PREDICTING FUTURE MUTATIONS FOR ESCAPE-RESISTANT VACCINES

The continuing mutation of COVID-19 (delta, lambda, omicron) during COVID-19 pandemic has shown the need for a new type of vaccine designs - one that is as dynamic and nimble as the virus it plans to protect against. Periodic reformulations similar to seasonal flu vaccines, is problematic for COVID-19 given its current rapid rate of mutation coupled with a high transmissibility, infectious asymptomatic patients, vaccine hesitancy, and potentially high mortality. This situation is not unique: emergent viruses experience diverse selection pressures fostering adaptations via new mutations. The current state of knowledge has no reliable tools to preempt such viruses: we do not know when or how new mutants will arise, and how to protect against them (Hum. vacc. & immunotherapeutics 16:286-294,'20),(Global Catas. Bio. Risks 75–83,10.1007/82_2019_179).Thus, there is a need of revolutionary conceptual breakthroughs to predict how a viral strain mutates in the wild under realistic selection, allowing for the design and testing of vaccines before the emergence event.

Unique Aspects: To achieve this goal, we formulated the methodological foundations for 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 driving the current pandemic and beyond. Since each 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. But we don't need to. We reduce the number of combinations by accounting for only those that occur along evolutionary trajectories – making calculation possible 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 predicted variants express correctly, are functional (bind to the human ACE2 receptor), and some are more resistant to antibody binding assays compared to the wild type. Using data from early days of the pandemic, we could preempt mutations that eventually arose in the delta variant. Testing the idea along a longer time-frame, we applied the same concept to Influenza drift. Here, this approach consistently out-performed WHO/CDC predictions for vaccine components with respect to how far removed the predictions were from the dominant strain in the future season (Medarxiv, 10.1101/2020.07.17.20156364).

Personnel: Thus, via a cross-disciplinary collaboration between Prof. Ishanu Chattopadhyay (mathematical modeling, information theory, machine learning) and Prof. Aaron Esser-Kahn (immunology, vaccine science), we envision a radically different approach to escape-resistant vaccine design. Beyond predicting likely future mutations in circulating strains, the goal of this proposal will be to build a platform technology which can be developed/tested toward (1) predicting mutations within a single individual as a potential source of novel variant emergence, and (2) develop a rank-ordering of sampled strains in animal reservoirs by risk of emergence (a capability well-beyond the state of the art). Such methods would form the nucleus of a burgeoning field of precision interventions in the animal reservoir to preemptively neutralize threats, *before the first human infection*.

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 a well-vetted test/design/test loop ultimately 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, including NIH and Biological Tech. office at DARPA.

Budget, Timeline: Conducted over a period of four years costing 1.5M USD, supporting study personnel (PI time + Postdoctoral time $\approx 60\%$, computational costs ($\approx 10\%$) and experimental costs ($\approx 30\%$), with some allocation for travel, and publication funding.

PROJECT DESCRIPTION

Overview: The COVID-19 pandemic, despite multiple vaccines, continues to be an ongoing challenge as new variants and potential escape mutants emerge. The current practice has no tools to predict, let alone preempt such emergence: we do not know when new mutants will arise, and how these mutants will differ in terms of pathogenicity, transmissibility and resistance to current 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 at the beginning, but also how specific mutations incrementally affect fitness. Ignoring the constraints needed 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 vaccine escape.

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

Peer Groups: Very recently, two articles investigated predicting pathogenicity from genomic sequences (Mollentze (PLoS biology 19: e3001390, '21)), and identifying current mutations which might dominate in future (Maher *et al.* (Science trans. med. eabk3445,'21)). 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 to human housekeeping genes hoping to identify viruses evading the human immune system, with limited performance (tagging incorrectly all SARS-related coronaviruses as pathogenic). And, Maher assumes mutations are mutually independent ignoring crucial epistatic and compensatory effects (Curr. opinion in struct. bio. 50: 18–25,'18), combining manually curated SARS-CoV-2-specific putative features via machine learning. Importantly, these approaches aim to predict point mutations, not addressing the next-level challenge of tracking dependent and compensatory mutations throughout a complete strain. Such ultra-high-dimensional sequence space predictions lie well beyond reach of our peers. While useful in hindsight analysis, these approaches cannot yet predict a new strain with an estimated probability, or predict whether it will pose a threat. Our method goes deeper into the fundamental principles underlying viral evolution, using information from each sequence to make strain-specific predictions. Thus, with sufficient data we can track any species, its future mutations and emergence events.

