PREDICTING FUTURE MUTATIONS FOR ESCAPE-RESISTANT VACCINES

With the continuing emergence of potential escape mutants (delta, lambda, omicron, the COVID-19 pandemic has cast doubt on the long term effectiveness of the current vaccine designs. Periodic reformulations similar to seasonal flu vaccines, is problematic for COVID-19 given its high transmissibility, infectious asymptomatic patients, vaccine hesitancy, and potentially high mortality. The COVID-19 situation is not unique: emergent viruses experience diverse selection pressures fostering adaptations via new mutations. The current state of knowledge has no reliable tool to preempt such variants: we do not know when or how new mutants will arise, and the phenotypic impact of future mutations with respect to pathogenicity, transmissibility and cross-reactivity to existing vaccines [1–7]. Thus, there is need of revolutionary conceptual breakthroughs to predict how a viral strain mutates in the wild under realistic selection, allowing escape-resistant vaccine designs.

Unique Aspects, Personnel: To achieve this goal, we formulated the methodological foundations to elicit a deep understanding of the evolutionary dynamics in the sequence/strain space. Our overarching vision, backed by pilot studies over the past year with limited intramural funding from the UChicago Big Ideas incubator, is to computationally interrogate evolutionary patterns underlying the current pandemic and beyond. Since each viral strain is but a single point in a $\approx \times 10^4$ dimensional space (SARS-CoV-2 genome $\approx 3 \times 10^4$ bases), we can never comprehensively explore the state space. Our biology-aware framework mitigates the combinatorial explosion met in naive simulation of future evolutionary trajectories, and makes such estimations manageable on high-performance computing clusters.

We have to-date predicted new mutations on the SARS-CoV-2 spike protein, and shown in in-vitro experiments that these computationally predicted variants express correctly, are functional (bind to the human ACE2 receptor), and some are more resistant to standard neutralization assays compared to the wild type strain. Using data from early 2020 when the pandemic was in its infancy, we could preempt mutations that eventually arose in the delta variant. Applying the same concept to Influenza drift, we consistently out-perform WHO/CDC predictions for vaccine components with respect to how far removed the predictions were from the dominant strain in the future season[8].

Thus, via a cross-disciplinary collaboration between Prof. Ishanu Chattopadhyay[9–12] (mathematical modeling, information theory, machine learning) and Prof. Aaron Esser-Kahn[12–14] (immunology, vaccine science), we envision a radically approach to escape-resistant vaccine design. Beyond predicting likely future mutations in circulating strains, our framework will enable predicting mutations within a single individual as a potential source of novel variant emergence, and also at-scale rank-ordering of sampled strains in animal reservoirs by risk of emergence (a capability well-beyond the state of the art), and eventually spark research into precision interventions in the animal reservoir to preemptively neutralize threats, *before the first human infection*.

#1. More details on what we are going to do?

Justifying Keck Support: Our vision entails risks; we are challenging a prevailing dogma, that future mutations, and variants, of a pathogen are intrinsically random and hence unpredictable. We have sufficient evidence to the contrary, and need Keck's support to validate our tool in well-designed binding, neutralization challenge assays fostering a paradigm shift in how we combat pandemics in future. While we have been turned down recently by NIH (FOA: AI21-035, Application id: 1 R21 AI169352-01), this study can fundamentally change the game, with high future interest from stakeholders.

Budget, Timeline: Conducted over a period of three years costing 1M USD, supporting study personnel (PI time + Postdoctoral time $\approx 33\%$, computational costs ($\approx 10\%$) and experimental costs ($\approx 50\%$).

PROJECT DESCRIPTION

Overview: While the case count for the COVID-19 pandemic has fallen globally with the roll out of multiple vaccines, new variants and potential escape mutants pose a new threat. The current practice has no reliable tool to preempt such emergence: we do not know when new mutants will arise, and the phenotypic impact of future mutations with respect to pathogenicity, transmissibility and cross-reactivity to existing vaccines. A key conceptual barrier is the missing ability to numerically estimate the likelihood of specific future mutations. Currently this likelihood is equated to sequence similarity, which is measured by how many mutations it takes to change one strain to another (the edit distance). In reality, the odds of one sequence mutating to another is a function of not just how many mutations they are apart to begin with, but also how specific mutations incrementally affect fitness. Ignoring the constraints arising from the need to conserve function makes any assessment of the mutation likelihood suspect. In this study we plan to computationally learn these complex and hitherto unknown evolutionary constraints from large sequence databases, enabling us to the chart trajectories of wild pathogens at scale. We propose to experimentally validate our approach in binding and neutralization assays, allowing us to leverage sequence and structural annotation databases, to predict when and how new strains are expected to appear, along with their impact on pathogenicity, and the potential for vaccine escape. Beyond escape-resistant vaccine design, our framework will enable rank-ordering collected animal strains at scale according to their true risk of emergence, making possible pro-active bio-surveillance.

