Contents lists available at ScienceDirect

BioSystems

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

Use of directed quasi-metric distances for quantifying the information of gene families

Steinar Thorvaldsen[a,*], Ola Hössjer[b]

[a] Dept. of Education, Division of Science, UiT the Arctic University of Norway, Norway
[b] Dept. of Mathematics, Stockholm University, Sweden

ARTICLE INFO

Keywords:
Self-information
Quasi-metrics
Mutual information
Functional information
Bioinformatics
Multi-category data

ABSTRACT

A large hindrance to analyzing information in genetic or protein sequence data has been a lack of a mathematical framework for doing so. In this paper, we present a multinomial probability space $\mathscr{X}$ as a general foundation for multicategory discrete data, where categories refer to variants/alleles of biosequences. The external information that is infused in order to generate a sample of such data is quantified as a distance on $\mathscr{X}$ between the prior distribution of data and the empirical distribution of the sample. A number of distances on $\mathscr{X}$ are treated. All of them have an information theoretic interpretation, reflecting the information that the sampling mechanism provides about which variants that have a selective advantage and therefore appear more frequently compared to prior expectations. This includes distances on $\mathscr{X}$ based on mutual information, conditional mutual information, active information, and functional information. The functional information distance is singled out as particularly useful. It is simple and has intuitive interpretations in terms of 1) a rejection sampling mechanism, where functional entities are retained, whereas non-functional categories are censored, and 2) evolutionary waiting times. The functional information is also a *quasi-metric* on $\mathscr{X}$, with information being measured in an asymmetric, mountainous landscape. This quasi-metric property is also retained for a robustified version of the functional information distance that allows for mutations in the sampling mechanism. The functional information quasi-metric has been applied with success on bioinformatics data sets, for proteins and sequence alignment of protein families.

1. Introduction

Biology has changed profoundly in the last two decades, transitioning from a descriptive science into a dynamic landscape of design and innovation driven by technological improvements. All life is based on genetic sequences stored in DNA, and today a large amount of sequence data is available in nucleotide and amino acid databases. An important aspect of analyzing such nucleotide or protein sequence data is to find a mathematical framework that makes it possible to extract information from these large datasets. Indeed, information has become a central concept of modern biology, and there is a common understanding that the informational aspect of life is fundamental (Godfrey-Smith and Sterelny, 2016; Walker and Davies, 2013; Griffiths, 2017). Claude Shannon's mathematical theory was a first important step towards developing a quantitative understanding of biology (Shannon, 1948), and his theory has been successfully applied to quantify and analyze nucleotide and amino acid sequences (Schneider, 2006; Schneider and Stephens, 1990). However, it is well known that Shannon's theory is not sufficient to describe the logic of biotic information in general (Thorvaldsen et al., 2024), since it does not take into account that biotic information is relative to its context in terms of semantics, function and codes within the living cell. This is to say that biotic information must be treated as relative to its context (Logan, 2012).

In this paper we will apply the notion of relative information in the context of a gene or protein family. In more detail, we develop a probabilistic framework that makes it possible to quantify the amount of external information that is required to generate a family of amino acids of specific form. The same challenge is faced for other bioinformatics datasets, whose components are different from amino acids. More specifically, we will quantify the amount of external information required to generate a multicategory dataset, where categories typically correspond to variants (or alleles) of biosequences. This external information is defined as a distance between prior expectations of the biosequence distribution, and the observed empirical distribution of the

[*] Corresponding author.
E-mail addresses: steinar.thorvaldsen@uit.no (S. Thorvaldsen), ola@math.su.se (O. Hössjer).

https://doi.org/10.1016/j.biosystems.2024.105256
Received 11 March 2024; Received in revised form 6 June 2024; Accepted 11 June 2024
Available online 12 June 2024

biosequences. In order to interpret such a distance, imagine a large pool or reservoir of biosequences distributed according to the prior, and that a small subset of sequences is sampled from this pool. The distance between the prior distribution of the reservoir and the distribution of the sample then corresponds to the amount of information that the sampling procedure that generated data mediates about the categories. The distance gets larger the more selection the sampling mechanism includes, so that categories with a higher fitness will have higher empirical frequencies compared to prior expectations. We consider distances based on conditional mutual information, mutual information, active information and functional information (FI). We argue that the latter functional information distance is particularly appealing, since it has a simple form, an intuitive interpretation in terms of FI (Szostak, 2003; Hazen et al., 2007), it naturally corresponds to a rejection sampling mechanism (Wells et al., 2004) that also incorporates mutations, and under certain assumptions it relates to an evolutionary waiting time (Durrett and Schmidt, 2008; Durrett et al., 2009; Behrens and Vingron, 2010; Sanford et al., 2015; Hössjer et al., 2021). The functional information distance is also a quasi-metric (Wilson, 1931), which means that it satisfies all properties of a regular metric except for symmetry. A quasi-metric corresponds to path lengths of walks in an asymmetric mountainous landscape, where the uphill distance between two points is larger than the downhill distance between these points. For the FI distance in particular, an uphill walk typically corresponds to increasing the frequencies of one or several categories, with low prior probabilities, by large multiplicative factors, whereas for a downhill walk, none of the categories have such a large multiplicative increase of its frequency. Within a biological context this has a natural interpretation, that it is more difficult to build up new structures within the living cell, than to remove existing ones.

Our paper is organized as follows: In Section 2 we introduce multinomial probability spaces with quasi-metrics for multicategory data. Various measures of information are defined in Section 3, whereas the sampling procedures that correspond to the abovementioned walks in a mountainous landscape are described in Section 4. This is used in Section 5 to define several distance measures between the prior and empirical distributions of data. Some protein datasets are analyzed with the FI quasi-metric in Section 6, and conclusions are presented in Section 7. Mathematical details are given in AppendicesA-E, and a summary of the most important concepts and mathematical notation are provided in Tables 1 and 2 respectively.

Table 1
Definitions and terminology.

Term	Description
Allele, category or type	A version of a particular biosequence, monomer or gene.
Base population or reservoir	A large collection of copies of a biosequence, monomer or gene.
Allele frequency	The relative abundance, in a population, of a particular allele
Allele frequency vector	A vector with frequencies of all alleles.
Sampling procedure	A description of how a subsample of biosequences is collected from the base population.
Rejection sampling	A sampling procedure where randomly chosen alleles are retained (not censored) in proportion to their fitness.
Selection	The alleles of the base population have different fitness, in proportion to how easily they are sampled. This definition of fitness does not involve the concept of generation or reproduction.
Information	The sampling procedure carries information about the frequencies of the sampled alleles. This information is positive if and only of the sampling procedure is not selectively neutral.
Functional information	For rejection sampling this is minus the base 2 logarithm (or the self information) of the fraction of retained (functional) biosequence copies.
Distance	A distance is defined between the allele frequency vector of the base population and the sample. The distance corresponds to how much information the sampling procedure carries about varying fitness of alleles.
Quasi-metric	A non-symmetric distance measure that satisfies the identity, point equality and triangle inequality.

Table 2
Mathematical notation.

Quantity	Description
K	Number of possible categories/variants/alleles.
X	A randomly chosen allele ($\in \{1,\ldots,K\}$).
k	Number of a particular allele ($\in \{1,\ldots,K\}$) and observed value of X.
$\mathscr{X}$	Space of multinomial probability vectors.
$\boldsymbol{R}$	Large reservoir of alleles.
$\boldsymbol{p} = (p_1,\ldots,p_K)$	Prior distribution of alleles ($\in \mathscr{X}$) drawn from $\boldsymbol{R}$.
L	Size of base population.
$X_{\text{samp}} = (X_1,\ldots,X_L)$	Base population (sample of size L drawn from $\boldsymbol{R}$).
Υ	Sampling space.
Y	Random sample ($\in \Upsilon$) for rejection sampling. Either the non-censored subsample $Y = (Y_1, \ldots, Y_M)$ of X_{samp} of size $1 \leq M \leq L$, or a binary censoring indicator $Y \in \{0,1\}$.
y	Observed value of Y.
$\boldsymbol{q}_y = \left(q_{1y},\ldots,q_{Ky}\right)$	Conditional distribution of alleles, given $Y = y$ ($\boldsymbol{q}_y \in \mathscr{X}$ and $q_{ky} = P(X = k\vert Y = y)$ for $k = 1,\ldots,K$).
$\boldsymbol{q} = (q_1,\ldots,q_K)$	Approximation of $\boldsymbol{q}_y$ for large populations.
$\boldsymbol{r}_y = (r_{1y},\ldots,r_{Ky})$	Vector with probabilities of retaining alleles $1,\ldots,K$ for rejection sampling, given observed subsample $Y = y$ of non-censored observations.
$\boldsymbol{r} = (r_1,\ldots,r_K)$	Approximation of $\boldsymbol{r}_y$ for large populations.
$\boldsymbol{N} = (N_1,\ldots,N_K)$	Composition vector, either for the base population ($\boldsymbol{N} = \boldsymbol{N}(X_{\text{samp}})$) or the non-censored subsample ($\boldsymbol{N} = \boldsymbol{N}(Y)$).
$d\left(\boldsymbol{p},\boldsymbol{q}_y\right)$	Distance between $\boldsymbol{p}$ and $\boldsymbol{q}_y$.
$d_{FI}\left(\boldsymbol{p},\boldsymbol{q}_y\right)$	Functional information distance between $\boldsymbol{p}$ and $\boldsymbol{q}_y$ ($= -\log_2(\boldsymbol{p} \bullet \boldsymbol{r}_y)$).
$I(k)$	Self-information of outcome k of X.
$H(X)$	Entropy of X.
$CMI(X;y) = d_{CMI}\left(\boldsymbol{p},\boldsymbol{q}_y\right)$	Conditional mutual information between X and observed value y of Y.
$MI(X,Y) = d_{MI}(\boldsymbol{p},\boldsymbol{q})$	Mutual information between X and Y ($= E[CMI(X;Y)]$).
$I^+(X;y) = d_{KL}\left(\boldsymbol{q}_y \Vert \boldsymbol{p}\right)$	Expected active information between X and observed value y of Y, or Kullback-Leibler distance between $\boldsymbol{p}$ and $\boldsymbol{q}_y$.
ε	Mutation probability.
Π	Mutation probability matrix.

2. Multinomial probability spaces and quasi-metrics

2.1. Multinomial probability spaces

In probability theory a K-dimensional *categorical distribution* denotes a discrete probability distribution that describes the possible results of a categorical random variable X that can take on one of K possible categories, with the probability of each category separately specified. There is no innate underlying ordering of these outcomes, but numerical labels $1,\ldots,K$ are attached in describing the distribution, were $\boldsymbol{p} = (p_1,\ldots,p_K)$ are the probabilities that satisfy $p_k \geq 0$ and

$$\sum_{k=1}^{K} p_k = 1.$$

We designate $\mathscr{X}$ the sample space of all such $\boldsymbol{p}$, often referred to as the standard $(K-1)$-dimensional simplex.

The categorical distribution of X will be denoted $\text{Cat}(\boldsymbol{p})$, and it is a general distribution over a K-way event. Suppose $X_{\text{samp}} = (X_1,\ldots,X_L)$ is a sample of independent and identically distributed random variables with marginal distribution $X_l \sim \text{Cat}(\boldsymbol{p})$. The total number of occurrences of the K outcomes are represented by the components of the composition vector $\boldsymbol{N}(X_{\text{samp}}) = (N_1,\ldots,N_K)$, with $N_k = \sum_{l=1}^{L} 1(X_l = k)$. This vector has a multinomial $\text{Mult}(L,\boldsymbol{p})$ distribution. Since $\boldsymbol{p} \in \mathscr{X}$ can be seen as a multinomial probability vector**,** we also refer to $\mathscr{X}$ as a multinomial probability space.

2.2. Genetics interpretation

In genetics, the categories of Section 2.1 represent K different variants (or alleles) of a portion of DNA (such as a genetic marker or a gene), or the possible amino acids at a particular site of a family of aligned protein sequences. In this context $\boldsymbol{p}$ is referred to as the vector of allele frequencies. An important topic of population genetics is to study how $\boldsymbol{p}$ changes forwards in time, prospectively, due to forces of selection, mutations, genetic drift and migration. This includes, for instance, the K-allele Wright-Fisher model and the K-allele Moran model (cf. Section 4.9 of Ewens, 2004, Section 8.1 of Durrett, 2008, and references therein). In spite of the success of population genetics, selection and fitness are theoretical constructs that are often difficult to interpret. Fitness is often defined, in a given context, in terms of reproductive schedules rather than biological functionality (Lewontin, 2003; Basener et al., 2021).

In this article we will not study the allele frequency process prospectively but instead have a retrospective approach. This makes it possible to attain a somewhat more empirical and long-term definition of fitness, which more easily opens up for interpretations in terms of biological function. In more detail, we assume that the allele frequency vector of a population, or large reservoir, has already been changed from $\boldsymbol{p}$ to $\boldsymbol{q}$ by some process that is pictured as sampling from the reservoir. This sampling procedure may implicitly involve the population genetic trajectory and the fitness landscape (Wright, 1932) along the forwards in time path from $\boldsymbol{p}$ to $\boldsymbol{q}$ (see Sections 4.3 and 7.3). Given that $\boldsymbol{q}$ has already been observed, we will look back in time and ask ourselves the following question: What information was infused into the sampling procedure in order to change the allele frequencies from $\boldsymbol{p}$ to $\boldsymbol{q}$ and is it possible to give this information a fitness interpretation? To answer this question, we will first define a distance measure between $\boldsymbol{p}$ to $\boldsymbol{q}$ and later on give this distance measure information theoretic and fitness interpretations.

2.3. Distances, quasi-metrics, and metrics

We will introduce a measure of *distance* $d(\boldsymbol{p},\boldsymbol{q})$ between pairs of elements $\boldsymbol{p}$ and $\boldsymbol{q}$ of the multinomial probability space $\mathscr{X}$**.** A metric (distance) space suggests that given two points of the space there should be a real number that measures the distance d between them. This is not straightforward, and several options based on the underlying axioms exist. The definition of a quasi-metric (or directed metric), (Wilson 1931; Stojmirovic 2005; Cobzas 2013; Khamsi 2015) and a metric are as follows:

Let $\mathscr{X}$ be a set, which in this paper will be taken as the multinomial probability space of dimension $K-1$. A *quasi-metric* d on $\mathscr{X}$ is a map $d:\mathscr{X}\times\mathscr{X}\rightarrow[0,\infty)$ such that for all $\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}\in\mathscr{X}$ the following conditions (i)-(iii) are satisfied:

