S.I.E.V.E.: Statistical Interactive Explorer of Vaccine Efficacy

Nick Kullman, Graham Clenaghan, and Wayne Yang

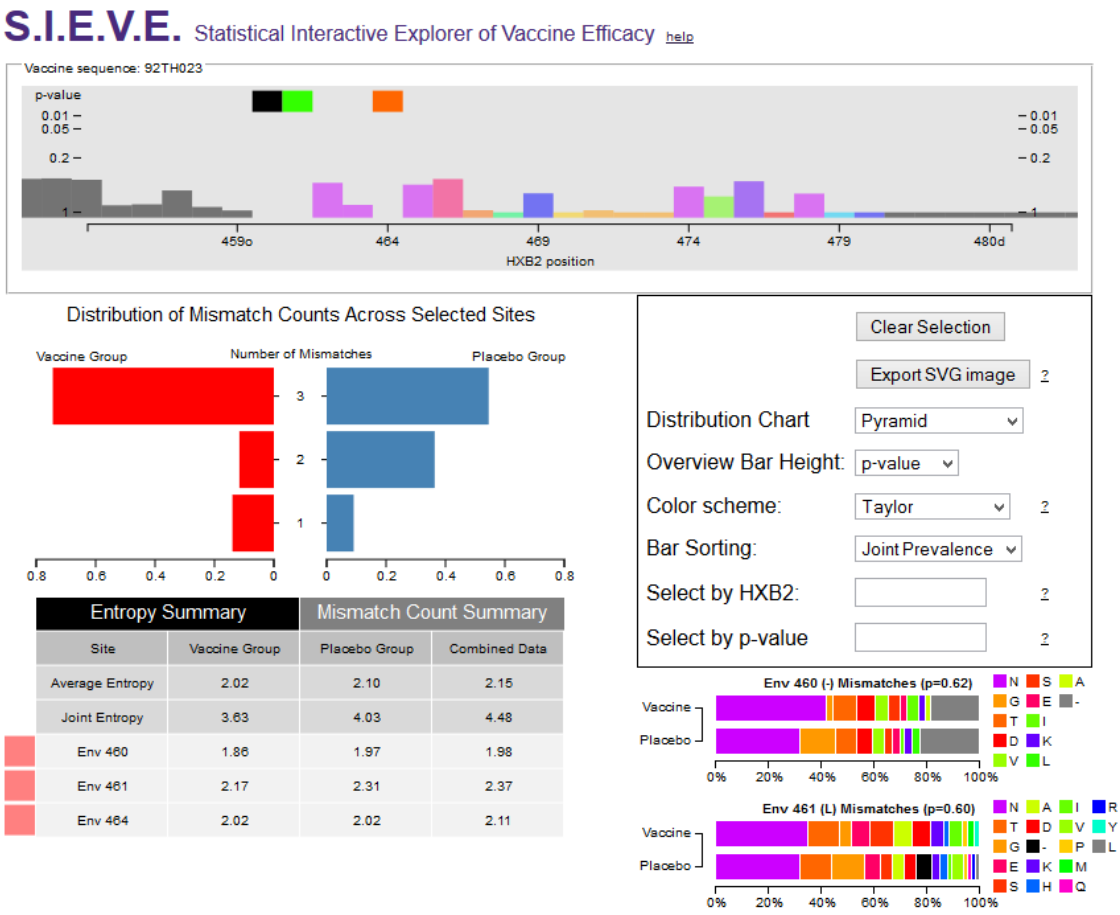


Fig. 1. Overview of the S.I.E.V.E. interface.

Abstract— In vaccine trials showing partial vaccine efficacy, researchers employ a statistical method known as sieve analysis to compare protein sequences of the viral strands found in infected study participants in placebo (unvaccinated) and vaccinated groups. The comparisons seek to find sites in the sequence where the unvaccinated and vaccinated participants exhibit significantly different distributions of amino acids. These differences may indicate vaccine-induced immune pressure on the virus, providing information valuable in constructing the next generation of vaccines. Detecting differences through site-wise comparison is difficult without the aid of visualizations. Constructing visualizations for comparisons is not currently automated and therefore labor intensive, as researchers must manually create charts for each site in the sequence for which they wish to make a comparison. Researchers also frequently study groups of sites simultaneously, making the manual chart-construction process even more challenging and laborious. The current obstacles to visualization inhibit researchers' ability to discover sites of significance and therefore slow the progress of vaccine development. We describe here our creation of SIEVE, a flexible, interactive tool facilitating the sequence comparison process of sieve analysis.

Index Terms—Sieve Analysis, Clinical Trials, Vaccine Efficacy.

1 Introduction

AIDS and other viral diseases like influenza and dengue claim more than a million lives each year. Vaccinations constructed to combat these diseases often exhibit partial efficacy. In vaccine trials showing partial vaccine efficacy, researchers employ a statistical method known as sieve analysis (see figure 2) to compare protein sequences of the viral strands found in infected study participants in placebo (unvaccinated) and vaccinated groups [2]. The comparisons seek to find sites in the sequence where the unvaccinated and vaccinated participants exhibit significantly different distributions of amino acids. These differences may indicate vaccine-induced immune pressure on the virus, providing information valuable in constructing the next generation of

- *Nick Kullman. E-mail: nkullman@uw.edu*
- *Graham Clenaghan. E-mail: clenagh@math.washington.edu.*
- *Wayne Yang. E-mail: wfyang@uw.edu*

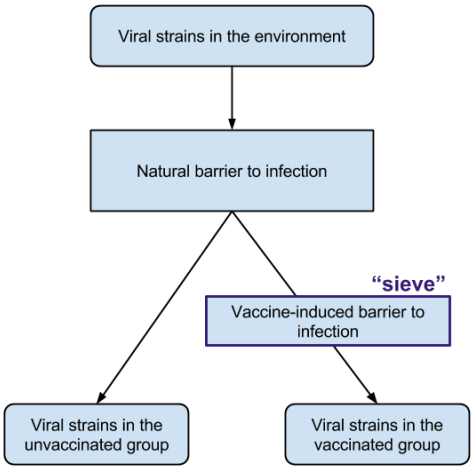


Fig. 2. Sieve analysis schematic.

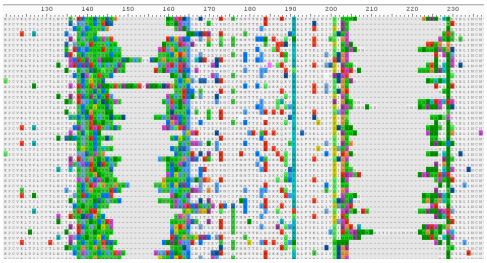


Fig. 3. AliView interface.

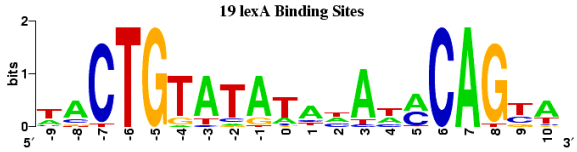


Fig. 4. Example WebLogo image.

vaccines. Detecting differences through site-wise comparison is difficult without the aid of visualizations. Constructing visualizations for comparisons is not currently automated and therefore labor intensive, as researchers must manually create charts for each site in the sequence for which they wish to make a comparison. Researchers also frequently study groups of sites simultaneously, making the manual chart-construction process even more challenging and laborious. The current obstacles to visualization inhibit researchers' ability to discover sites of significance and therefore slow the progress of vaccine development. To facilitate sequence comparison and the discovery of sites exhibiting potential vaccine-induced immune pressure we created SIEVE. SIEVE allows for rapid comparison of amino acid distributions at any site or group of sites selected by the researcher. It allows rapid changes in selection, providing both aggregate and site-specific information updated dynamically as the selection changes. In the following sections we describe previous work related to SIEVE and introduce the methods our team used to create it. In section 4 we more fully describe the SIEVE tool. In section 5 we discuss the positive feedback from early adopters of SIEVE. We conclude with a brief discussion of future possibilities for development.

2 Related Work

Visualizations for sieve analysis exist, but none of them completely fulfill the task as they do not allow for easy graphic manipulation nor the exploration of data at the level needed for effective vaccine efficacy analysis.

2.1 Sequence Data Visualization Tools

There are several existing tools for the visualization of genomic/proteomic sequence data. Some of these tools tend to provide a very detailed display of the alignments of sequences for a particular gene/protein from multiple patients. While this approach allows a user to see all of their sequence data at once, it does not provide quick or easy analysis of particular sites in the sequence across patients or subsets of patients as is required in sieve analysis. The example shown in figure 3 is from a software called Aliview [4]. Clearly, understanding the pattern of mutations and any relationships to treatment status for a particular site in the sequence is very difficult to do in this view with any precision.