Relevant Efforts: The Big Ideas Generator (BIG) program at the University of Chicago has funded our initial work, with substantial interest going forward.

Peer Groups: Very recently, two articles have explored the possibility of predicting pathogenicity from genomic sequences (Mollentze[15]), and forecasting which amongst observed mutations will dominate the circulating population (Maher *et al.*[16]). While these questions overlap with our framework, our approach is distinct, and vastly more ambitious both intellectually and in scope. Mollentze uses classical sequence similarity; extended to include similarity to human housekeeping genes hoping to identify viruses evading the human immune system more easily, with poor performance (tagging incorrectly all SARS-related coronaviruses as potentially pathogenic), and un-actionable specificity. And, Maher outright assumes mutations to be independent, with SARS-CoV-2 specific manually selected features the authors hack together via machine learning. Importantly, these approaches only aim to predict point mutations, with the gargantuan complexity of tracking a more complete strain (or the RBD fragment) through an ultra-high-dimensional sequence space lying well beyond reach. Thus, even the question if a yet-to-be-seen strain is indeed a valid biological encoding of a virus (which is simpler to determining risk posed by such future variants) cannot answered by our peers, limiting such approaches to analyzing mutations already seen, or strains already collected. Additionally, generalizability and actionability is suspect, given that Maher's features are SARS-CoV-2 specific, and Mollentze's similarity to house-keeping genes might not be universal.

Goals & Methodology: We infer a recursive forest of predictors (the Q-net) that maximally extracts dependency information between mutations, and can preempt new complete strains that have never been seen before, but nevertheless represent a valid genomic sequence. Our framework is generalizable, we do not need to manually curate features that apply to individual viruses, resulting in unprecedented scalability. Our planned goals are:

#2. more details here?

□ **Aim 1: Validate a biologically meaningful metric (the q-distance) for comparison of genomic sequences. (6 months)** Combining novel machine learning, and information theory, we will characterize patterns of mutations from large sequence databases that constrain evolutionary trajectories, to inform a biologically relevant metric of similarity between genomic

sequences. We hypothesize this q-distance to adapt to specific organisms, its background population, and realistic selection pressures. We plan to demonstrate these results in expression and functional assays, showing that smaller q-distance indicates similar phenotypes.

□ **Aim 2: Develop and validate algorithm for preempting possible future variants (18 month)** With tractable function-aware sampling (q-sampling) of the neighborhood of an observed strain in ultra-high-dimensional possibility space, we will preempt: 1) future likely mutations 2) probability of spontaneous jump via specific mutations, and 3) likely variants arising within specific time-frames in the wild. Validate that predicted mutations/strains are biologically plausible, expressing functional proteins – in silico and in laboratory assays, piloted with the spike protein for SARS-CoV-2 and haemagglutinnin (HA) for Influenza A.

□ **Aim 3. Preempt and characterize escape variants. (24 month)** Preempt escape variants, via characterizing future mutations that evade standard antibody neutralization assays, and thus are candidate escape mutants for SARS-CoV-2 and Influenza A.

□ **Aim 4. Evaluate the ability of the proposed tools to define the emergence edge identifying animal strains poised to emerge into humans with high transmissibility/pathogenicity. (24 month)** Piloted with emerging Influenza A strains, we will compare our predictions against CDC-developed Influenza Risk Assessment Tool (IRAT) scores [17] for influenza variants, and aim to characterize all Influenza A sequences (and predicted variants) in public databases. If validated, such a tool will vastly accelerate such risk assessment, cutting down the time required for individual strains from weeks/months to under a second.

#3. Need to discuss the methods

Impact: Tools for adhoc quantification of genomic similarity [18–23] are **not inherently biologically meaningful** – a smaller edit distance between two strains does not necessarily imply that a feasible wild trajectory exists from one to the other. Our algorithms are the first of its kind to learn the **appropriate metric of comparison from data**, without assuming any model of DNA/RNA or amino acid substitution, or a genealogical tree a priori, and is designed to be aware of the impact of the host environment and background epidemiology. A successful completion of this study will have profound impact on biosurveillance strategies, and what we do with the products of such efforts. With risk-wise rank-ordering of newly collected strains, we can better judge pandemic risks, quantify the odds of a particular strain spilling to humans, and estimate its potential to lead to a global pandemic. And for strains already circulating in the human population, we will preempt variants, and their odds of escaping current vaccines.

#4. need to have figure?

Fundraising: To date 85K has been allocated to this projects. Describe what amounts of funding have been committed to this project to date, including institutional funding. Cite pending proposals, amounts requested, sources and expected notification dates. If no other external support will be sought for this project, please explain.One NIH proposal has been turned down, as stated earlier. DARPA BTO, NIAID.