(i) $d(\boldsymbol{p},\boldsymbol{p}) = 0$; (Identity)
(ii) $d(\boldsymbol{p},\boldsymbol{q}) = d(\boldsymbol{q},\boldsymbol{p}) = 0 \Rightarrow \boldsymbol{p} = \boldsymbol{q}$; (Point equality)
(iii) $d(\boldsymbol{p},\boldsymbol{s}) \leq d(\boldsymbol{p},\boldsymbol{q}) + d(\boldsymbol{q},\boldsymbol{s})$; (Triangle inequality)

If d satisfies (i)-(iii) and additionally

(iv) $d(\boldsymbol{p},\boldsymbol{q}) = d(\boldsymbol{q},\boldsymbol{p})$; (Symmetry)

we say that d is a *metric* on $\mathscr{X}$, and $\mathscr{X}$ is a “standard” metric space. Well known examples of metrics on $\mathscr{X}$ include the total variation distance

$$d_1(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\sum_{k=1}^{K}|q_k - p_k|,$$

the Hellinger distance

$$d_2(\boldsymbol{p},\boldsymbol{q}) = \left(\frac{1}{2}\sum\nolimits_{k=1}^{K}\left(\sqrt{q_k} - \sqrt{p_k}\right)^2\right)^{1/2},$$

and the maximum distance

$$d_\infty(\boldsymbol{p},\boldsymbol{q}) = \max\nolimits_{1\leq k\leq K}|q_k - p_k|.$$

Although these three distances have a simple form, they do not take into account the way in which $\boldsymbol{q}$ is obtained form $\boldsymbol{p}$ through a sampling procedure. They are also symmetric, and in our setting when the first argument of d corresponds to a prior distribution and the second argument is the empirical distribution of data, this turns out to be a disadvantage. Avoiding the condition of symmetry allows us to distinguish between the distance from $\boldsymbol{p}$ to $\boldsymbol{q}$ and that from $\boldsymbol{q}$ to $\boldsymbol{p}$. This is intuitively useful if $d(\boldsymbol{p},\boldsymbol{q})$ is to measure the amount of effort to go from point $\boldsymbol{p}$ to point $\boldsymbol{q}$ in a landscape of mountains and valleys, where it usually takes more effort to go up than down. This involves asymmetric notions of "cost" which arise naturally from the observation that it is harder to walk uphill than downhill. In the same way it takes more effort to build up information, than to dissolve it. This intuitive picture of quasi-metrics $d(\boldsymbol{p},\boldsymbol{q})$ is confirmed by general results (see for instance Stojmirovic, 2005) stating that a quasi-metrics can be represented as path distances on weighted directed graphs. In this context, for a space $\mathscr{X}$ with a quasi-metric d, the geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$ is the path from $\boldsymbol{p}$ to $\boldsymbol{q}$ that has the shortest cumulative distance.

For these reasons, in this article we will search for distances $d(\boldsymbol{p},\boldsymbol{q})$ that a) are quasi-metrics, and b) have a natural interpretation in terms obtaining $\boldsymbol{q}$ from $\boldsymbol{p}$ through a sampling procedure. Before defining such distances in Section 5, we will first introduce measures of information and sampling procedures in Sections 3 and 4 respectively.

3. Basic measures of information

Information theory is a vast subject, and we refer to Burgin (2010) and Cover and Thomas (2006) for extensive overviews. In this section we will define different ways of quantifying how much information the sample $Y\in\Upsilon$ mediates about a category $X\sim\mathrm{Cat}(\boldsymbol{p})$ from the base population X_{samp}. Loosely speaking the stronger the selection component of the sampling mechanism is, the more information Y carries about X, in the sense that the conditional distribution of X given Y departs more from the prior distribution of X. It is assumed that the sampling mechanism belongs to some space Υ, the form of which will be specified in Section 4.

3.1. Self-information and entropy

We will start by introducing some information measures for X, before sampling takes place. What is commonly referred to as *self-information* in information theory can be applied to the probabilities of the vector $\boldsymbol{p}$. This is a measure of information content or ‘surprisal’ of a particular outcome k of X*:*

$$I\,(k) \overset{\mathrm{def}}{=} \log_2\frac{1}{p_k} = -\log_2(p_k).$$

This definition states that the surprisal of an outcome corresponds to the amount of self-information carried by that outcome. In information theory, the expected *self-information*

$$H(X) = H^{(\boldsymbol{p})} = -\sum_{k=1}^{K} p_k\log_2(p_k)$$

of a random variable $X\sim\mathrm{Cat}(\boldsymbol{p})$ is used to express the mutual information of X with itself, and it equals the entropy of the random variable (Shannon, 1948). This is the reason that entropy is sometimes referred to as expected self-information.

