EVALUATING THE RUGGEDNESS OF PROTEIN FITNESS LANDSCAPES

MAHAKARAN SANDHU, ADAM MATER, COLIN JACKSON*

Abstract. This is the abstract.

1 Introduction

Relevance/importance
Similar work
Aims
Outline

2 Results

2.1 Explanation of the metrics.

(1) Number of local maxima.
(2) Random walk methods
(3) Roughness to slope ratio
(4) Fourier expansion

2.2 Validation on Synthetic Landscapes.

Explanation of synthetic landscapes used.

2.2.1 *Accounting for sequence length.*

In real-world empirical datasets there is a need to be able to measure ruggedness for different lengths N and amino acid/nucleotide alphabets A. Szendro and colleagues[1] effectively circumvented this problem for binary alleles A and few N by calculating ruggedness metrics systematically over subgraphs of sequence space; empirical protein datasets have far larger values of $|A|$ and N, and this approach quickly becomes computationally intractable (see SI Subgraph Problem). Therefore, we sought alternative approaches. The idiosyncrasies of the individual metrics called for tailored normalisation approaches; we address each individually below.

Number of local maxima. N_{max} is a global measure of a landscape's ruggedness that is dependent not only on ruggedness but also N and A[1,2]. As N and A increase, N_{max} also increases, despite constant ruggedness (proof?/ref? see SI). The maximum number of N_{max} is given by $\max(N_{max}) = \frac{|A|^N}{N+1}$ (see SI Properties of fitness landscapes). If we assume that underlying ruggedness is constant for increasing N, A, then a straightforward normalisation strategy is to divide the calculated N_{max} by $\max(N_{max})$ given N, A. We tested this strategy for NK and RMF[3] landscapes.

Random walk methods. These should be really easy because of ergodic theorem etc, and because this is a local property

r/s ratio. This will be hard

Fourier. this may also be hard (although when we distill down to 'percent contribution' of higher order terms it doesn't really seem like it should be that much more difficult.[4]

2.2.2 *Accounting for fitness scale*

2.2.3 *Accounting for non-linearity*

2.2.4 *Accounting for incomplete datasets/sampling*

2.3 Validation on Real Datasets

3 Discussion

Discussion of properties and limitations of the metrics
Discussion of other metrics and future work
Discussion of surrounding literature (e.g. global epistasis, minimum epistasis etc)
Implications for evolution and directed evolution

Implications for ML

The assumption that ruggedness is an intensive property of fitness landscapes (i.e. ruggedness is constant for arbitrary partitions of the landscape) is valid only in the case of statistically isotropic landscapes; in anisotropic landscapes, ruggedness may differ in different regions of the landscape. In an anistotropic landscape, we would expect the distribution of N_{max}) to also be anisotropic. Using N_{max} it may therefore be possible to map out regions of high ruggedness, and use dimensionality reduction techniques (such as those proposed by McLandlish Evolution. 2011 Jun; 65(6): 1544–1558.) to visualise the distribution of ruggedness in an empirical landscape.

4 Methods

5 Supporting Information

5.0.1*Ruggedness itself*

5.1 Properties of fitness landscapes

Property 1. ($\max N_{max}$ *as a function of* A, N*)* The probability P_m that a given sequence is a local maximum is the probability that it has higher rank-order than any of its Q one-mutant neighbours[2,5,6]

$$P_m = \frac{1}{Q+1} = \frac{1}{N(|A|-1)+1} \tag{1}$$

It then follows that the maximum possible number of local optima is given by

$$\max N_{max} = \frac{|A|^N}{N(|A|-1)+1} \tag{2}$$

where $|A|^N$ is the expression for the size of the sequence space $|S|$. This result is applicable to any HoC landscape (is it? ref? proof?). In the HoC case, we assume that underlying ruggedness is maximal, even though the dimensionality of the space changes with A and N; how N_{max} changes with K is mathematically complex (cite cite cite). While we do not offer a proof for this conjecture, we appeal to intuition (below).

