Main text

Statistical inference to optimize anti-SARS-CoV2 antibody design and uncover viral resistance phenotypes

Natalie Dullerud1, Tea Freedman-Susskind 2, Vanessa D. Jonsson3,4

1 Department of Mathematics, University of Southern California, Los Angeles, CA

2 Department of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA

3 Department of Hematology, City of Hope National Medical Center, Duarte, CA

4 Department of Quantitative and Computational Medicine, City of Hope National Medical Center, Duarte, CA

**Abstract**

SARS-CoV2 antibody therapy

**Main**

**satlasso uncovers antibody genome locations that confer better binding**

We developed a statistical model to uncover correlations between anti-SARS-CoV2 antibody sequence data recovered from convalescent COVID-19 patients and SARS-CoV2 neutralization data. In order to obtain a model that could explain antibody neutralization data with a minimal set of residues while taking into account irregular experimental data corresponding to the limit of neutralization assays (Figure 1), we extended the well-known least absolute shrinkage and selection operator (Lasso) (cite). Our model, the saturated Lasso (satlasso), returns a set of amino acid residues and their antibody sequence location that explains viral neutralization data by minimizing error between model and experimental data, penalizing large models, and appropriately weighing errors corresponding to saturated experimental data. A mathematical description of the model can be found in Materials and Methods. Model selection was performed with 5-fold cross validation (cite).

To assess the generalization of our satlasso model, we compare the first xx largest magnitude regressors to experimentally derived data1 sets and find that many antibody residues known to be critical to SARS-CoV2 neutralization are identified (Figure 1).

**Random forest predicts viral infectivity hotspots**

To uncover infectivity hotspots, we applied random forest classifier on data from cite bloom.

**Computing Gibbs free energy models of antibody and SARS-CoV2 mutations**

We develop a biophysical model based on Gibbs free energy of binding derived from energy minimization calculations on structural information to quantify fitness differences in antibody neutralization.

We hypothesize that a virus’ capacity to infect and be neutralized by specific neutralizing antibodies (NAbs) can be approximated by differences in Gibbs free energy of binding associated to de novo point mutations on the spike protein complexed with the ACE2 receptor and antibody structure. To compute fitness landscapes relating to viral replication and antibody neutralization, we apply an empirical force field, Fold X , to evaluate the effect of point mutations on the stability, folding and dynamics on detailed S-protein/ACE2 and S protein/NAb molecular structures.

**Combination antibody design**

The potential for intra patient viral evolution is possible due to antibody selective pressure (cite regeneron, cite mouse model B cells).

To estimate how mutations of the virus might affect new antibody design, we calculated differences in Gibbs free energies between mutations of the virus and antibody designs predicted by our algorithm. We compared these mutations to the mutations that evolved from the Regeneron paper and found x. We compared the mutations found in the Bloom paper,

**Discussion**

Recent advances in the identification and engineering of anti- HIV-1 antibodies have produced a large set of detailed molecular structures and neutralization data generated against a broad panel of HIV-1 strains. Recent computational analysis of antibody neutralization data has been successful in categorizing antibodies with respect to their neutralization activity [5], extracting the identities of RBD residues that are necessary for neutralization [17] and uncovering antibody epitopes [11]. Here, we report a computational methodology that utilizes antibody neutralization data and structural information to construct SAR fitness landscapes and reason about the dynamics of resistance. It consists of the interpretation of neutralization data using statistical inference, the construction of fitness landscapes using computational chemistry and the development of biophysical and mathematical model to capture the dynamics of replication, mutation and selection.

Our statistical model is able to uncover critical residues involved in antibody neutralization and is consistent with recent studies in antibody resistance. It does not identify the entire structural epitope involved in the protein-protein contact. Rather, satlasso identifies an epitope that is involved in the function of the protein-protein interaction, in our case neutralization.

**Methods**

**Data used**

PDB structures were used C105 antibody (PDB 6XCM),

**Random forest generator**

**Satlasso**

We define the saturated least absolute shrinkage and selection operator (satlasso) and formulate it as a convex op- timization problem. We consider that the antibody neutral- ization data X = Xs + Xu is comprised of saturated data Xs , corresponding to IC50s of very resistant viruses, and of un- saturated data Xu (Figure 1). Observe n predictor response pairs (xi,yi) where xi ∈ Rp and y ∈ R. Forming X ∈ Rn×p, X = Xu + Xs with standardized columns, the saturated lasso, (satlasso) is an estimator defined by the following convex optimization problem:

minimize β∈Rp

1 1􏰊n  
λ1 ||yu − Xuβ||2 + λ2 wi|βi| + λ3 max(ys − Xsβ, 0),

to the ordinary least squares problem for subset j. Model se- lection was performed by 5 fold-cross validation. Our results exclude mutations that are located in the signal peptide re- gion, transmembrane and intracellular regions corresponding to HXBC2-numbered residues (1-30), (685-706) and (706-end) respectively. These locations on Env are not exposed to anti- body binding and we assume that are not subject to selective pressure by antibody.

**Dimension reduction and clustering**

**Combined Statistical analysis**

**Data availability**

Original virus infectivity and antibody neutralization data was deposited in ensemble and scRNA-seq data is deposited in the GEO and are available under accession number GSE12345. Exome-sequencing data are deposited in the Sequence Read Archive (SRA) and are available under accession number ABCDEF. Bulk TCR-seq data can be accessed through the ImmuneACCESS database of Adaptive Biotechnologies (URL). All other relevant data are available in extended data tables.

**Code availability**

Satlasso code is available at https://github.com/vdjonsson/satlasso. Gibbs energy analysis code is available at [https://github.com/vdjonsson/gibbs\_fitness](https://github.com/vdjonsson/gibbs_fitnesssc). Code to generate figures for the paper can be found at https://github/com/vdjonsson/paper\_covab.

**References**

**Acknowledgements**

We thank Anthony West and Pamela Bjorkman. This work was supported by an award from the National Cancer Institute of the National Institutes of Health (K12CA001727). N.D. and T.F.S. were supported by the Norton Anonymous donor fund. T.F.S. was also supported by SURF . The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Author Information**

**Affiliations**

**Contributions**

V.D.J. conceived the project. V.D.J., T.S. and N.D. analyzed data. V.D.J., T.F.S and N.D. implemented analytic and computational pipelines. First manuscript draft: V.D.J. and N.D. Final manuscript: V.D.J. and N.D.

**Corresponding authors**

Correspondence to Vanessa D. Jonsson.

**Ethics declarations**

**Competing interests**

**Extended Data**

**Supplementary Information**

Tables: