No Pain, No Gain:
Relating information gain and selective loads

Kerr Lab Meeting (4/7) Talk Summary

Introduction

The genome carries information akin to a blueprint for constructing an organism and regulating its form, function, and behavior. Genetic information is what is transmitted across generations, and, through mutation and selection, this information is that which evolves. But getting information into the genome is "no pain, no gain." Populations cannot gain genetic information about their environment (i.e., adapt) without paying the Darwinian cost of individuals dying or failing to reproduce.

Previous theory, first presented by Haldane and Kimura, has shown that the rate of genetic information acquisition is bounded by the "substitutional load". Substitutional load can be interpreted as the cumulative depression of population growth rate during selection relative to its optimal growth rate. Or, in other terms, the number of less-favorable individuals that are removed from the population by natural selection. Basically, it follows from this result that in order to put each bit into the genome, half of the population must fall victim to selection.

Theoretical Background

Briefly, the relationship between substitutional load and information gain that was first communicated by Haldane and Kimura is as follows:

Consider a single locus with two alleles, A_0 and A_1, where the respective relative fitnesses are 1 and $1-s$. Let the frequency of allele A_0 at time t be denoted $p_0(t)$, and similarly for the frequency of A_1. We can consider allele A_0 to be a superior allele that emerges in the population at time $t=0$ with initial frequency $p_0(0)$ and is subsequently selected to fixation.

Let us define the *substitutional load* as the cumulative depression in population fitness that is accrued over the course of the substitution of allele A_1 by allele A_0. This can be expressed as:

$$L = \int_0^\infty s p_1(t) dt$$

This integral for the value of the substitutional load can be shown to evaluate to

$$L = -\log p_0(0)$$

Now, let us define the *information gain* in the population as the relative entropy, or Kullback-Leibler divergence, between the allele frequency distributions before and after selection:

$$I = D(\mathbf{p}^{(0)} || \mathbf{p}^{(\infty)}) = \sum_i p_i(\infty) \frac{p_i(\infty)}{p_i^{(0)}}$$

This expression for the information gain can be shown to evaluate to

$$I = -\log p_0(0)$$

Thus, the amount of information accrued during selection is equal to the substitutional load incurred by selection. In other words, the number of bits encoded by selection is proportional to the fold decrease in fitness the population must pay while selecting for those bits. Interestingly, these quantities depend only on the initial frequency of the superior allele and are independent of the strength of selection. This is Haldane and Kimura's key result.

It is of interest to extend this analysis to investigate the relationship between selection and information in more sophisticated evolutionary scenarios. Now consider a generalization of the above model, where a given locus has m segregating alleles, and each allele A_i is associated with its own selection coefficient, which may now vary according to frequency dependence, $s_i(\mathbf{p})$. Let allele A_0 be the most fit allele (for notational convenience). Information gain, I, is defined as before, and substitutional load is now defined assuming an additive contribution of each allele to the load:

$$L = \sum_{i=0}^{m} \int_0^\infty s_i(\mathbf{p}) p_i(t) dt$$

I have shown that Haldane and Kimura's result continues to hold in this general model, given certain assumptions. That is, the information gain is equal to substitutional load, and these quantities are independent of all selection coefficients and frequency dependence relationships — so long as the optimal allele has a constant selection coefficient $s_0 = 0$ (i.e., no frequency dependence in the fitness of the optimal allele). However, the exact values of the information gain and substitutional load, while equal/proportional, may not equal $-\log p_0^{(0)}$ in cases where frequency dependence prevents the optimal allele from fixing. If the initially optimal allele experiences frequency dependence, then there is no general relationship between load and information. In this case, substitutional load, as defined, may or may not diverge depending on the nature of frequency dependence, and it is worth questioning whether the classical definition of substitutional load has meaning in such cases.

Experimental Validation

This theory clarifies a critical relationship between two of evolutionary biology's most fundamental quantities - long term growth rates and genetic information. However, this relationship has not yet been explored empirically. It would be valuable to measure how close actual living populations come to realizing the predicted parity between information gain and subsitutional load in order to gain insights about how the theory should be extended to account for any observed deviations from the theoretical predictions.

I am beginning work with Peter and Olivia to test these predictions in bacterial evolution experiments. We will be competing *E. coli* β-lactam resistant TEM-1 variants in the presence of a drug. Initial experiments will involve 2-4 variants to get a good understanding of how the predicted theoretical relationship holds with a small number of alleles and with the benefit of relatively high temporal resolution. Ultimately, I plan to use data from the deep mutational scanning experiments, which will involve sim5,000 variants, to assess the theory in an extreme population diversity case.