Statistical isotropy, homogeneity, ruggedness and N_{max}*.* Suppose that a HoC landscape is statistically isotropic and homogeneous. This implies that any equally-sized connected subgraph/partition $P \subseteq \mathcal{G}(V, E)$ of the fitness graph will have the same expectation of local maxima, $\mathbb{E}(N_{max})$, i.e. the density of local maxima is constant throughout the landscape, $\rho(N_{max})$. This further implies that the naive probability that any vertex V is a local maximum is identical throughout the landscape. This intuition follows from Property 1. We might venture to say that the *ruggedness* with respect to N_{max} of the landscape is uniform throughout the landscape. (One caveat is that even though the landscape is isotropic in terms of N_{max}, the magnitude difference in different regions of the landscapes may not be – this may be evaluated by taking the second-order calculation of the average magnitude difference of local maxima with respect to 1-mutant neighbours and then taking the distribution of the calculated quantity – a complex distribution would indicate anisotropy).

However, it is unclear how we might think about comparing ruggedness in differently-sized landscapes, specifically, different $|A|$ and N. Supposing statistical isotropy and homogeneity, we would have to scale the partition $P \subseteq \mathcal{G}(V, E)$ by some factor that takes into account the dimensionality of the space. From

a naive graph-theoretic perspective, if the density of N_{max} and the degree distribution is constant, adding new nodes to a graph ought not to change $\rho(N_{max})$. The total number of maxima will increase, but we can easily obtain $\rho(N_{max})$ by normalising for the increased number of nodes. This simplistic case, however, does not hold in a Hamming graph, where the degree distribution changes as well as the number of nodes. Thus, normalising for constant $\rho(N_{max})$ must take into account the changed degree distribution. In particular, the probability that any given vertex is a local maxima changes as $\frac{1}{N(|A|-1)+1}$. Therefore, as N and $|A|$ increase, the probability decreases as a linear function of both N and $|A|$. However, the total number of maxima increases dramatically as $|A|^N$. Therefore, to satisfactorily compare ruggedness between different sized landscapes, it seems necessary to correct for the probability decrease, as well as the dramatic increase in space size. This result allows us to compute $\rho(N_{max})$ for different sized HoC landscapes. To compare ruggedness values for non-HoC landscpaes (i.e. most real landscapes), we calculate the ratio $N_{max}^{real}/N_{max}^{HoC}$, giving us a proxy for ruggedness as a fraction of total possible ruggedness with respect to N_{max} given $|A|$ and N; through this relation, we can compare ratios for differently-sized landscapes because we know how HoC N_{max} can be compared between differently-sized landscapes.

Is the above argument valid? Ask J. Krug?