3.2. Conditional mutual information, mutual information, and Rokhlin measures

We will refer to

$$\begin{aligned} CMI(X;y) &= \Delta H(X;y) = H(X) - H(X|y) \\ &= -\sum_{k=1}^{K} p_k \log_2(p_k) + \sum_{k=1}^{K} q_{ky} \log_2\left(q_{ky}\right) \end{aligned} \tag{1}$$

as the conditional mutual information between $X \sim \text{Cat}(\boldsymbol{p})$ and an observed value y of the sampling mechanism $Y \in \Upsilon$, see for instance Thorvaldsen and Hössjer (2023) and references therein. This is the reduction in uncertainty about X that the observation $Y = y$ conveys, and it involves the prior distribution $p_k = P(X = k)$ as well as the conditional distribution $q_{ky} = P(X = k|Y = y)$ of X given $Y = y$. The index y of the sampling distribution $\boldsymbol{q}_y = \left(q_{1y}, \ldots, q_{Ky}\right)$ indicates that this distribution depends of the actual outcome of the sample Y. However, when the size of the sample Y gets large, it assumed that $\boldsymbol{q}_y$ approaches some limit $\boldsymbol{q} = (q_1, \ldots, q_K)$ that only depends on the sampling mechanism and not the outcome of the sample.

Mutual information $\text{MI}(X, Y) = E[\Delta H(X; Y)]$ was introduced by Shannon (1948). In contrast to conditional mutual information, mutual information refers to the *expected* reduction in the uncertainty of X that Y mediates. It captures all dependencies between X and Y, and it measures how much the Shannon uncertainty for one of these two random variables is expected to decrease when knowledge of another random variable is taken into account. High mutual information indicates a large reduction in uncertainty. In order to compute the mutual information between X and Y we need to know p_k, $P(Y = y)$ and q_{ky} for all possible outcomes k and y of X and Y. Given these quantities, the mutual information can be found through

$$\begin{aligned} \text{MI}(X, Y) &= H(X) - E[H(X|Y)] \\ &= -\sum_{k=1}^{K} p_k \log_2(p_k) + \sum_{y\in\Upsilon} P(Y = y) \sum_{k=1}^{K} q_{ky} \log_2\left(q_{ky}\right), \end{aligned} \tag{2}$$

or equivalently

$$\text{MI}(X, Y) = H(X) + H(Y) - H(X, Y), \tag{3}$$

where $H(Y) = -\sum_{y\in\Upsilon} P(Y = y) \log_2[P(Y = y)]$ is the entropy of Y and $H(X, Y)$ is the joint entropy of X and Y. It follows from equations (2) and (3) that $\text{MI}(X, Y)$ is also the expected value (with respect to the joint variation of X and Y) of the so called pointwise mutual information

$$\log_2\left[\frac{P(X = k, Y = y)}{P(X = k)P(Y = y)}\right] = \log_2\left[\frac{q_{ky}}{p_k}\right],$$

introduced by Fano (1961).

Another notion of information, somewhat related to MI, is the Rokhlin measure (Rokhlin 1967; Srivastava and Khare 1999).

$$\text{R}(X, Y) = H(X, Y) - \ 0.5[H(X) + H(Y)]. \tag{4}$$

3.3. Active information

Active information, I^+, was introduced by Dembski and Marks to handle infusion of knowledge in random search algorithms (Dembski and Marks 2009a, 2009b). It was later applied to population genetics by Díaz-Pachón and Marks (2020), whereas Díaz-Pachón and Hössjer (2022) used active information in order to model fine-tuning. In our context, $I^+(k) = \log_2\left(q_{ky}/p_k\right)$ is the active information associated with a change of the frequency of outcome k from the prior probability p_k to the observed relative frequency q_{ky}. Note in particular that $I^+(k)$ equals the pointwise mutual information.

Taking an average with respect to all possible outcomes of X, a change in the frequency distribution from $\boldsymbol{p}$ to $\boldsymbol{q}_y$, corresponds to the expected active information

$$\begin{aligned} I^+(X;y) &= E[I^+(X)|Y = y] = E_{\boldsymbol{q}_y}^{(\boldsymbol{p})} - E_{\boldsymbol{q}_y}^{\left(\boldsymbol{q}_y\right)} \\ &= -\boldsymbol{q}_y \bullet \log_2 \boldsymbol{p} + \boldsymbol{q}_y \bullet \log_2 \boldsymbol{q}_y = \sum_{k=1}^{K} q_{ky} \log_2 \frac{q_{ky}}{p_k}, \end{aligned} \tag{5}$$

where $E_{\boldsymbol{q}_y}^{(\boldsymbol{p})}$ is the cross-entropy of $\boldsymbol{p}$ relative $\boldsymbol{q}_y$, $\log_2 \boldsymbol{p} = (\log_2 p_1, \ldots, \log_2 p_K)$, $\bullet$ refers to the scalar product between two vectors of equal length K and $E_{\boldsymbol{p}}^{(\boldsymbol{p})} = H^{(\boldsymbol{p})}$. Motivated by continuity, we define $0\bullet\log 0 = 0$ for all terms in (5) such that $q_{ky} = 0$.