Other tools are geared towards specific analyses of sequence data. The image in figure 4 was generated using WebLogo [3], which is an online tool that can be used to parse sequence data and generate plots. In addition to not supporting the types of analysis needed, such as statistical comparison of two treatment groups, the interface to WebLogo, pictured in figure 5 is not conducive to exploratory data analysis since the user has to be quite specific about what is actually plotted. If there are sites that are known to be interesting, then a researcher can use WebLogo to construct plots, but rapid changes in site-selection and the corresponding visualizations are impossible.

Another tool that came up in our discussions with Fred-Hutchison staff scientist Dr. Andrew Gartland is the HIV Genome Browser, which allows users to browse reference data hosted online [1]. The browser is highly interactive and allows a user to explore the available sequences. However, this visualization does not easily allow a user to compare across patients and treatment groups as required by sieve analysis.

2.2 Sieve Analysis Graphics

The main inspiration for the project is [2], in which the authors perform sieve analysis on the RV144 HIV vaccine trial. Graphics created for this paper include an overview of the HIV genome with statistically significant sites annotated, figure 6, and charts of individual sites showing mismatch prevalence, figure 7. The authors generated these graphics using a python script which requires manual coding in order to create each plot. The SIEVE tool described here aims to streamline the process of sieve analysis by allowing interactive exploration of sequence data and automatic generation of relevant charts.

3 Methods

Our goals in designing the interface were the following:

- Sequence viewing software should be somewhat familiar to researchers.

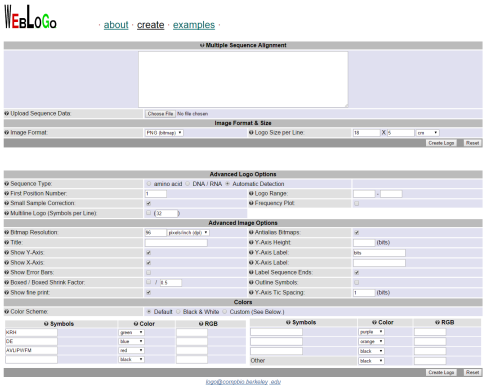


Fig. 5. Interface to WebLogo.

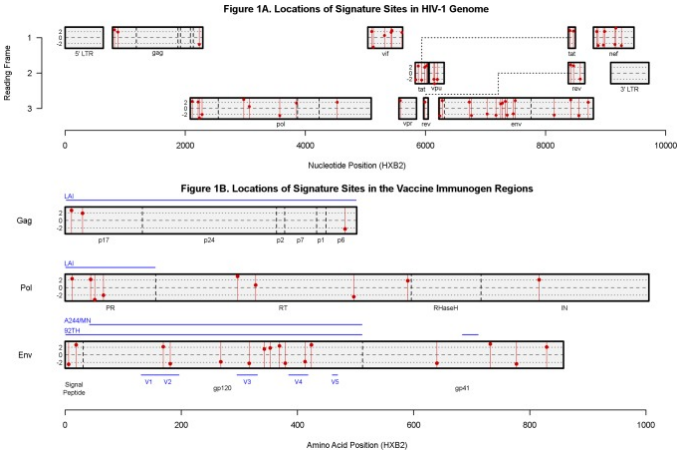


Fig. 6. Overview of sequenced proteins.

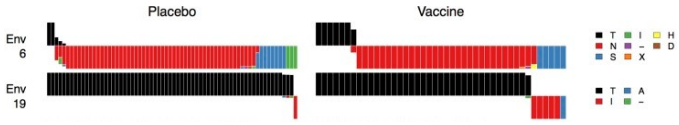


Fig. 7. Visualization of distribution of mutations at a particular site.

- Sites of potential interest should be quickly identifiable.
- Sites of known interest should be easily found.

To accomplish this, we decided to break the tool into three parts: an overview showing a summary of the entire data and the ability to drill down and select particular sites, and then once a selection is made, both a summary of the overall statistics and relevant charts for each selected site.

3.1 Selection

The selection interface (figure 8) aims to allow quick navigation through the sequence while highlighting sites of potential interest. Each bar represents an amino acid, and the sequence can be zoomed and panned to move through the data fluidly. The height of the bars representing amino acids is used to encode statistics about the site: p-value and entropy are currently supported and other user-specified statistics could be included in future extensions of SIEVE.

A selection is made by holding shift, clicking and dragging across the bars. A selected bar is distinguished from non-selected bars through color (opacity) and position encoding - selected bars are raised above the heights of non-selected bars. Trials during development of SIEVE showed that this choice of encoding allowed for detection of selected sites even when the zoom extent is at minimum and the entire overview sequence is visible. Because of this, the user can comfortably select multiple groups of sites across the genome. Since proteins fold, non-contiguous groups in the genome may be spatially close, so such selections are necessary for studying recognition of HIV by antibodies.

