Computational Analysis of Somatic Hypermutation

PENNSTATE
Department of Computer Science and Engineering
College of Engineering

Benjamin J. Lengerich
Adviser: Raj Acharya
Department of Computer Science and Engineering,
The Pennsylvania State University

Introduction

The human immune system protects against a diverse set of antigens by dynamically generating an antibody to pair with each potential antigen. As shown in Figure 1, immunoglobulin diversity is generated through a two stage process: (1) V(D)J recombination followed by (2) somatic hypermutation.

Unfortunately, the targeting of somatic hypermutation is poorly understood. Controlled through a balance of error-prone (through cytosine deamination by activation induced deaminase) and high-fidelity DNA repair, somatic hypermutation preferentially targets some *Ig* loci[1], known as hypervariable regions. However, there is currently no generative model of the distribution of hypervariable regions, leading to difficulty understanding immune responses.

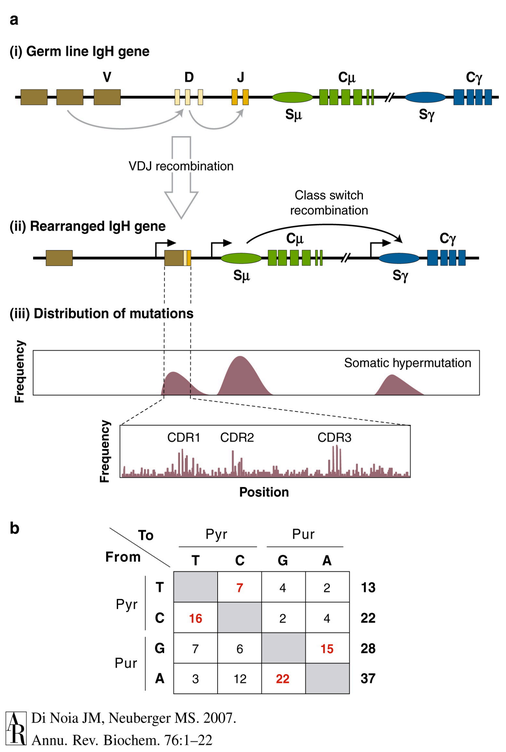

Figure 1. The mechanisms of V(D)J recombination and somatic hypermutation.

Objectives

The motivation for this project is 4-fold:

1. To unearth characteristics of somatic hypermutation targeting.
2. To generate a dataset for further investigation of somatic hypermutation.
3. To increase accuracy of antigen and immune response models.
4. To advance toward a comprehensive and dynamic anti-viral software by advancing requisite computational models of immune responses.

Materials and Methods

First, as no publicly available datasets have been developed for this purpose, a dataset of somatic hypermutation rates had to be constructed. Using the Stanford_S22 dataset[2] (a standard for benchmarking the performance of human antibody gene alignment tools) that consists of 13,153 sequence reads and over 3,000,000 base pairs, hypermutation rates were calculated. After calculating the most closely aligned germline recombination, each base pair was compared to the corresponding germline location to identify mutations. Mutations were recorded as 1, conserved nucleotides as 0. This created a dataset of over 3,000,000 base pairs annotated by hypermutation rate.

To discover motifs associated with non-hypervariable regions, the non-hypervariable regions were extracted. These regions were defined as segments at least 5 base pairs long with no mutations, along with the surrounding 10 base pairs. DREME[3] was then used to identify putative motifs enriched in the non-hypervariable regions.

To label V/D/J segments, a conditional random field (implemented by CRF++) was used. After training on a synthetic dataset, it was tested for accuracy and generalizability on the Stanford_S22 dataset. While training the conditional random field was computationally intensive ($\mathbf{O}(TM^2)$), using the trained model to label sequences is efficient ($\mathbf{O}(M)$). Thus, the use of the conditional random field allowed for extensive statistical tests (implemented in Python) to be performed on labeled sequences, of which a few significant results are shown in the Results section.

Results

While hypervariable regions have been extensively studied for motifs, the complementary non-hypervariable regions have not. By analyzing non-hypervariable regions for common nucleotide patterns, 19 significant (e-value < 0.0001) motifs were discovered. The 4 with highest enrichment are displayed in Table 1. The number and significance of these motifs suggest that there may be protective factors that bind to prevent hypermutation.

Motif	E-Value
GSAGAC	4.2e-972
TCSGTGAA	2.7e-414
ATACTAC	7.5e-301
GCCC	2.1e-199

Table 1. Nucleotide motifs discovered to be enriched in non-hypervariable regions. E-value is a measure of the strength of the signal, values less than 0.0001 are generally considered to be significant [2].

Furthermore, hypermutation rate was found to vary significantly with segment type. As seen in Figure 2, hypermutation rates for D and J segments are stable over the length of the segment. In contrast, hypermutation rates in V segments follow a distinct pattern with 4 peaks. These peaks appear to correspond to previously identified Complementarity Determining Regions(CDR) 1, 2, and 3[4]. This interpretation might suggest that the previously identified CDR1 is actually composed of two smaller CDRs.

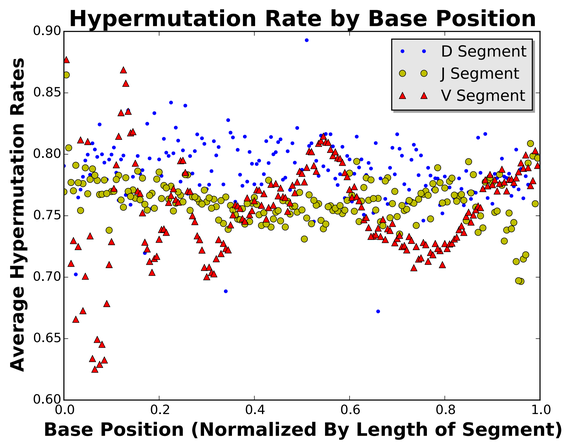

Figure 2. Hypermutation rates by base position for each segment. Rates shown are the average of 13153 sequence reads.

Finally, statistical analysis suggests quantitative differences between the three segments types. As shown in Table 2, the mean running average of hypermutation rates is highest in D segments. Though this difference is not large, coupled with a reduced standard deviation, it may be enough to improve state-of-the-art probabilistic identification of segment boundaries.

Segment Type	Mean Running Average of Hypermutation Rates	Standard Deviation of Hypermutation Rates
V	0.750959	0.149849
D	0.785855	0.075827
J	0.761050	0.132611

Table 2. Statistics for the hypermutation rates of each type of segment. Running averages are of length 9.

Conclusions

There are three main findings of interest for future work. First, a dataset has been built to facilitate analysis of hypermutation rates. Second, 19 motifs have been identified as putative causes of non-hypervariable regions, suggesting potential for targeted experiments. Finally, significant differences in the distribution of hypermutation rates have been observed between segment types. Work is currently being done to combine these observations with probabilistic methods to improve predictions of gene segment boundaries and immune response modeling in general. This will enable future advances in fields as diverse as personalized medicine and anti-viral software.

Acknowledgements

Many thanks to Raunaq Malhotra for the helpful scripts and insightful discussions. Source code and more information is available at www.blengerich.github.io/4-thesis.

Literature Cited

1. Liu, Man and Schatz, David G. 2009. "Balancing AID and DNA repair during somatic hypermutation." *Trends in Immunology* 30:173-181.
2. Jackson et al, "Benchmarking the performance of human antibody gene alignment utilities using a 454 sequence dataset." *Bioninformatics,* 26(24):3129-3130
3. Timothy L. Bailey, "DREME: Motif Discovery in transcription factor ChIP-seq data"", *Bioinformatics*, 27(12):1653-1659, 2011.
4. William E. Paul (2008). *Fundamental Immunology* (6th ed.). Lippincott Williams & Wilkins.