4. Rejection sampling and some new measures of information

4.1. Rejection sampling mechanisms

We will consider a subsample $Y = (Y_1, \ldots, Y_M)$ of $X_{\text{samp}} = (X_1, \ldots, X_L)$ of size $M = M(Y) \leq L$ that belongs to a sample space

$$\Upsilon = \bigcup_{M=1}^{L} \{1, \ldots, K\}^M.$$

The total number of occurrences of the K outcomes are represented by the components of the composition vector $\boldsymbol{N}(Y) = (N_1, \ldots, N_K)$, with $N_k = \sum_{m=1}^{M} 1(Y_m = k)$. Let y refer to the observed value of the sampling scheme Y, whereas $\boldsymbol{q}_y \in \mathscr{X}$ is the corresponding observed value of $\boldsymbol{N}/M$. The elements of the vector $\boldsymbol{q}_y = \left(q_{1y}, \ldots, q_{Ky}\right)$ are the relative frequency counts for the K different categories, with an interpretation $q_{ky} = P(X = k|Y = y)$, where X is a randomly chosen member of the subsample Y. As mentioned in Section 3, when the size M of the sample gets large, it is assumed that $\boldsymbol{q}_y$ approaches a limit $\boldsymbol{q} = (q_1, \ldots, q_K)$.

One way of obtaining the subsample $Y = (Y_1, \ldots, Y_M)$ from $X_{\text{samp}} = (X_1, \ldots, X_L)$ is through rejection sampling (Wells et al., 2004), with proposal distribution $\boldsymbol{p}$ and a target distribution $\boldsymbol{q}$. It is hypothetically assumed that X_{samp} is obtained through sampling L times from a very large reservoir $\boldsymbol{R}$, whose elements have one of K categories, with a distribution $\boldsymbol{p}$. The subsample $Y = Y_{\text{nc}}$ consists of $M \leq L$ non-censored observations. It is obtained through a rejection or censoring mechanism, with

$$r_k = \frac{q_k/p_k}{\max(q_1/p_1, \ldots, q_K/p_K)} \tag{6}$$

the probability of retaining (not censoring) each sampled copy X_l of $\boldsymbol{R}$ that equals k. The larger r_k is the larger is the selective advantage of category k in the sampling process by which Y is drawn. Equation (6) implies that each non-censored copy is distributed as

$$P(X_l = k|X_l \text{ non} - \text{censored}) = \frac{p_k r_k}{\sum_{l=1}^{K} p_l r_l} = q_k,$$

when the randomness of the base population X_{samp} is accounted for. From this it follows that

$$\boldsymbol{q}_y \in \text{Mult}(M, \boldsymbol{q})\Big/M \tag{7}$$

is the empirical distribution of categories from the observed sample y with marginal frequencies $q_{ky} = P(X = k|Y = y)$, where X is a randomly chosen copy from the non-censored subsample. Analogously,

$$r_{ky} = \frac{q_{ky}\Big/p_k}{\max\Big(q_{1y}\Big/p_1,\dots,q_{Ky}\Big/p_K\Big)} \tag{8}$$

can be seen as the fraction of sampled copies from $\boldsymbol{R}$ with outcome k that are retained (not censored). Note in particular that $M = L$ and $\boldsymbol{q}_y \in \text{Mult}(L,\boldsymbol{p})/L \approx \boldsymbol{p}$ if none-of the sampled copies are rejected, so that $r_k = 1$ and $r_{ky} \approx 1$ when L is large, for $k = 1,\dots,K$.

Regarding the fraction $\boldsymbol{p} \bullet \boldsymbol{r_y}$ of retained copies as functional, this gives rise to the functional information

$$FI = -\log_2(\boldsymbol{p} \bullet \boldsymbol{r_y}), \tag{9}$$

see Szostak (2003), Hazen et al. (2007), and Thorvaldsen and Hössjer (2023) for more details. The rejection sampling mechanism (6)–(7) will be extended in Section 5.3.2 to not only include selection through censoring, but also mutations.

4.2. Sampling in terms of censoring indicators

Consider the rejection sampling mechanism of Section 4.1, and assume that a very large number L of copies $X_{\text{samp}} = (X_1,\dots,X_L)$ are drawn from the reservoir $\boldsymbol{R}$, with subsamples Y_{nc} and Y_{c} of non-censored and censored copies of lengths $M_{\text{nc}} \approx Lr$ and $M_{\text{c}} = L - M_{\text{nc}} \approx L(1-r)$ respectively, where $r = \sum_{k=1}^{K} r_k p_k$ is the probability that a randomly chosen element drawn from $\boldsymbol{R}$ is retained. Suppose we want the sampling mechanism to account for whether a copy drawn from the reservoir $\boldsymbol{R}$ was censored or not. This can be modelled as a binary censoring variable $Y \in \boldsymbol{\Upsilon} = \{0,1\}$, where $Y = 0$ and $Y = 1$ correspond to a censored or non-censored copy X of the base population X_{samp} respectively. From this it follows that $P(Y=0) = P(X \in Y_{\text{c}}) = 1 - r$ and $P(Y = 1) = P(X \in Y_{\text{nc}}) = r$. Since L, and hence also $M = M_{\text{nc}}$, is large, if follows from (7) that the composition vectors of the non-censored and censored samples satisfy

$$\boldsymbol{N}(Y_{\text{nc}}) \approx L(r_1p_1,\dots,r_Kp_K) = Lr(q_1,\dots,q_K)$$

and

$$\boldsymbol{N}(Y_{\text{c}}) \approx L((1-r_1)p_1,\dots,(1-r_K)p_K) = L(1-r)(q_{10},\dots,q_{K0}).$$

