Novelty Search Promotes Antigenic Diversity in Microbial Pathogens

Brandon Ely[1], Winston Koh[2], Eamen Ho[2], Tasmina M. Hassan[2], Anh V. Pham[2], and Weigang Qiu[1,2, 3]

[1]The Graduate Center & [2]Hunter College of City University of New York; [3]Weil Cornell Medical College

Abstract

Driven by host–pathogen coevolution, cell surface antigens are often the fastest evolving parts of a microbial pathogen. The persistent evolutionary impetus for novel antigen variants suggests the utility of novelty-seeking algorithms in predicting antigen diversification in microbial pathogens. In contrast to traditional genetic algorithms maximizing variant fitness, novelty-seeking algorithms optimize variant novelty. Here, we designed and implemented three evolutionary algorithms (fitness-seeking, novelty-seeking, and hybrid) and evaluated their performances in 10 simulated and 2 empirically derived antigen fitness landscapes. The hybrid walks combining fitness- and novelty-seeking strategies overcame the limitations of each algorithm alone, and consistently reached global fitness peaks. Thus, hybrid walks provide a model for microbial pathogens escaping host immunity without compromising variant fitness. Biological processes facilitating novelty-seeking evolution in natural pathogen populations include hypermutability, recombination, wide dispersal, and immune-compromised hosts. The high efficiency of the hybrid algorithm improves the evolutionary predictability of novel antigen variants. We propose the design of escape-proof vaccines based on high-fitness variants covering a majority of the basins of attraction on the fitness landscape representing all potential variants of a microbial antigen.

Evolutionary Algorithms

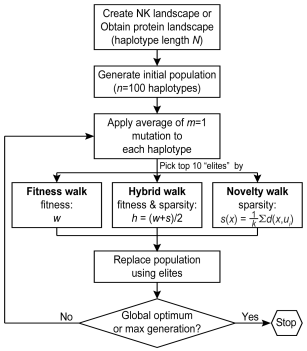

Fig 1 – Evolutionary Alogorithms
For evolution on a simulated landscape, an initial population of antigen variants (with n individuals) were randomly generated. Elites were defined by either the fitness score (fitness walk, left), the novelty score (novelty walk, right), or a linear combination of fitness and novelty scores (hybrid walk, middle). Novelty of a haplotype (x) is measured by sparsity, calculated as average distance to its k nearest neighbors (u).

Human influenza NA Empirical Landscape

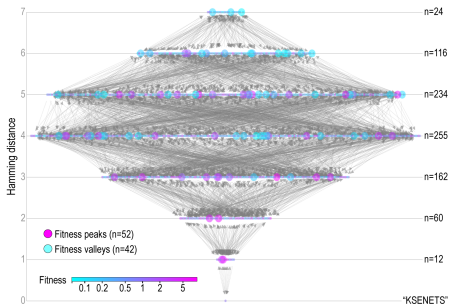

Fig 4 – Nueraminidase (NA) fitness landscape Landscape for 7 epitope sites from HK19 genetic background (n = 864 nodes). Wildtype epitope sequence on bottom level (hamming distance = 0).

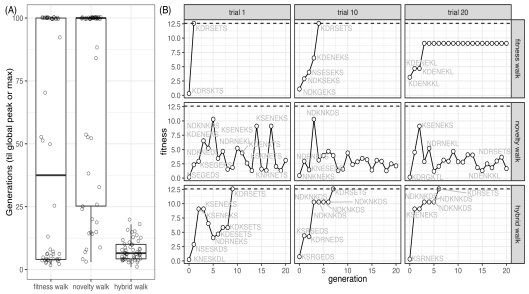

Fig 5 – Evolutionary walks on the NA landscape
(A) Number of generations when a population (represented by a data point) of 100 NA 7-amino acid epitopes reached either the global peak or the maximum generation limit (g = 100). Evolutionary walks were repeated 50 times for each algorithm. **(B)** Fitness values of the fittest individual over generations from three randomly chosen independent walks.

