Visualization of effects of mutations in the *P. falciparum* Kelch13 protein on antimalarial resistance

Brian Gruessner

1 Introduction

The outward spread of drug resistance to the antimalarial drug artemisinin from Southeast Asia is an issue of major concern to human health [10]. The current strategy to contain resistance is to carefully monitor patient response to treatment and to direct resources to areas where treatment failures and resistant mutants of *Plasmodium falciparum* (parasite that causes malaria) are found [13]. In *P. falciparum*, mutations in the *kelch13* gene can disrupt the normal function of the Kelch13 protein and confer drug resistance to the parasite [11]. While there is some information known about the protein function, structure, and important sites, it is still not straightforward to predict which mutations will cause drug resistance. This issue is highlighted in the instance of the discovery of "A578S" mutation that led to a temporary scare that drug resistance had spread to Africa, which carries most of the world's malaria burden [14]. To address this concern, this visualization tool will generate plots of known amino acid substitutions based on location in the protein, degree of change to the protein caused by the mutation, and degree of drug resistance conferred by the mutation. This will allow for a search for patterns int eh data that can be used to predict the effects of newly discovered changes in Kelch13 in the future. Finally, this tool could be easily adapted to predict preservation of protein function against amino acid substitutions in any protein of interest in biology and medicine. This adaptation could have implications in, for instance, inheritable disease, cancer biology, and non-human functional genomics.

2 Summary

This tool will visualize the properties of known amino acid substitutions in Kelch13 to aide in the search for predictive patterns in mutations that confer antimalarial resistance.

3 Project Type

This vis will be designed for data visualization and analysis in the field of computational biology. It is designed to generate hypotheses.

4 Audience

This vis would be geared towards malaria-oriented computational biologists. They would have a clear understanding of the technical metrics used in this vis. The question being addressed would be a useful one for malaria scientists. Specifically, this tool could be of use to scientists monitoring the spread of artemisinin resistance [3,5,12] who occasionally discover novel mutations and could benefit from a quick initial assessment of the likelihood that it represents a new resistant variant. For the vis to succeed in this goal, there will need to be an option insert a hypothetical mutation point into the vis to estimate its mutation potential. Finally, the skeleton of the code could readily be adapted for functional probes of any protein of interest to computational biologists. For this to be feasible, the project will rely

on metrics and terms that can be applied universally in biology. Substitution of genes and new analyses will then rely on simply switching out data tables and a few labels.

5 Approach

5.1 Details

Abstraction of data: To generate this visualization, it is first necessary to explore what data is needed and to categorize this data. Examination of datasets have revealed a typical pattern of presentation of data on Kelch13 amino acid substitutions and resistance potential. A typical point of data contains three types of information: it contains the location in the protein of the substitution, the specific substitution that occurs, and the degree of resistance that the mutation confers. As such, this visualization must effectively express three channels of data. Each channel can be expressed through multiple metrics (see *Table 1*). For instance, position can be expressed as the amino acid position in the protein, which is essentially a string of amino acids. This could also be expressed through the secondary structure present at a given point (small structures in protein). There is also tertiary structure, which would represent large regions of the protein that can be identified in 3D models of the protein. In Kelch13, the key tertiary structures would be its "propeller blades" [1, 3]. "Substitution Dissimilarity" measures how drastic a change the amino acid change represents. More drastic changes are more likely to upset the larger protein structure, though this likelihood will likely interact with the position channel. Differences in amino acid acidity/charge, measured in pKa, and polarity, measured by the hydrophobicity index, can alter how the protein interacts with itself and other molecules, potentially changing its structure and ability to function. Alternately, the BLOSUM score represents how likely a substitution from one amino acid to another is between species, based on the rationale that no change or small changes are more common than drastically dissimilar substitutions. Finally, there is the "Function Change" metric. For the malaria vis, this output will represent artemisinin drug resistance. This vis will primarily use the "parasite clearance half-time", which measures how quickly the parasite disappears from the patient's blood over the course of treatment. This is because of the existence of a large database of mutations that uses this metric [5]. However, the laboratory "ring stage survival" metric also estimates the resistance well [16]. The "Function Change" metric will be set up so that it could be adapted to represent the function of any protein through an appropriate clinical test or test-tube function assay.