In particular, a randomly chosen member X of the non-censored and censored subsamples are distributed as $P(X = k|Y_{\text{nc}}) = q_{k1} = q_k$ and $P(X = k|Y_{\text{c}}) = q_{k0} = p_k(1-r_k)/(1-r)$ respectively. The more the vectors $\boldsymbol{q}_1 = \Big(q_{11},\dots,q_{K1}\Big) = \boldsymbol{q}$ and $\boldsymbol{q}_0 = \Big(q_{10},\dots,q_{K0}\Big)$ differ, the more selection the sampling mechanism includes.

4.3. Evolutionary trajectory sampling

Consider a population that evolves of time, with $L(t)$ the number of individuals at time $t \geq 0$, with $L = L(0)$. Let $X_l(t) \in \{1,\dots,K\}$ refer to the category of individual l at time t. The population $X_{\text{samp}} = \{X_1(0),\dots,X_L(0)\}$ at time 0 is drawn from a large reservoir $\boldsymbol{R}$ of individuals with distribution $\boldsymbol{p}$, whereas the future dynamics of the population is determined by selection. The population is sampled at some (possibly random) time point T. The size of the sampled population is $M = L(T)$, the sample is $Y = (Y_1,\dots,Y_M) = (X_1(T),\dots,X_M(T))$, and the sample space Υ is the same as in Section 4.1.

Let $\boldsymbol{N}_t = (N_{t1},\dots,N_{tK})$, with $N_{tk} = \sum_{l=1}^{L(t)} 1(X_l(t) = k)$, refer to the composition vector at time t. Then $\boldsymbol{p}_t = \boldsymbol{N}_t/L(t) \in \mathscr{X}$ is the genetic composition of the population at time t, whereas $\boldsymbol{q}_y = \boldsymbol{p}_T$ is the genetic composition of the sample, taken at time T.

We will assume that $\boldsymbol{N}_t$ is a multitype death process. This is a Markov model in continuous time, where individuals of type k die at rate $\lambda_k \geq 0$, independently of the other individuals. Consequently, N_{tk} decreases by 1 at rate $\lambda_k N_{tk}$. Since λ_k depends on k, this makes it possible to include selection, since the smaller λ_k is, the higher is the fitness of type k individuals.

The Markov process algorithm can be viewed as a special case of the rejection sampling algorithm of Section 4.1. Indeed, the probability of retaining an individual of type k at time T is

$$r_k = \int_0^\infty e^{-\lambda_k t} f(t)dt, \tag{10}$$

where f is the density function of T. In order for (4) to conform with the requirement $\max(r_1,\dots,r_K) = 1$ imposed by (6), we will assume that $\lambda_{\min} = \min(\lambda_1,\dots,\lambda_K) = 0$.

5. Main results: information based distances

In this section we regard $\boldsymbol{p} = (p_1,\dots,p_K) \in \mathscr{X}$ as a multinomial probability vector that corresponds to a prior assumption on the distribution of data $X \in \text{Cat}(\boldsymbol{p})$ with K categories, whereas $\boldsymbol{q}_y = \Big(q_{1y},\dots,q_{Ky}\Big) \in \mathscr{X}$ summarizes the conditional distribution $q_{ky} = P(X = k|y)$ of X given data y. We will make use of the measures of information of Section 3 and the sampling mechanisms of Section 4 in order to introduce a number of distance measures $d\Big(\boldsymbol{p},\boldsymbol{q}_y\Big)$ that quantify how much data change the prior assumptions on the distribution of X. In particular, we will investigate which of these distance measures that qualify as quasi-metrics, as defined in Section 2.3.

5.1. The conditional mutual information approach

A conditional version of the commonly used *mutual information* from information theory may be applied to multinomial probability space $\mathscr{X}$. It follows from (1) that the conditional mutual information corresponds to a distance

$$d_{CMI}\Big(\boldsymbol{p},\boldsymbol{q}_y\Big) = \Delta H(X;y) = H(X) - H(X|Y=y) = -\boldsymbol{p} \bullet \log_2 \boldsymbol{p} + \boldsymbol{q}_y \bullet \log_2 \boldsymbol{q}_y$$

between $\boldsymbol{p}$ and $\boldsymbol{q}_y$. Despite of its useful information theoretic interpretation, this distance is skew-symmetric and since it takes on negative values it does not qualify as a quasi-metric.