Goals & Methodology: We computationally infer a collection of cross-dependent predictors (the Q-net) that maximally extracts dependency information between mutations & motifs. We can preempt complete strains that have never been seen before, but nevertheless represent a valid genomic sequence. Our framework is general. With no manually curated features for individual viral species, our *sequence only* model lends robust scalability. Our goals (Fig. 1):

□ **Aim 1: Validate a meaningful comparison (q-distance) for genomic sequences. (12 mo, Y1)** Combining novel machine learning, and information theory, we will characterize mutation patterns from large sequence databases (Nuc Acids Res 45:D482–D490, 16) that constrain evolutionary trajectories and reveal selection pressures, to inform a biologically meaningful species-specific adaptive metric of sequence similarity. Using SARS-CoV-2 & Influenza A as model organisms, we will validate the q-distance using past trajectories of dominant strains, showing closer sequences in this metric are more predictive of phenotype than edit distance.

□ **Aim 2: Develop+validate algorithm preempting future variants (24 mo, Y1-Y2)** 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 specified time-frames in the wild. Validate that predicted mutations/strains are biologically plausible, expressing functional proteins, both in silico and in laboratory assays, piloted with the spike protein for SARS-CoV-2 and Hemagglutinin (HA) for Influenza A. In each case, we will predict 10^3 in silico sequences of each protein, validate folding using standard software (Nat Struct & Mol Bio 28:869–870, '21), then screen top-candidates for in-vitro assays.

□ **Aim 3. Preempt and characterize escape variants. (36 mo, Y2-Y4)** 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. Taking the 1,000 pseudo-virus expressing proteins , we will down-select for proteins that (1) bind either ACE-2 or Sialic Acids (est 100), and then (2) escape the binding of panels of sera from convalescent patients and vaccine recipients (estimate 10-50). These ≈10-50 protein sequences will then be encoded as model mRNA vaccines and their antibody responses evaluated. The success metric here is to show that our tools significantly reduce escape odds. This characterization will also feed-back to fine-tune q-sampling to directly predict, escape mutants with high probability.

□ **Aim 4. Define the emergence edge identifying animal strains poised to emerge into humans with high transmissibility/pathogenicity. (24 mo, Y3-Y4)** Piloted with emerging Influenza A strains, compare predictions against CDC-developed Influenza Risk Assessment Tool (IRAT) scores for flu variants, and characterize all Influenza A sequences (and predicted variants) in public databases. If validated, we will vastly accelerate risk assessment, cutting down the time required for individual strains from weeks/months to $<$ 1sec.

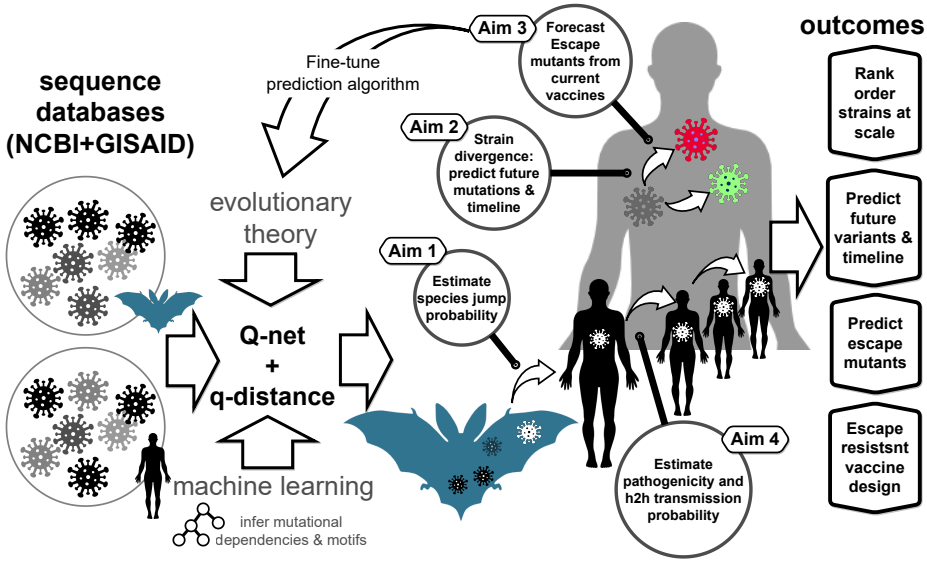

Fig. 1. Conceptual framework and outcomes: General approach to scalably learn mutational dependencies to predict future escape variants and proactive surveillance

