Novelty Search Drives Pathogen Diversification

by

Brandon Ely1, Winston Koh2, Tasmina Hassan2, Anh Pham2, Eamen Ho2, and Weigang Qiu1,2,\*

1Department of Biology, Graduate Center, City University of New York, USA

2Department of Biological Sciences, Hunter College, City University of New York, USA

3Department of Physiology and Biophysics & Institute for Computational Biomedicine, Weill Cornell Medical College, New York, USA

\* Correspondence: Weigang Qiu <wqiu@hunter.cuny.edu>

# ABSTRACT

Cellular or viral surface antigens of microbial pathogens evolve rapidly and adaptively as a result of host-pathogen coevolution. The “strain theory”, for example, posits that antigen variants coexisting in (or successively emerging from) a pathogen population are antigenically distinct from one another to confer immune escape. Novel antigen variants may also emerge by enhancing pathogen transmissibility through better binding with host cell receptors. Fore-sighted pandemic prevention by predicting the course of adaptive sequence evolution of pathogen antigens, however, remains an unfulfilled challenge. Here we simulated adaptive sequence evolution on computationally and empirically derived fitness landscapes as a way of forecasting the emergence of new antigen variants. Specifically, we designed and implemented three evolutionary algorithms (the objective, novelty, and combination searches) and evaluated their performances on simulated and empirical fitness landscapes. The simulated fitness landscapes consisted of binary strings representing antigen sequence variants, the fitness of which were specified by models of additive or epistatic interactions among mutation sites. Empirical fitness landscapes were derived from studies of deep mutation scans of pathogen antigens, including the GP160 envelop protein of human immunodeficient virus (HIV), the neuraminidase (HA) of human influenza virus, the GB1 protein of Streptococcal bacteria, and the receptor-binding domain (RBD) of SARS-CoV-2 spike protein. On simulated as well as empirical landscapes, novelty and combination searches outperformed the objective-only search by evolving sequence variants that reached globally optimal fitness peaks more often and more quickly. The advantage of novelty and combination searches increased with the ruggedness of fitness landscape, a measure of epistatic complexity. We conclude that pathogen variants emerge rapidly through a process akin to the novelty search algorithm, by which a pathogen population overcomes low-fitness valleys through hyper-mutability, weak selection (e.g., in immune-compromised hosts), or a combination of such genetic and population processes. The efficacy of novelty search algorithms in reaching global peaks on rugged fitness landscapes makes it a promising tool to predict the emergence of new pathogen variants, provided that a fitness landscape be accurately measured or computed and that the novelty search algorithms be optimized and implemented at a large scale.

# SIGNIFICANCE STATEMENT

(to do)

# INTRODUCTION

(to do)

# Data, Methods and Algorithms

## Simulated fitness landscapes

* Random
* Additive
* NK
* Mt Fuji
* polynormial

## Empirical landscapes

* GB1 4-site landscape
* GP160 of HIV
* HA of influenza
* RBD of SARS-CoV-2

## Evolutionary algorithms

* Objective-only
* Novelty-only
* Objective-Novelty combination

## Data and code availability

* Github: “nov-search”

# RESULTS

## Simulated landscapes

(to do)

## Performances on simulated landscapes

(to do)

## Empirical landscapes & performance measures

(to do)

# DISCUSSION

## Implications

(to do)

## Insights

(to do)

## Future directions

(to do)

# Author contributions

(to do)

# Acknowledgments

Brandon Ely is supported in part by the Doctoral Program in Biology of the Graduate Center, the City University of New York.

# References

# Figure legends

## Fig 1. Simulated fitness landscapes

(to do)

## **Fig 2.** Performance of search algorithms on simulated landscapes

(to do)

## **Fig 3.** Empirical fitness landscapes

(to do)

## Fig 4. Performances of search algorithms on empirical fitness landscapes

(to do)

## Fig 5. Gene duplications and losses on the linear plasmid lp54

(to do)

# List of supplementary materials

Table S1. (to do)

Table S2. (to do)

Table S3. (to do)

Fig S1. (to do)

Fig S2. (to do)

Fig S3. (to do)