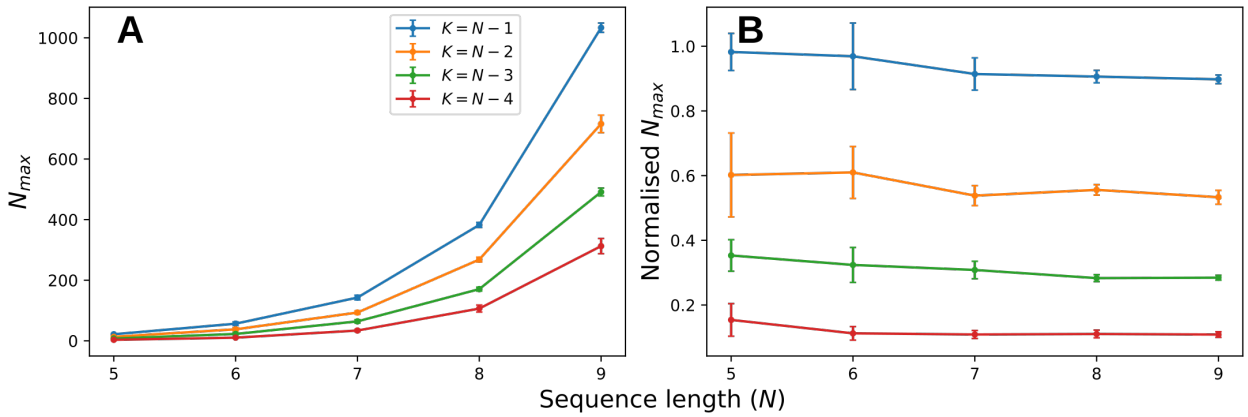

FIGURE 1. **(A)** N_{max} at increasing sequence lengths N and different K values (legend); K values are defined in terms of subtraction from N to give roughly similar ruggedness N/K values; $K = N - 1$ gives the maximally-ruggedn HoC case. **(B)** Normalised N_{max} values with ruggedness correction. See Methods for detailed normalised procedure. Differences within K values are insignificant; between K values, differences are significant. We note that there is a non-linear relationship between increasing K and normalised N_{max}, the exact nature of of which is outside the scope of this work and needs to be checked. **(C)** This will be un-normalised changes in Nmax as alphabet increases.

We further note that N_{max} as a metric loses information regarding the isotropy or lack thereof of the fitness landscape; it assumes that the landscape is isotropic. To circumvent this limitation, we have implemented the average height of fitness maxima.

5.2 Subgraph Problem

Definition 1. (Sequence space subgraphs) A subgraph of size $m < L$ is the set of all combinations of the alleles at m of the L loci at a fixed state of the remaining $L-m$ 'background' loci. There are $\binom{L}{m}$ subsets of m loci and (for the bi-allelic case) 2^{L-m} backgrounds; and thus the total number of subgraphs is $2^{(L-m)}\binom{L}{m}$.

If there are A possible alleles, then there are A^{L-m} backgrounds, and the total number of subgraphs is $A^{(L-m)}\binom{L}{m}$; each subgraph will have A^m sequences, for a total of $A^m A^{(L-m)}\binom{L}{m} = A^L$ sequences (Krug, J. and de Visser, JAG, personal communication).

Problem Consider a protein of length N, where N is nontrivially large (e.g. $N > 100$). For this example, let $N = 100$. While the sequence space S has size A^N, we assume here that we are dealing with an incomplete experimental dataset $D \in S$. Let the the structure of mutations in D be random, i.e. any given position $p_i \in N$ will be found mutated at some sequence $d \in D$. We desire to generate a set of subgraphs Σ with size $m = 4$ from the experimental dataset D.

Solution 1. (Non-heuristic sorting algorithm) Generating all possible subgraphs Σ requires enumerating a ruleset for membership in any given subgraph $\sigma \in \Sigma$, and then performing a search through the dataset D for sequences d that fulfill the membership ruleset for a given $\sigma \in \Sigma$. In our example, the size of Σ is $A^{(L-m)}\binom{L}{m} = 20^{(100-4)}\binom{100}{4} = 20^{96} \times 3921225$ (here $L = N$), which is clearly intractable, even if only generating empty subgraphs containing only the ruleset. Note that if this program were run on an incomplete experimental dataset D, many of the subgraphs would be empty.

Solution 2. (Heuristic sorting algorithm) Generating an incomplete subgraph set Σ^* with valid but incomplete subgraphs σ^* can be acheived by enumerating an incomplete set of rulesets $r^* \in R$ that correspond to $\sigma \in \Sigma$, and then performing a search through the dataset D for sequences d that fulfill some $r \in r^*$. This is a tractable program when $\text{size}(r^*) << \text{size}(R)$.

References

[1] Ivan G Szendro et al. "Quantitative analyses of empirical fitness landscapes". In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.01 (Jan. 2013), P01005.

[2] Stuart A. Kauffman. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford University Press, 1993.

[3] Johannes Neidhart, Ivan G Szendro, and Joachim Krug. "Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model". In: *Genetics* 198.2 (Aug. 2014), pp. 699–721.

[4] Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. "Exact results for amplitude spectra of fitness landscapes". In: *Journal of Theoretical Biology* 332 (Sept. 2013), pp. 218–227.

[5] Stuart Kauffman and Simon Levin. "Towards a general theory of adaptive walks on rugged landscapes". In: *Journal of Theoretical Biology* 128.1 (Sept. 1987), pp. 11–45.

[6] Noah A. Rosenberg. "A sharp minimum on the mean number of steps taken in adaptive walks". In: *Journal of Theoretical Biology* 237.1 (Nov. 2005), pp. 17–22.