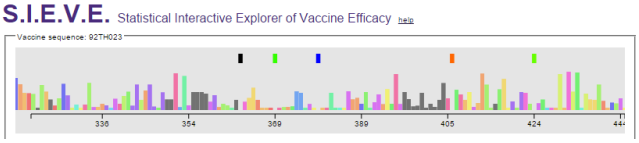


Fig. 8. Selection interface.

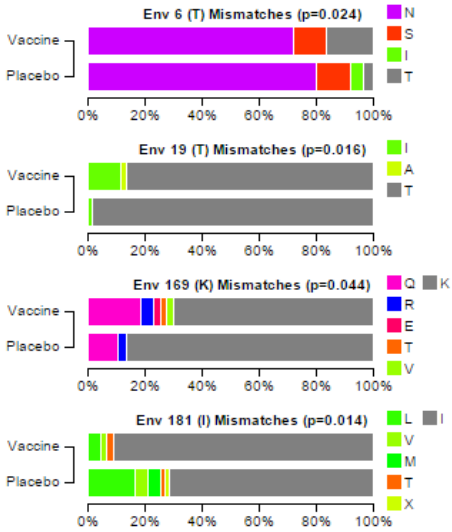


Fig. 9. Site-wise mismatch charts.

3.2 Site-wise Comparisons

Once a selection is made, stacked bar charts are generated for each site in the selection (see figure 9). The stacked bar charts show the distribution of amino acids at the site for both the vaccine and placebo groups. The presence of an amino acid at a site different from the vaccine's amino acid at that site is known as a mismatch. Mismatches have proven to be a useful gauge of a site's importance in sieve analysis [2]. The stacked bar chart figures were designed as a direct replacement for those appearing in figure 7, but allowing easier comparison between the two groups.

The color scheme used throughout is a known encoding for amino acids which gives chemically similar acids similar colors. Other options are provided, including colors used in WebLogo and a 20-category color scheme included in D3.js which maximizes the perceptual difference between acids.

If a researcher finds an interesting chart, this part of the tool can be exported as an SVG, preserving the formatting options. The exported file can be easily annotated using an SVG editor and converted to other formats for use in publications.

3.3 Group Data

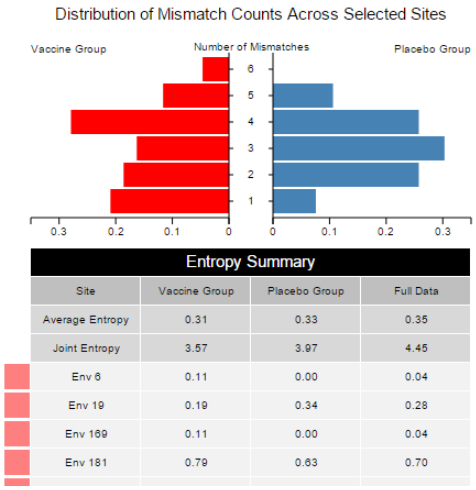
Group statistics about a selection are shown in two ways: a chart representing the distribution of mismatch counts, and a table showing summary statistics such as joint entropy, as well as site-specific statistics for all sites selected. These are seen in figure 10.

The data set in the chart is the number of mismatched amino acids in the current selection for each patient. There are two options to view this data, either as a pyramid chart (as shown in figure 10) showing a histogram of mismatch counts, or as a box plot.

The table in figure 10 can either show statistics about entropy, which measures how much variety there is in the number of amino acids seen at a particular site or selection, or the raw mismatch counts.

3.4 Configuration Options and Alternate Selection

A small number of configuration options are available to the user to increase the flexibility of the tool while preventing it from being overwhelming. These options are displayed in figure 11. Of particular importance is the ability to change the measure encoded by the height of the bars in the selection chart. These measures are used to indicate sites where interesting behavior is observed, so being able to view



Entropy Summary

Site	Vaccine Group	Placebo Group	Full Data
Average Entropy	0.31	0.33	0.35
Joint Entropy	3.57	3.97	4.45
Env 6	0.11	0.00	0.04
Env 19	0.19	0.34	0.28
Env 169	0.11	0.00	0.04
Env 181	0.79	0.63	0.70

Fig. 10. Mismatch distribution and entropy summary.

