

THE CHANNEL CAPACITY OF SELECTIVE BREEDING: ULTIMATE LIMITS ON THE AMOUNT OF INFORMATION MAINTAINABLE IN THE GENOME

Watkins C.J.C.H.

Department of Computer Science, Royal Holloway, University of London,
Egham Hill, Egham, Surrey TW20 0EX, United Kingdom, e-mail: C.Watkins@cs.rhul.ac.uk

Key words: *evolution, information theory, channel capacity, selective breeding, genotype, sexual reproduction, asexual reproduction, genetic architecture*

Resume

Motivation: Genomes contain the information for constructing organisms. In some sense, this information is put in by selection, and it is degraded by mutation and genetic drift. It is natural to pose some basic questions of principle. What is the maximum amount of information that can be maintained at mutation-selection equilibrium? Can organisms in principle become arbitrarily complex, or, for a given mutation rate and intensity of selection, is there a limit to the amount of information that can even in principle be maintained in the genome? What method of encoding information in the genome allows the greatest amount of information to be maintained for the least selective cost? Indeed, how can the “information from selection” even in principle be defined? As far as we are aware, these questions have not previously been satisfactorily posed nor answered.

Results: We make these questions precise and answer them by modelling selective breeding as a communication channel. We show how the information-theoretic capacity of this channel provides a measure of the amount of information that can be put into or maintained in the genome by selection. The channel capacity is computed for some simple genetic models. A striking result is that for a large population in mutation-selection equilibrium, the amount of information that can be maintained using a diffuse encoding analogous to an error-correcting code is vastly greater than the amount of information that can be maintained if information is encoded as exact sequences of nucleotides.

Introduction

Organisms are shaped by selective breeding, where the selection may be natural or artificial. In some sense, selective breeding introduces and maintains the large amounts of information necessary to construct complex organisms. It is natural to ask some basic questions about the total amount of information that could be maintained in the genomes of a species through selection, whether natural or artificial.

First, how can the information that is the result of selection even in principle be defined? In real populations of members of a species, there are many common features of the genomes that are accidental and of no adaptive significance. How can we define, even in principle, the amount of information in the genomes of a species that is the result of selection?

Next, the amount of information that may be maintained will depend upon

- The mode of reproduction (sexual or asexual).
- The mutation rate per locus per generation.
- The intensity of selection (very intense selection may maintain more information).
- The method of encoding of the information in the genome.

In classical population genetics, such as in the textbook of Crow and Kimura (1970), the influence of mutations on an organism is described in terms of to what degree the mutation is advantageous or deleterious. However, there is a different and complementary view: how does selection affect the amount of information in the genome as a whole? Some mutations become fixed, others are lost: in a large genome in which many mutations occur, selection influences the fate of mutations statistically, but will not determine the fate of all mutations that occur. In slightly influencing the fates of many mutations, how does selection influence the amount of information stored in the genome as a whole?

In particular, it has been rediscovered many times that sexual reproduction eliminates deleterious mutations far more efficiently than asexual reproduction. In genetics, Crow and Kimura (1979) is an early paper, while Kondrashov (1988) provides a review: in computer science, independent analyses are Muhlenbein (1993), Baum (1995), and Mackay (1999).

We give an alternative informational analysis, and we show that the maximum maintainable amount of information in the genome is much larger with sexual than with asexual reproduction, and that selection can be more effective when applied to large numbers of loci each with small effect than when it is applied to a small number of loci, each with large effect.

Model

Selective breeding as a communication channel

Very briefly, we regard selective breeding as a communication channel in which the “message sent” is the rule for selecting which organisms in each generation to breed from, and the “message received” is a single organism sampled once mutation-selection equilibrium is reached. The information transmissible from selection rule to final organism is a measure of the maximum extent to which the genome of the organism could be influenced by selective breeding: it is a measure of the “adaptive capacity” of the organisms. The “message received” is defined to a single organism because we are interested in the information from selection that is present in the entire final breeding population. For example, the genetic information that makes a poodle a poodle must be present in almost all of a breeding population of poodles.

A simple genetic model

We use a simplified model, in which genomes are fixed-length vectors with elements that are 0 or 1. We assume that there is a large population in which all loci are in full linkage equilibrium at all times, so that for a randomly sampled genome, the values at all loci are statistically independent. (Approximate results for small populations may be obtained using the standard diffusion approximations given in Crow and Kimura (1970)).

Let the length of all genomes be L . Let the mutation rate U be defined as the fraction of loci per generation that change from 0 to 1, or from 1 to 0: mutations in each direction are assumed to be equally likely. All mutations are assumed to be point mutations that occur independently at all loci: no mutations that are insertions or deletions occur.