Impact: Adhoc quantification of genomic similarity (Bioinformatics 14:817–818, '98) (Elife 3: e03568, '14) is not meaningful – a smaller edit distance between two strains does not imply a feasible wild trajectory from one to the other. Our algorithms are a first to learn the **appropriate metric of comparison**, without assuming any model of DNA/RNA substitution, or a genealogical tree a priori, and are designed to be aware of the impact of the host environment and background epidemiology. This study, if successful, 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 1) better judge pandemic risks, 2) quantify the odds of a particular strain spilling to humans, and 3) estimate its potential to lead to a global pandemic. And for strains already circulating in the human population, we can potentially 4) preempt variants, 5) their timeline of emergence, 6) their odds of escaping current vaccines, and ultimately 7) design escape-resistant vaccines.

Fundraising: To date ≈100K allocated (BIG funding + PI development fund). We have pending proposals at NSF (PIPP, 1M USD, summer '22) and NIH (R21, 400K, Fall '22).

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. Geoffrey Lynn
Senior Vice President, Synthetic Immunotherapies at Vaccitech
1812 Ashland Ave
Baltimore, MD 21205
Bethesda, Maryland, United States
email: Geoffrey.Lynn@vaccitech.us

Dr. Lynn is an expert in synthetic chemistry and cellular immunology with research interest in precision immunotherapies for complex diseases.

5. Danny Altmann
5S5C Hammersmith Hospital
Hammersmith Campus
72 Du Cane Rd, London W12 0HS, United Kingdom
phone: +44 (0)20 3313 8212
email: d.altmann@imperial.ac.uk

Prof. Altmann is an a well-known immunologist with research interest in the immunology of infectious disease including severe bacterial infections.

REFERENCES

H. Bagdonas, C. A. Fogarty, E. Fadda, and J. Agirre. The case for post-predictional modifications in the alphafold protein structure database. *Nature Structural & Molecular Biology*, 28(11):869–870, 2021.

CDC. Influenza risk assessment tool (IRAT) https://www.cdc.gov/flu/pandemic-resources/national-strategy/risk-assessment.htm. Dept. of HHS Report (no DOI). https://www.cdc.gov/flu/pandemic-resources/pdf/CDC-IRAT-Virus-Report.pdf. (Accessed on 07/02/2021).

J. Fair and J. Fair. *Viral Forecasting, Pathogen Cataloging, and Disease Ecosystem Mapping: Measuring Returns on Investments*, pages 75–83. Springer International Publishing, Cham, 2019. ISBN 978-3-030-36311-6. doi: 10.1007/82_2019_179.

X. Gou, X. Wu, Y. Shi, K. Zhang, and J. Huang. A systematic review and meta-analysis of cross-reactivity of antibodies induced by h7 influenza vaccine. *Human vaccines & immunotherapeutics*, 16(2):286–294, 2020.

E. L. Hatcher, S. A. Zhdanov, Y. Bao, O. Blinkova, E. P. Nawrocki, Y. Ostapchuck, A. A. Schäffer, and J. R. Brister. Virus variation resource – improved response to emergent viral outbreaks. *Nucleic Acids Research*, 45(D1):D482–D490, nov 2016. doi: 10.1093/nar/gkw1065.

J. Li, T. Li, and I. Chattopadhyay. Preparing for the next pandemic: Learning wild mutational patterns at scale for analyzing sequence divergence in novel pathogens. *medRxiv*, 2020. doi: 10.1101/2020.07.17.20156364.

M. C. Maher, I. Bartha, S. Weaver, J. Di Iulio, E. Ferri, L. Soriaga, F. A. Lempp, B. L. Hie, B. Bryson, B. Berger, et al. Predicting the mutational drivers of future sars-cov-2 variants of concern. *Science translational medicine*, page eabk3445, 2021.

N. Mollentze, S. A. Babayan, and D. G. Streicker. Identifying and prioritizing potential human-infecting viruses from their genome sequences. *PLoS biology*, 19(9):e3001390, 2021.

R. A. Neher, C. A. Russell, and B. I. Shraiman. Predicting evolution from the shape of genealogical trees. *Elife*, 3:e03568, 2014.

D. Posada and K. A. Crandall. Modeltest: testing the model of dna substitution. *Bioinformatics (Oxford, England)*, 14(9):817–818, 1998.

J. F. Storz. Compensatory mutations and epistasis for protein function. *Current opinion in structural biology*, 50:18–25, 2018.