Streptococcal GB1 Empirical Landscape

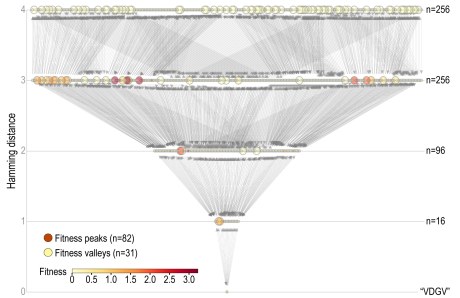

Fig 6 – Subset of GB1 fitness landscape
Nodes (n = 625) represent haplotypes. The full GB1 landscape (not shown) consists of 20^4 = 160,000 haplotypes, 6409 fitness peaks, and 37,089 fitness valleys.

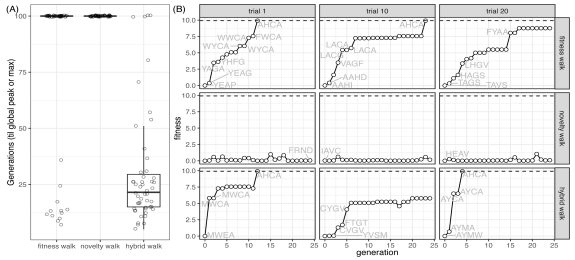

Fig 7 – Evolutionary walks on the full GB1 landscape
(A) Number of generations when a population (represented by a data point) of 100 GB1 4-amino acid peptides reached either the global peak or the maximum generation limit (g = 100). Evolutionary walks were repeated 50 times for each algorithm. **(B)** Fitness values of the fittest individual over generations from three randomly chosen independent walks.

NK Simulated Landscapes

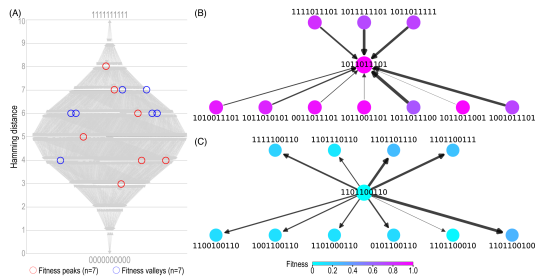

Fig 2 – A Simulated NK Landscape
(A) Multipartite renderings of a simulated fitness landscape using the NK model of mutation interactions (N =10, K = 2). Nodes (n = 1024) are genotypes represented by 10-bit binary strings, edges (n = 5120 in total) connect two genotypes separated by a Hamming distance of d = 1. Nodes separated from the global fitness peak (B) and valley (C) by 1 degree.

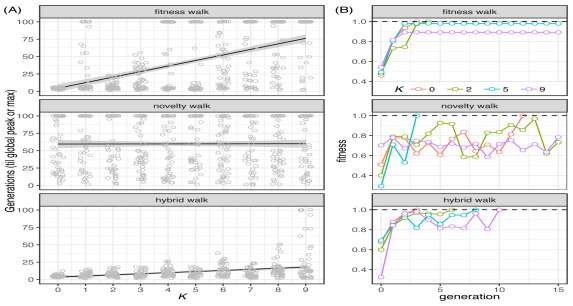

Fig 3 – Evolutionary walks on NK landscapes
(A) Simulation of adaptive evolution of a microbial population on ten NK landscapes. Each point represents the number of generations (y-axis) either when the population reached the global fitness peak (g < 100) or when the simulation reached the generation limit (g = 100). **(B)** Fitness values of the fittest individuals at each generation.

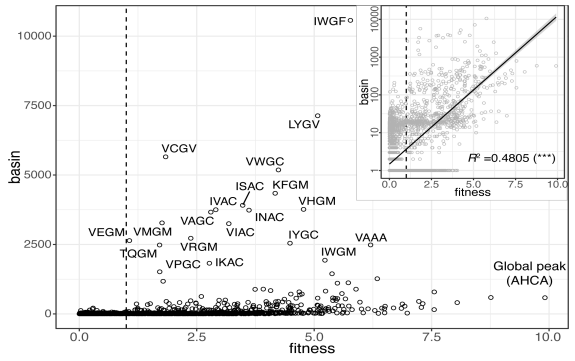

Fig 8 – Local optimal network
A plot of 6409 fitness peaks on the full GB1 land- scape, with the position of a peak (a dot) represented by its fitness (x-axis) and size of basin of at- traction (y-axis). A dashed line marks the fitness value of the wildtype ("VDGV", w = 1). Top 20 haplotypes with basin size >1500 are labeled.