#5. Need 5, please suggest

KNOWLEDGEABLE EXPERTS IN THE FIELD

1. Peter Hraber
Theoretical Biology & Biophysics Group, T-6
Theoretical Division
Los Alamos National Laboratory
PO Box 1663, MS K710
Los Alamos, NM 87545
phone: 505 665 7491
email: phraber@lanl.gov

Dr. Hraber is an expert in theoretical biology and biophysics, focusing in computational immunology, evolution, and statistical genetics, and is well-suited to evaluate the interplay of mathematical modeling, evolutionary dynamics and immunological aspects of the proposed project.

---

2. Patrick Wilson
Assistant Professor
Drukier Institute for Children's Health
Weill Cornell Medicine
1300 York Avenue
New York, NY 10065
phone: 212 746 4111
email: pcw4001@med.cornell.edu

Prof. Wilson is an immunologist with extensive experience in characterization of human immune responses, with definitive work in influenza vaccine designs and B cell biology.

---

3. Balaji Manicassamy
Associate Professor of Microbiology and Immunology
Iowa State University
3-430 Bowen Science Building
51 Newton Rd
Iowa City, IA 52242
phone: 319 335 7590
email: balaji-manicassamy@uiowa.edu

Prof. Manicassamy is an expert in influenza viruses and respiratory pathogens, with extensive experience in reverse genetics and pathogenesis.

4.

5.

REFERENCES

[1] Gou, X., Wu, X., Shi, Y., Zhang, K. & Huang, J. A systematic review and meta-analysis of cross-reactivity of antibodies induced by h7 influenza vaccine. *Human vaccines & immunotherapeutics* **16**, 286–294 (2020).

[2] Hannenhalli, S. & Pevzner, P. Transforming cabbage into turnip.(polynomial algorithm for sorting signed permutations by reversals). dept. of computer science and engineering, penn state university. Tech. Rep., Technical Report CSE-95-004 (1995).

[3] Jean, G. & Nikolski, M. Genome rearrangements: a correct algorithm for optimal capping. *Information Processing Letters* **104**, 14–20 (2007).

[4] Ozery-Flato, M. & Shamir, R. Two notes on genome rearrangement. *Journal of Bioinformatics and Computational Biology* **1**, 71–94 (2003).

[5] Tesler, G. Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences* **65**, 587–609 (2002).

[6] Shao, M. & Lin, Y. Approximating the edit distance for genomes with duplicate genes under dcj, insertion and deletion. *BMC bioinformatics* **13**, S13 (2012).

[7] Fair, J. & Fair, J. Viral forecasting, pathogen cataloging, and disease ecosystem mapping: Measuring returns on investments (2019).

[8] Li, J., Li, T. & Chattopadhyay, I. Preparing for the next pandemic: Learning wild mutational patterns at scale for analyzing sequence divergence in novel pathogens. *medRxiv* (2020). URL https://www.medrxiv.org/content/early/2020/07/20/2020.07.17.20156364. https://www.medrxiv.org/content/early/2020/07/20/2020.07.17.20156364.full.pdf.

[9] Chattopadhyay, I. & Lipson, H. Data smashing: uncovering lurking order in data. *Journal of The Royal Society Interface* **11**, 20140826 (2014).

[10] Chattopadhyay, I., Kiciman, E., Elliott, J. W., Shaman, J. L. & Rzhetsky, A. Conjunction of factors triggering waves of seasonal influenza. *Elife* **7**, e30756 (2018).

[11] Huang, Y. & Chattopadhyay, I. Universal risk phenotype of us counties for flu-like transmission to improve county-specific covid-19 incidence forecasts. *PLoS computational biology* **17**, e1009363 (2021).

[12] Onishchenko, D. *et al.* Reduced false positives in autism screening via digital biomarkers inferred from deep comorbidity patterns. *Science advances* **7**, eabf0354 (2021).

[13] Moser, B. *et al.* Small molecule nf-kb inhibitors as immune potentiators for enhancement of vaccine adjuvants. *ChemRxiv* **10** (2019).

[14] Manna, S., Maiti, S., Shen, J., Du, W. & Esser-Kahn, A. P. Pathogen-like nanoassemblies of covalently linked tlr agonists enhance cd8 and nk cell-mediated antitumor immunity. *ACS central science* **6**, 2071–2078 (2020).

[15] Mollentze, N., Babayan, S. A. & Streicker, D. G. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS biology* **19**, e3001390 (2021).

[16] Maher, M. C. *et al.* Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science translational medicine* eabk3445 (2021).

[17] CDC. Influenza risk assessment tool (irat) — pandemic influenza (flu) — cdc. https://www.cdc.gov/flu/pandemic-resources/national-strategy/risk-assessment.htm. (Accessed on 07/02/2021).

[18] Posada, D. & Crandall, K. A. Modeltest: testing the model of dna substitution. *Bioinformatics (Oxford, England)* **14**, 817–818 (1998).