensitivity metric was employed. This metric is essentially functional output above the normal baseline divided by a normalized BLOSUM Score, a score of how common mutations are between species.

"Uncommon" mutations tend to more drastically alter protein function and render organisms unviable, hence making them unobservable. The threshold was adjusted to identify the most sensitive structures. Fig-

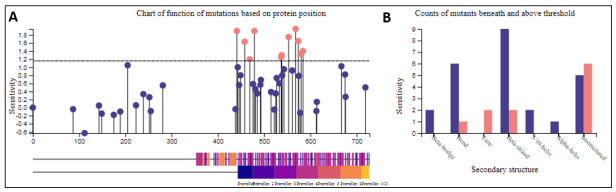

Figure 3: Figure 3: Sensitivity plots show regions most susceptible to mutations

ure 3A infers that Propellers 1 and 3 seem to be the most sensitive to mutation while unstructured regions are, interestingly, the most sensitive secondary-structure regions (Figure 3B). The lower-right scatterplot relates mutation dissimilarity to function output. In this vis, there were not clear relations between the two. The most apparent conclusion from this finding is that position is a far more important indicator of disruptive mutation. There was an interesting isolated resistant-conferring mutation in the earlier portion of the protein (Figure 4A, highlighted point). The highlight/tooltip function of MutFuncVis was utilized here to further analyze the point. The highlighting of this point revealed its identity (I205T), and its location in the scatterplot. Analysis of this mutation type showed a moderate BLOSUM Score (Figure 4B), and minimal changes to pKa(4C) or hydrophobicity(4D). Considering that this mutation does not appear to be dramatic, its position may present an important, poorly-characterized domain in the protein.

4.2 Validation:

To validate the conclusions drawn in this Vis, four outside characterized mutants were entered into the simulation option in MutFuncVis [3]. These mutants were not among those in the dataset used to create this vis. An analysis was performed to predict whether the mutation would confer resistance. This test was performed blinded to the actual determined results. The Q613L mutation is shown as an example (Figure

