Inference of a large empirical fitness landscape from high-throughput sequencing of the *E. coli lac* promoter

J. Otwinowski[1] and Ilya Nemenman[2]

***Short Abstract* — We infer a large fitness landscape from high-throughput sequence data from the E. coli lac promoter region with ~200k sequences. The sequences are associated with measurements of transcriptional activity. Utilizing multivariate regression and L1 regularization (LASSO), we find the best linear and quadratic approximations to fit the data. We find the fitness landscape to be largely smooth and additive, with a small amount of synergistic epistasis. Our method also reveals the locations of binding sites, and their interactions without any prior knowledge and without any difficult optimization steps.**

***Keywords* — Fitness landscape, lasso, CRP-RNAP, lac promoter, epistasis**

PURPOSE

The relationships among genotype, phenotype and fitness are often complex and difficult to untangle. Because it is a map from a high dimensional space to a real valued number, it known sometimes as the fitness landscape. One can measure the fitnesses of different phenotypes or genotypes and to learn large scale properties of the landscape such as the amount of epistasis [1,2], the presence of stabilizing selection [3], and the reproducibility of evolutionary paths [4, 5]. Pitt et al [6] have measured the fitness landscape of ~107 RNA sequences, and is to our knowledge the most detailed fitness landscape measured. Here we quantify a genotype-phenotype map using data from high-throughput sequencing and by applying multivariate linear regression with regularization. The data consists of mutanegized transcriptional regulatory sequences from the E. coli lac promoter [7]. In total ~200,000 lac promoters were mutagenized in a 75 bp region containing the cAMP receptor protein (CRP) and RNA polymerase (RNAP) binding sites (-75:-1) over six experiments. Each sequence has an assigned bin corresponding to different fluorescence levels which indicate transcriptional activity. The relationship between transcriptional activity and fitness depends on the environment, which in this case is the presence or absence of cAMP or lactose. However the fitness, or the growth rate, and the transcriptional activity to enable the metabolism of lactose are likely to be correlated in the presence of large amounts of lactose, and anti-correlated in the presence of catabolite suppression mediated by cAMP. Therefore, we consider transcriptional activity of the promoter as a direct measure of its fitness.

[1]Department of Physics, Emory University, Atlanta, GA 30322. E-mail: jotwino@emory.edu

[2]Departments of Physics and Biology, Emory University, Atlanta, GA 30322. E-mail: ilya.nemenman@emory.edu

METHODS AND RESULTS

We regress the fitness on the presence of individual mutations assuming no epistasis among them and then assuming pairwise epistasis. We estimate the fit error through cross-validation and other tests. We find the fitness landscape to be largely smooth and additive, with a small amount of antagonistic epistasis. Non-epistatic contributions to fitness account for about half of the variance in the data. While pairwise epistatic interactions account for ~10% or less, their effect is statistically significant. We believe this to be the first estimation of the effects of epistasis among many individual nucleotide mutations over a large portion of a regulatory region. We infer the landscape for different environmental conditions (where lac expression is needed, and where it is not), and we find changes in the specificity of RNAP binding. We show that, compared to random sequences, the wild type provides an efficient regulation of transcription in both environments. Our method reveals the locations of binding sites, and their interactions without biophysical modeling and without any difficult optimization steps.

REFERENCES

[1] I.G. Szendro, M.F. Schenk, J. Franke, J. Krug, and J. de Visser. Quantitative analyses of empirical fitness landscapes. arXiv preprint arXiv:1202.4378 (2012).

[2] M Costanzo et al. The genetic landscape of a cell. Science, 327(5964):425-31, 2010.

[3] Ruth G Shaw and Charles J Geyer. Inferring fitness landscapes. Evolution; international journal of organic evolution, 64(9):2510-20, September 2010.

[4] Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. Nature, 445(7126):383-6, January 2007.

[5] Daniel M Weinreich, Nigel F Delaney, Mark A Depristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science (New York, N.Y.), 312(5770):111-4, April 2006.

[6] Jason N Pitt and Adrian R Ferré-D'Amaré. Rapid construction of empirical RNA fitness landscapes. Science (New York, N.Y.), 330(6002):376-9, October 2010.

[7] Kinney, J.B., Murugan, A., Callan, C.G. & Cox, E.C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. PNAS (2010).