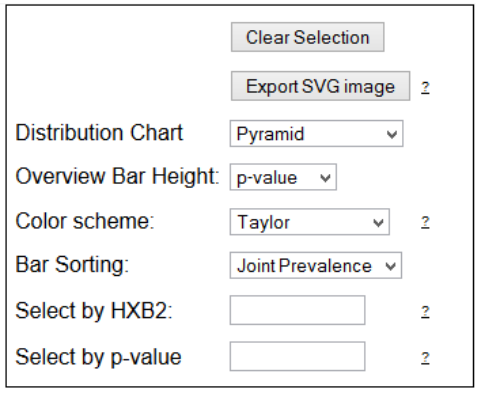


Fig. 11. Configuration Options

a quick overview of what sites are important according to different metrics is critical to the task of exploratory analysis.

The two text fields in figure 11 are alternative selection mechanisms. In the first text field, a user can select sites by entering them as a comma separated list of positions or ranges of positions (based on the indexing used in the HXB2 reference HIV strain). The second text field selects all sites with a p-value less than or equal to the threshold entered. So, if a user enters 0.05 into the second field, the selection will be changed to the set of sites with $p < 0.05$.

3.5 Tools

We chose to accomplish this using the D3.js visualization library so that our web-based interface would be as accessible as possible to researchers in the field. In addition, we planned features to aid in editing and exporting the graphic for inclusion in sieve analysis papers and we hoped to design it be easily extensible by researchers to add new features as they saw fit, such as different statistical measures and different data sets.

4 Results

The final product is a lightweight tool designed for the Google Chrome web browser which requires only aligned sequence data in the form of FASTA files (a common format for nucleotide or peptide sequence data) and a CSV file of p-values. Besides the p-values, all mismatch and mutation information is computed in the browser using Javascript. While this may cause slowdowns when selecting upwards of five hundred sites at a time, the intended use generally involves less than a hundred sites in a single selection. The benefit of this strategy is that it requires much less preprocessing by the user and is sufficient for the two major use cases for researchers.

The first use case is to use SIEVE as a quick reference for known sites of interest. For example, if a researcher is interested in showing a colleague information regarding the pattern of mutations at a particular site or set of sites, SIEVE can facilitate rapid access which circumvents having to dig through their hard drive to find relevant figures. The researcher can either use the selection bar (figure 8) to navigate to a particular site or simply enter the index into the selection text field.

The second major use case is to streamline the process of exploratory analysis by allowing a researcher to identify potentially interesting sites based on entropy or p-value and then immediately generating comparative plots and statistics. In this case, the primary mechanism for deciding which sites are the measures encoded as height in the selection bar (figure 8). Once those sites have been identified, adding them to the selection by holding shift, clicking, and dragging immediately brings up information (see figure 9) which can help decide if that particular site is of actual biological interest.

5 Discussion

Feedback from the vaccine researchers we worked with has been extremely positive, and they have expressed excitement for using the tool on future studies. In one instance, while demonstrating an early prototype, one researcher had a specific site which was significant in a paper he recently read. He was able to navigate the overview to the site and select it, and compare our charts to those in the paper from a different cohort. After this instance, we added the ”Select by HXB2” option to support tasks such as these.

6 Future Work

Several extensions to this project are possible. The first extension would be to make swapping data sets easier in order to be able to use SIEVE on a greater variety of studies. Part of this process would involve setting up a framework which would allow for the uploading of files. Another part of this process might involve moving some calculations to the server side. This might reduce the file requirements down to the bare minimum of the aligned sequences.

Another potentially interesting extension would be to integrate other sources of data. One might consider incorporating some tool to select groups of sites which are known to be related biologically. For example, being able to select the sites associated with known epitopes (parts of the virus recognized by the immune system) listed in a database could be a valuable tool.

ACKNOWLEDGMENTS

The authors wish to thank Andrew Gartland and Allan DeCamp for providing the project idea, relevant data, and insightful feedback on SIEVE and also the Data Visualization course instructor Jeff Heer and assistants Jeff Snyder and Dominik Moritz.

REFERENCES

[1] Hiv genome browser. http://www.hiv.lanl.gov/.

[2] P. T. Edlefsen, M. Rolland, T. Hertz, S. Tovanabutra, A. J. Gartland, A. C. deCamp, C. A. Magaret, H. Ahmed, R. Gottardo, M. Juraska, et al. Comprehensive sieve analysis of breakthrough hiv-1 sequences in the rv144 vaccine efficacy trial. *AIDS research and human retroviruses*, 30(S1):A25–A26, 2014.

[3] J. C. G. Crooks, G. Hon and S. Brenner. Weblogo: A sequence logo generator. *Genome Research*, 14, 2004.

[4] A. Larsson. Aliview: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30, 2014.