Journal of Theoretical Biology 419 (2017) 159–183

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

From typical sequences to typical genotypes

Omri Tal*, Tat Dat Tran, Jacobus Portegies

Max-Planck-Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany

ARTICLE INFO

Keywords:
Typical sequences
Typical genotypes
Population entropy rate
Population cross entropy rate
Classification

ABSTRACT

We demonstrate an application of a core notion of information theory, typical sequences and their related properties, to analysis of population genetic data. Based on the *asymptotic equipartition property* (AEP) for nonstationary discrete-time sources producing independent symbols, we introduce the concepts of *typical genotypes* and *population entropy* and *cross entropy rate*. We analyze three perspectives on typical genotypes: a set perspective on the interplay of typical sets of genotypes from two populations, a geometric perspective on their structure in high dimensional space, and a statistical learning perspective on the prospects of constructing typical-set based classifiers. In particular, we show that such classifiers have a surprising resilience to noise originating from small population samples, and highlight the potential for further links between inference and communication.

1. Introduction

We are drowning in information and starving for knowledge. - John Naisbitt.

In this paper we identify several intrinsic properties of long stretches of genetic sequences from multiple populations that justify an information theoretic approach in their analysis. Our central observation is that long genotypes consisting of polymorphisms from a source population may be considered as sequences of discrete symbols generated by a 'nonstationary source', where the capacity to sequence long stretches of genomes is congruent with the use of large block sizes in the design of communication channels. Rather than arising *temporally as an ordered sequence of* symbols in a communication channel, genetic sequences are non-temporal linear outputs of a sequencing scheme. This perspective ultimately enables the utilization of important information-theoretic asymptotic properties in the analysis of population genetic data.

Specifically, we introduce the concept of *typical genotypes* for a population, analogous to the core notion of typical sequences in information theory. These are genotypes one typically expects to encounter in a given population and are likely to represent the population very well. We analyze these typical genotypes from various perspectives. We show that it is possible that a genotype is typical to two different populations at once and give an algorithm that can quickly decide whether mutual typicality occurs, given standard population models.

Crucially, we identify conditions in which it is *likely* that mutual typicality occurs asymptotically, that is, for genotypes consisting of a very high number of variants. What we observe, however, is that in this case, only a very small portion of typical genotypes for one population is typical for another. This immediately suggests a classification scheme based on typical sets. We introduce two such typical-set based classifiers and show that their error rates decay exponentially fast, as one would expect from a good classifier. Moreover, we show that such classifiers generally perform well even in the presence of sampling noise arising from small training sets.

From a mathematical point of view, a recurring difficulty is the nonstationarity of the source distribution, or in other words, that the markers vary in their frequency across loci. This prevents us from directly utilizing some of the standard results in information theory that apply to stationary sources, and required us to find more refined mathematical arguments instead.

1.1. *Typical sequences and the* asymptotic equipartition property

Information Theory (historically, Communication Theory) is at core concerned with the transmission of messages through a noisy channel as efficiently and reliably as possible. This primarily involves two themes, data *compression* (aka, *source coding*) and error correction (aka, *channel coding*). The former theme is mainly concerned with the attainable limits to data compression for particular source distributions, while the latter involves the limits of information transfer rate across a given noisy channel. Both themes rely intrinsically on the notion of 'typical sequences'.

A key insight of Shannon, the *asymptotic equipartition property* (AEP) forms the basis of many of the proofs in information theory. The

* Corresponding author.
E-mail addresses: Omri.Tal@mis.mpg.de (O. Tal), trandat@mis.mpg.de (T.D. Tran), jacobus.portegies@mis.mpg.de (J. Portegies).

http://dx.doi.org/10.1016/j.jtbi.2017.02.010
Received 13 October 2016; Accepted 7 February 2017
Available online 12 February 2017
0022-5193/ © 2017 Elsevier Ltd. All rights reserved.

property can be roughly paraphrased as “Almost everything is almost equally probable”, and is essentially based on the law of large numbers with respect to long sequences from a random source. Stated as a limit, for any sequence of i.i.d. random variables X_i distributed according to X we have,

$$\lim_{n\to\infty} Pr\left[\left|-\frac{1}{n}\log_2 p(X_1, X_2, \ldots, X_n) - H(X)\right| < \varepsilon\right] = 1 \quad \forall\, \varepsilon > 0. \tag{1}$$

This property is expressed in terms of the information-theoretic notion of *empirical entropy*. This denotes the negative normalized log probability of a sequence x, an entity better suited for analysis than $p(x)$. This property leads naturally to the idea of typical sequences, which has its origins in Shannon's original ground-breaking 1948 paper (Shannon, 1948). This notion forms the heart of the central insights of Shannon with respect to the possibility of reliable signal communication, and features in the actual theorems and their formal proofs. The definition of a typical set $A_\varepsilon^{(n)}$ with respect a distribution source X, its entropy $H(X)$, a (small) $\varepsilon > 0$ and a (large) n, entails the set of all sequences of length n that may be generated by X such that,

$$2^{-n[H(X)+\varepsilon]} \le p(x_1, \ldots, x_n) \le 2^{-n[H(X)-\varepsilon]} \tag{2}$$

where $p(x_1, x_2, \ldots, x_n)$ denotes the probability of any particular sequence from X.