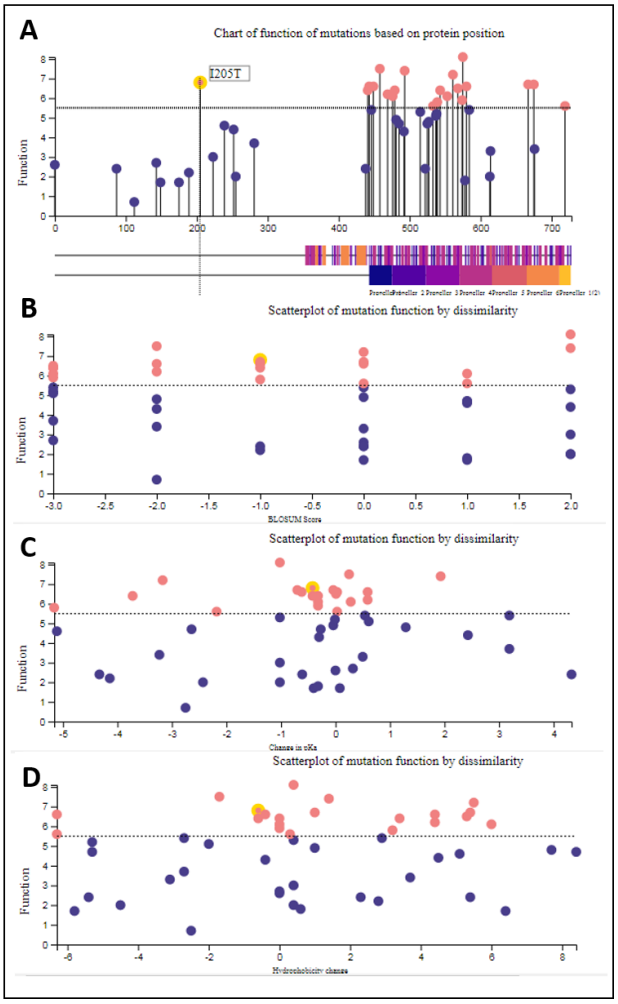

Figure 4: Figure 4: Analysis of dissimilarity metrics in the I205T mutant

5), wherein it is found that the mutation is found in an insensitive region of the protein in the highlighted region of the lollipop chart and is located in a relatively insensitive secondary structure as shown in the highlighted bar chart region. However, it is a moderately severe mutation, as suggested by the highlighted scatterplot region. Given the insensitive region of the protein, it was correctly predicted that this mutant did not confer mutations. The other three mutations were analyzed similarly. The results are shown in Table 1. Three of the four mutants were correctly predicted to be either drug resistant or not. The one incorrect guess, D353Y, was predicted not to confer resistant status when it did. It was in a region far from any other characterized mutants, which made its analysis difficult. This error underscores limitations of this tool that must be considered during its use. It relies on characterized mutant data, and its effectiveness appears to be heavily dependent upon coverage of the protein or protein area with data. The vis tool PlotProtein avoids this problem by using protein alignments with other species to reveal "conserved" amino acids, which would be sensitive to mutation [15]. However, it does not base its findings on experimentally derived data. The need for comprehensive data must be balanced with the power of using experimentally validated functional output data when drawing conclusions from this domain of vis tools.

Table 1: Result of Validation Experiment

Mutation	Confers Resistance?	
	Analysis	Actual
F395Y	NO	NO
Q613L	NO	NO
D353Y	NO	YES
F673I	YES	YES

Figure 5: Table 1: Results of validation test

5 Design Choices:

Following Schneiderman's mantra of "Overview first, zoom and filter, details on demand" [17] , the design is centered around the lollipop chart, which initially shows the full set of mutations. Alternately, the scatterplot shows the full set of mutations for dissimilarity-oriented approaches to mutations. The lollipop chart can then be zoomed to analyze trends in subregions. The bar chart also helps to break down the complex data in the lollipop chart into simpler subsets. Finally, the highlight function allows for individual mutations to be dissected, as shown with the I205T mutation previously. The function threshold for both the scatterplot and lollipop chart were linked together. This allows changes in one chart to be applied to the

other. While there may be some advantage to independently manipulating these thresholds in different charts, vis unity and coherence was considered a more important factor. While this tool was developed to examine the Kelch13 plasmodia protein, it has been coded in a way wherein data for many kinds of proteins can be fed into it using standard formats. It should be adaptable to other instances protein functionality visualization. The simulation appropriately identifies the relevant regions in the various plots for subjective analysis. An algorithm that could generate predicted functional output form a given mutation would be ideal for this tool. However, time constraints did not allow for this feature.

6 Technical Achievements:

As predicted, one of the greater challenges in this project was getting all of the necessary data together for simultaneous use. Apart from physically collecting the information, it was necessary to read it into the program in a way that would allow for simultaneous use of many datasets. This was ultimately accomplished by "daisy-chaining" d3.csv.then functions, wherein the last line of one started the function to read in the next. Ultimately, all of the data was read into global arrays. There were many interactive elements that required extensive debugging to get working across all charts in the vis. In a related challenge, the elements of the visualization were all successfully interlinked to lend coherence to the tool. The use of divs allowed for common html tools to be used alongside complex d3 charts and functions. This was particularly important for the simulation input function.

7 Conclusions:

Overall, MutFuncVis represents an effective tool to explore mutant data to find patterns that have some basis in experimental findings. For predicting the effects of mutants, it appears to be somewhat effective. It should also be considered that this analysis is, to a large extent, subjective. Different researchers may review the same simulated mutation and come to different conclusions. This simulation would benefit strongly from regression analysis tools to predict a function output from a given mutation. Overall, MutFuncVis represents a new approach to modelling protein function by visualization of experimentally determined function over the entire protein and its spectrum of amino acid substitutions. Here it has been demonstrated in the field of antimalarial drug resistance. However, this tool could easily be adapted to approach a variety of important challenges in functional proteomics.