The exact form of the constraint on intensity is not important: we will assume there is truncation selection in which the fittest 50% of genomes are selected to breed the next generation. Subject to this constraint, it turns out that maximum channel capacity is achieved with selection rules of the form: for some ideal genome \mathbf{g} , select the 50% of genomes in the population that agree in most positions (are closer in Hamming distance) to \mathbf{g} . There are 2^L possible “ideal” genomes \mathbf{g} , and hence 2^L possible selection rules of this form. Two extreme forms of genetic encoding are easy to analyse:

1. Encoding information as an exact nucleotide sequence

Suppose that we require the genomes of all members of the population to be, with high probability, exactly equal to \mathbf{g} . For small U , the number of mutations in a child-genome will be Poisson distributed with mean UL . For truncation selection to restore the population to purity, the fraction of offspring with no mutations must be greater than $1/2$. We require therefore that $e^{-LU} \geq 1/2$, which implies that $L \leq \frac{\ln 2}{U}$. In this selection regime, each \mathbf{g} yields a distinct equilibrium population, so

for given U and optimising over L , the channel capacity is $\frac{\ln 2}{U}$ bits.

2. Encoding information as a long but highly variable nucleotide sequence

Assume that the population is large. For large L , in equilibrium the population will be polymorphic, and a randomly drawn genome will not agree with \mathbf{g} at all loci. Let a locus at which a genome agrees with \mathbf{g} be termed an *agreement*. Let the mean fraction of agreements in genomes drawn from the equilibrium population be p , where $1 > p > 1/2$. We calculate p as follows. First, note that for a large enough population, at *all* loci the fraction of alleles identical to the corresponding allele in \mathbf{g} will be close to p . The variance of the fraction of agreements in individual genomes drawn from the population is

therefore $\sigma^2 = \frac{p(1-p)}{L}$. In each generation, 50% truncation selection increases p by $\frac{\sigma}{\sqrt{2\pi}}$, and mutation reduces p

towards $1/2$ by $2U(p - 1/2)$. The equilibrium equation is therefore

$$2U(p - 1/2) = \sqrt{\frac{p(1-p)}{2\pi L}}. \text{ Solving for } p \text{ we obtain } p - 1/2 = \frac{1}{2\sqrt{8\pi LU^2 + 1}} \approx \frac{1}{4U\sqrt{2\pi L}} \text{ when } LU^2 \gg \frac{1}{8\pi}.$$

The entropy of the equilibrium population is $LH(p)$ bits, where H is the entropy function $H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$. Very briefly, in this case the channel capacity

is $\text{Capacity} = L(1 - H(p)) \approx \frac{1}{16\pi(\ln 2)U^2} \propto \frac{1}{U^2}$ for p close to $1/2$. Far more information can thus be maintained with

large L and p close to $1/2$ than with small L and p equal to 1 .

3. Asexual reproduction

Very briefly, for asexual reproduction with the same selection rules and selection intensity, the analysis is similar, except that the variance of the number of agreements is $O(LU)$. Solving for p , we obtain $p - 1/2 = O((LU)^{-1/2})$, so that the amount of information is $O(U^{-1})$ whether for exact or diffuse encodings.

Discussion

This simple abstract analysis gives two striking and surprising results. First, in sexual reproduction, the manner in which information is encoded strongly affects the amount of information that can be maintained. In principle, vastly more information may be maintained at equilibrium if information is encoded diffusely in very long genomes. For such diffuse encoding, the amount of information that can be stored is proportional to U^2 , whereas if genetic information is encoded as an exact sequence of 0s and 1s, the amount that can be stored is proportional to U^{-1} , which is very much less for small U .

Second, in asexual reproduction the maintainable information is proportional to U^{-1} , even for diffuse encodings.

These results are obtained by a rather abstract argument and must be interpreted with care: they are upper limits on the amount of information that could conceivably be maintained in equilibrium, using the most favourable possible genetic

encodings and the most efficient possible selection rules. Nevertheless, where channel capacity exists it is likely to be used. The genomes of sexual eukaryotes are in general larger than those of the asexual prokaryotes: it is tempting to speculate that the "junk" DNA of sexual eukaryotes may encode some useful information in a very diffuse way, whereas the genomes of asexual prokaryotes are compact since there can be no advantage to a diffuse encoding.

Acknowledgements

We acknowledge helpful conversations with David McAllester and Quaid Morris.

References

1. Baum E.B., Boneh D., Garrett C. (1995) On Genetic Algorithms, COLT 95: Proceedings of the Eighth Annual Conference on Computational Learning Theory, ACM, New York. 230-239.
2. Cover T.M., Thomas J.A. (1991) Elements of Information Theory, Wiley-Interscience, New York.
3. Crow J.F., Kimura M. (1970) An Introduction to Population Genetics Theory, Harper and Row, New York.
4. Crow J.F., Kimura M. (1979) Efficiency of Truncation Selection. Proc. Natl Acad. Sci. USA. 76:396-9.
5. Kondrashov A.S. (1988) Deleterious Mutations and the Evolution of Sexual Reproduction. Nature. 336 (6198): 435-440.
6. Mackay D.J.C. (1999) Rate of Acquisition of Information of a Species subjected to Natural Selection, unpublished, available from <http://www.mrao.cam.ac.uk/~mackay>
7. Muhlenbein H., Schlierkamp-Vosen D. (1993) Predictive Models for the Breeder Genetic Algorithm 1. Continuous parameter optimisation. Evolutionary Computation. 1: 25-50.