If the source is binary and *stationary* it is intuitive to spot sequences that are possibly typical. For instance, say we have a binary independent and identically distributed (i.i.d) source with a probability for “1” of 0.1, then the sequence 00001000100000000000001000000011 seems very possibly typical (as it has roughly 10% 1 s), while the sequence 01101001101100101111101001001011 is most probably not. Note that typical sequences are not the most probable ones; evidently, the most probable for this source is 00000000000000000000000000000000.

The interesting and useful properties of typical sets are a result of the AEP, and are thus *asymptotic* in nature: they obtain for large enough n, given any small arbitrary ‘threshold’ ε. Formally, for any $\varepsilon > 0$ arbitrarily small, n can be chosen sufficiently large such that:

(a) the probability of a sequence from X being drawn from $A_\varepsilon^{(n)}$ is greater than $1 - \varepsilon$, and
(b) $(1 - \varepsilon)2^{n(H(X)-\varepsilon)} \le |A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$.

Thus at high dimensionality ($n \gg 1$), the typical set has probability nearly 1, the number of elements in the typical set is nearly $2^{nH(X)}$, and consequently all elements of the typical set are nearly equiprobable with a probability tending to $2^{-nH(X)}$ (Cover and Thomas, 2006, Theorem 3.1.2).

The set of all sequences of length n is then commonly divided into two sets, the *typical set*, where the *sample entropy* or the *empirical entropy*, denoting the negative normalized log probability of a sequence, is in close proximity (ε) to the true entropy of the source per Eq. (2), and the non-typical set, which contains the other sequences (Fig. 1). We shall focus our attention on the typical set and any property that is true in high probability for typical sequences will determine the behavior of almost any long sequence sampled from the distribution.

1.2. The population model

We consider for simplicity two *haploid* populations P and Q that are in linkage equilibrium (LE) across loci, and where genotypes constitute in a sequence of *Single Nucleotide Polymorphisms* (SNPs). A SNP is the most common type of genetic variant – a single base pair mutation at a specific locus usually consisting of two alleles (the rare/minor allele frequency is >1%). Each SNP X_i is coded 0 or 1 arbitrarily, and SNPs from population P have frequencies (probability that $X_i = 1$) p_i while those from population Q have frequencies q_i. Closely following practical settings, we assume some arbitrary small cut-off frequency for SNP frequencies, such that frequencies in any population cannot be arbitrarily close to fixation, $0 < \delta < p_i, q_i < 1 - \delta$. Each genotype population sample is essentially a long sequence of biallelic SNPs, e.g., GCGCCGGGCGCCGGCGCGGGGG, which is then binary coded according to the convention above, e.g., 0101100010110010100000. The probability of such a genotype $x = (x_1, x_2, \ldots, x_n)$ from P is then $p(x) = (1 - p_1)p_2(1 - p_3)p_4p_5\ldots p_n$. We first assume the SNP frequencies are fully known (as if an infinite population sample is used in the learning stage), and later on relax this assumption in the section on small-sample related noise. Finally, for analyzing properties in expectation and deriving asymptotic statements we assume p_i and q_i are sampled i.i.d. from frequency distributions. For making explicit calculations and numerical simulations we employ a parameterized Beta distribution for SNP frequencies, such that $p_i \sim B(\alpha_P, \beta_P)$, $q_i \sim B(\alpha_Q, \beta_Q)$, as is standard in population genetic analysis (Rannala and Mountain, 1997). The use of a common Beta model for allele frequencies was adopted for both its mathematical simplicity and goodness of fit to empirical distributions from natural populations, and is by no means a prerequisite for arriving at our main results. Finally, to simulate our results, we sample SNP frequencies from these distributions and then sample long genotypes from the multivariate Bernoulli distribution for populations P and Q that are parameterized by p_i and q_i, i: $1\ldots n$, respectively.

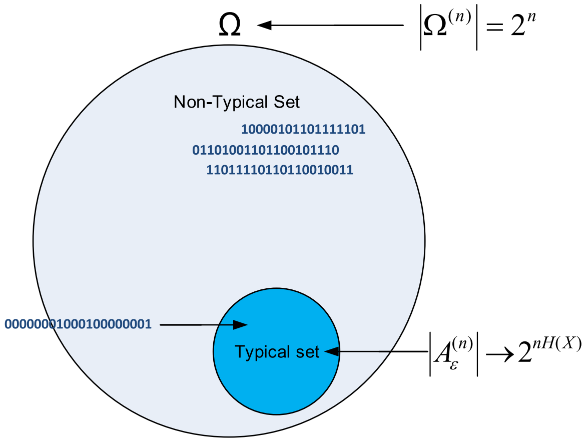

Fig. 1. The universe of all possible sequences with respect to a source distribution in a high dimensional space can be divided into two exclusive subsets, typical and non-typical. Here, we illustrate one typical sequence and a few very non-typical sequences corresponding to an i.i.d. source with probability of 0.1 for “1” for some small epsilon and high n.

1.3. Properties of sequences of genetic variants

Population SNP data have several interesting ‘set-typicality’ properties that may render them amenable to information theoretic analysis:

(a) SNPs are typically bi-valued, simplifying their representation as sequences of *binary* symbols from a communication source.
(b) The standard assumption of *linkage equilibrium* within local populations translates to a statistical independence of X_i, which in turn enables the applicability of the AEP (for sources with independent symbols).
(c) SNPs have typically differing frequencies across loci, such that in conjunction with the assumption of *linkage equilibrium* populations are analogous to ‘nonstationary’ sources with independent symbols, in turn allowing the use of advanced forms of the AEP, as we shall see.
(d) The recent availability of very large number of SNPs from high-throughput sequencing of genomes enables the consideration of