References

[1] F. Ariey, B. Witkowski, C. Amaratunga, J. Beghain, A.-C. Langlois, N. Khim, S. Kim, V. Duru, C. Bouchier, L. Ma, et al. A molecular marker of artemisinin-resistant plasmodium falciparum malaria. *Nature*, 505(7481):50, 2014.

[2] W. W. de Leng, C. G. Gadellaa-van Hooijdonk, F. A. Barendregt-Smouter, M. J. Koudijs, I. Nijman, J. W. Hinrichs, E. Cuppen, S. van Lieshout, R. D. Loberg, M. de Jonge, et al. Targeted next generation sequencing as a reliable diagnostic assay for the detection of somatic mutations in tumours using minimal dna amounts from formalin fixed paraffin embedded material. *PloS one*, 11(2):e0149405, 2016.

[3] M. P. falciparum Community Project et al. Genomic epidemiology of artemisinin resistant malaria. *elife*, 5:e08714, 2016.

[4] N. C. for Biotechnology Information. Blosum62, nd.

[5] W. K. G.-P. S. Group et al. Association of mutations in the plasmodium falciparum kelch13 gene (pf3d7_1343700) with parasite clearance rates after artemisinin-based treatments via wwarn individual patient data meta-analysis. *BMC medicine*, 17(1):1, 2019.

[6] I. Hunt and R. Spinney. Table of pka and pi values, 2006.

[7] H. Jakubowski. Hydrophobic- ity indices for amino acids, nd.

[8] D. Jiang, W. Tempel, P. Loppnau, S. Graslund, H. He, M. Ravichandran, A. Seitova, C. Arrowsmith, A. Edwards, C. Bountra, M. El Bakkouri, G. Senisterra, K. Osman, D. Lovato, R. Hui,

A. Hutchinson, Y. Lin, and S. G. C. (SGC). 4yy8: Crystal structure analysis of kelch protein from plasmodium falciparum, 2015.

[9] L. Kocincová, M. Jarešová, J. Byška, J. Parulek, H. Hauser, and B. Kozlíková. Comparative visualization of protein secondary structures. *BMC bioinformatics*, 18(2):23, 2017.

[10] A. Mbengue, S. Bhattacharjee, T. Pandharkar, H. Liu, G. Estiu, R. V. Stahelin, S. S. Rizk, D. L. Njimoh, Y. Ryan, K. Chotivanich, et al. A molecular mechanism of artemisinin resistance in plasmodium falciparum malaria. *Nature*, 520(7549):683, 2015.

[11] S. Meshnick. Artemisinin resistance in southeast asia. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 63(11):1527, 2016.

[12] J. Muwanguzi, G. Henriques, P. Sawa, T. Bousema, C. J. Sutherland, and K. B. Beshir. Lack of k13 mutations in plasmodium falciparum persisting after artemisinin combination therapy treatment of kenyan children. *Malaria journal*, 15(1):36, 2016.

[13] W. H. Organization. World malaria report 2017. *Geneva*, CC BY-NC-SA 3.0 IGO, 2017.

[14] S. A. Silveira, A. V. Fassio, V. M. Gonçalves-Almeida, E. B. de Lima, Y. T. Barcelos, F. F. Aburjaile, L. M. Rodrigues, W. Meira Jr, and R. C. de Melo-Minardi. Vermont: Visualizing mutations and their effects on protein physicochemical and topological property conservation. In *BMC proceedings*, volume 8, page S4. BioMed Central, 2014.

[15] T. Turner. Plot protein: visualization of mutations. *Journal of clinical bioinformatics*, 3(1):14, 2013.

[16] V. Vacic, P. R. Markwick, C. J. Oldfield, X. Zhao, C. Haynes, V. N. Uversky, and L. M. Iakoucheva. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS computational biology*, 8(10):e1002709, 2012.

[17] J. S. Yi, Y. ah Kang, and J. Stasko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007.