X: Position	Y: Substitution dissimilarity	Z: Function change (resistance)
X1: Amino acid position	Y1: BLOSUM score (adjusted)	Z1: Parasite clearance half-time
X2: Secondary structure	Y2: Change in acidity *(ΔpKa)*	Z2: Ring-stage survival %
X3: Tertiary structure	Y3: Change in hydrophobicity	

Figure 1: Table 1: Channels for Data

Gathering of data: The paper [5] dataset of protein substitutions for the Kelch13 protein and their clinical clearance rates for patients worldwide. Given its importance, the 3D protein structure of Kelch13 has been experimentally determined [8], and its secondary [8] and tertiary [3] motifs are available. BLOSUM scores [4] and general amino acid properties like pI [6] and hydrophobicity indices [7] are publicly available. One challenge for this vis will be validation. This will require finding characterized mutations not in the main dataset and determining whether the patterns in the vis can predict its characteristics. A separate comb for mutations was found in [3].

Presentation of data: Mutational data is often portrayed as a lollipop plot. A representative example is shown below from [2]. It demonstrates the effective visualization of many metrics and submetrics at once. The lollipop plot can show protein position and associated higher structures, a second metric can be encoded by vertical position (like degree of resistance), and other metrics can be encoded by the lollipop character, including shape and color. This type of plot is an effective choice for showing the entire mutation dataset before filtering and zooming. Realistically, this plot would plot positional data against resistance data to explore the hypothesis that position determines resistance. Alternately, a resistance /

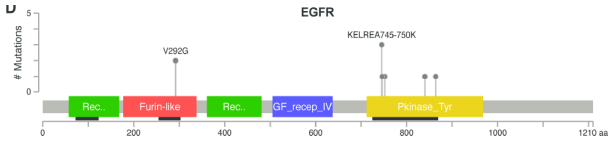

Figure 2: Figure 1: Lollipop plot. This plot is an appealing choice for the bioinformatic domain because of its ability to efficiently fit multidimensional mutation data to meaningful positional data.

dissimilarity vertical metric would plot out a "sensitivity curve" of sorts. The vertical component could be given a cutoff metric that could be manually adjusted, represented by a dashed line. For instance, the default cutoff ratio for "resistant malaria" would be a clearance halftime of 5.5 hours, as suggested in [5]. The ratio of mutations, by structure, above and below this cutoff line would then be presented in a bar graph. To explore the hypothesis that the nature of the mutation determines resistance, a scatterplot of dissimilarity metric vs resistance will be created. Dot color will be determined by positional structure. Using the bar graph from the previous paragraph, structures could be triggered on or off as a filter on this graph. Finally, to aide in characterization of novel mutations, this vis should have a means to calculate and locate positions and dissimilarity metrics for novel mutations. This could be performed using pre-existing databases, with positions/dissimilarity rows/columns in the chart being shaded. The user would be left to draw conclusions based on the patterns in the graphs.

5.2 Evidence for Success

Previous attempts to visualize Kelch13 placed dots on a 3D structure of the protein to represent mutations with high levels of resistance [1]. The vis suggested a clustering of points mostly around three of the six major tertiary structures but did not further analyze the mutations. This attempt, while by no means exhaustive, suggests that there are patterns to be found. Other papers have successfully found patterns across a wide range of proteins by analyzing the relationship between function and secondary structure [9, 15]. This type of vis has some precedent in the literature, but this project will build on previous attempts by attempting a more thorough analysis that allows for interaction with the data.

6 Best-case Impact Statement

In the best-case scenario, this vis will aide in the discovery of new patterns involved in mutations associated with antimalarial drug resistance. This will allow for more informed decisions concerning resource allocation for the suppression of drug-resistant malaria when novel mutations arise.

7 Major Milestones

-Gathering and organization of data: All of the necessary information has already been found. Organization will require the parsing of mutation names into position and amino acid substitution and calculation of dissimilarity properties based on the substitution. Additionally, secondary and tertiary data and processing for Kelch13 will be necessary. This will require the use of multiple datasets at once. Given the manageable size of the mutation dataset (about 50 members), these calculations are not expected to substantially slow the program.

-Creation of the lollipop plot: This plot utilizes structure graphics and likely a brush option in addition to a conventional plot grid.

-Development of supporting Bar chart and scatterplot.

-Development of tools to iterate through metrics of channels. This will likely be performed through a small clickable list at the top of the vis, organized like in *Table 1.*

-Development of a novel mutation introduction tool. This will intake a substitution through a standardized nomenclature, calculate the associated metrics of the substitution (excluding resistance), and display its location on the charts.

-Validation. The current plan is to find recently discovered, characterized mutations in new studies and compare their findings to what could be predicted by this vis. A dataset with mutations not found in [5] has

been found [3]. However, its resistance data is given as a binary yes/no value. Finding a suitable way to compare the empirical results with what the vis would independently predict will be an important challenge in this vis

8 Obstacles

8.1 Major obstacles

-Development of a brush tool. This looks technically intimidating but is likely necessary to fit the lollipop chart on the screen.

-Generating functions to parse input data and calculate metrics using various datasets.

-Creation of function to add in new positional and substitution points from user input.

8.2 Minor obstacles

-Creation of the extensive interaction between the charts.

-Development of basic regression analysis. This would add some statistical weight to my vis but will be technically demanding to implement.

-Development of the code in a context that will allow it to be adapted for use outside of antimalarial resistance.

9 Resources Needed

The necessary data has already been gathered for the core of this project. This includes the sequence of the Kelch13 protein and its secondary and tertiary structures, two datasets of P. falciparum Kelch13 mutants found in patients with corresponding resistance information that contain mutually exclusive mutations, a BLOSUM amino acid substitution matrix, amino acid property information such as hydrophobicity and pI. Ideally, an additional dataset outside of malaria with a large amount of gene mutation info and functional output will also be collected and attempted to be incorporated into the working vis to demonstrate the adaptability of the project, but this is a secondary goal.

10 Related Publications

[5]: This paper provides a large, well-annotated dataset of discovered K13 substitutions with specific functional output (resistance). However, as is apparent by flipping through the paper, the article is very table-heavy and does not effectively visualize the data to draw conclusions.

[8]: This lab group discovered the 3D structure of the Kelch13 protein and allowed for the determination of the secondary structure. Their work was instrumental to the visualization in [1] and the naming of the key tertiary structures [3] and for having reliable information for this Kelch13 project.

[3]: The MalariaGen paper produced a second independent dataset of mutations with resistant/sensitive characterization, which can be used to validate the effectiveness of this vis vs. [5]. The group also identified regions of major tertiary structures in Kelch13, allowing me to fill out the positional data for the protein.

[1]: This group created an early visualization of resistant mutations. Positions were overlaid on the 3D structure of the protein [8] to reveal that they were skewed towards one half of the Kelch13. While this is an interesting trend that could potentially be replicated in this project, the importance of substitution of amino acid was not explored in this graphic.

[2]: While not malaria-related, this vis provided an effective and interesting idea for a visualization that could successfully accomplish the goals of this project. The lollipop plot that this paper utilized (*Figure 1*) is effective at representing multiple dimensions of data over robust positional data.

11 Define Success

Success will be defined as the creation of a visualization that can generate hypotheses about features of missense mutations in the Kelch13 protein that predict artemisinin drug resistance. Ideally these hypotheses should be backed up by more validated data not used to generate the visualization.

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—a 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] Y. Lubell, A. Dondorp, P. J. Guérin, T. Drake, S. Meek, E. Ashley, N. P. Day, N. J. White, and L. J. White. Artemisinin resistance–modelling the potential human and economic costs. *Malaria journal*, 13(1):452, 2014.

[11] 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.

[12] D. Ménard, N. Khim, J. Beghain, A. A. Adegnika, M. Shafiul-Alam, O. Amodu, G. Rahim-Awab, C. Barnadas, A. Berry, Y. Boum, et al. A worldwide map of plasmodium falciparum k13-propeller polymorphisms. *New England Journal of Medicine*, 374(25):2453–2464, 2016.

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

[14] 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.

[15] 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.

[16] B. Witkowski, N. Khim, P. Chim, S. Kim, S. Ke, N. Kloeung, S. Chy, S. Duong, R. Leang, P. Ringwald, et al. Reduced artemisinin susceptibility of plasmodium falciparum ring stages in western cambodia. *Antimicrobial agents and chemotherapy*, 57(2):914–923, 2013.