5.2. The active information approach

The expected active information was introduced in (5). It is equivalent to the Kullback–Leibler divergence

$$d_{KL}\Big(\boldsymbol{q}_y\Big\|\boldsymbol{p}\Big) = E[I^+(X)|Y=y] = -\boldsymbol{q}_y \bullet \log_2 \boldsymbol{p} + \boldsymbol{q}_y \bullet \log_2 \boldsymbol{q}_y$$

between $\boldsymbol{p}$ to $\boldsymbol{q}_y$ (Kullback and Leibler 1951). Although some terms of the expression for the expected active information are negative when $\boldsymbol{p} \neq \boldsymbol{q}_y$, the sum is always non-negative and quantifies a directed 'distance', or relative information, between two probability distributions over the same sample space, with $d_{KL} = 0$ being the most similar (the probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}_y$ are identical). This can be phrased as $d_{KL}\Big(\boldsymbol{q}_y\Big\|\boldsymbol{p}\Big) \geq 0$, with $d_{KL}\Big(\boldsymbol{q}_y\Big\|\boldsymbol{p}\Big) = 0$ if and only if $\boldsymbol{p} = \boldsymbol{q}_y$, so that d_{KL} satisfies the range, identity and point equality properties of Section 2.3. The distance d_{KL} is also asymmetric, and it has many useful properties, with applications in statistics and data mining. However, the Kullback–Leibler divergence does not satisfy the triangle inequality (iii) of Section 2.3 (Cover and Thomas, 2006). Thus d_{KL} does not qualify as a quasi-metric of a probability space.

5.3. The Functional information approach

5.3.1. Functional information without mutations

The third model is *Functional information*. Recalling the censoring mechanism of Section 4.1, it gives rise to the measure

$$d_{FI}\left(\boldsymbol{p},\boldsymbol{q}_y\right) = -\log_2\left(\boldsymbol{p}\bullet\boldsymbol{r}_y\right) = \log_2\max\left(\frac{q_{1y}}{p_1},\ldots,\frac{q_{Ky}}{p_K}\right) - \log_2\sum_{k=1}^{K}q_{ky} = \max\log_2\left(\frac{q_{1y}}{p_1},\ldots,\frac{q_{Ky}}{p_K}\right) \qquad (11)$$

of information (Thorvaldsen and Hössjer 2023), where $\boldsymbol{r}_y = \left(\boldsymbol{r}_{1y},\ldots,\boldsymbol{r}_{Ky}\right)$ is a vector containing the non-censoring fractions (8) for all outcomes $k = 1,\ldots,K$, given an observed value y of the censoring mechanism. Note that $d_{FI}\left(\boldsymbol{p},\boldsymbol{q}_y\right)$ differs from the Kullback-Leibler divergence $d_{KL}\left(\boldsymbol{q}_y\middle\|\boldsymbol{p}\right)$ in that a weighted summation of all $\log_2\left(q_{yk}/p_k\right)$ for $k = 1,\ldots,K$, with weights q_{ky}, is replaced by a maximum operation.

An expression analogous to (11) also appears in the definition of functional information, as introduced by Jack Szostak, in an important bioinformatical paper in *Nature* (Szostak, 2003; Hazen et al., 2007). Szostak and his colleagues specified *functional information* in terms of a gene string as $-\log_2$ of the tiny fraction of functional sequences that have fitness values (activity of a biopolymer) greater than a specified value (Hazen et al., 2007). This is the probability that a random sequence will encode if the non-censored sequences are defined as functional, whereas the censored sequences are non-functional. With this interpretation of d_{FI}, it follows that $\boldsymbol{p}\bullet\boldsymbol{r}_y$ is the probability that a randomly drawn copy of the reservoir is functional. The following result states that d_{FI} is a quasi-metric, and its proof can be found in Appendix A.

Proposition 1. *Let $\mathscr{X}$ be a multinomial probability space, and let $\boldsymbol{p},\boldsymbol{q}\in\mathscr{X}$. If all entries of $\boldsymbol{p}$ are positive ($p_k > 0$) the distance $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ as defined above is finite, and it satisfies the properties of a quasi-metric metric. The geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$ is the directed curve*

$$l(\boldsymbol{p},\boldsymbol{q}) = \{\exp[\log(\boldsymbol{p}) + u(\log(\boldsymbol{q}) - \log(\boldsymbol{p}))];\, 0\le u\le 1\},$$

with u ranging from 0 to 1.

Note that $l(\boldsymbol{p},\boldsymbol{q})$ is not a straight line. And in spite of the fact that d_{FI} is not symmetric, the geodesic $l(\boldsymbol{q},\boldsymbol{p})$ has the same graph as $l(\boldsymbol{p},\boldsymbol{q})$, although it is traversed in the opposite direction, from $\boldsymbol{q}$ to $\boldsymbol{p}$. As mentioned in Section 2.3, in the mathematical literature asymmetric quasi-metric spaces are often described as "mountainous" spaces, since the effort of going upward to the top of a mountain is not the same as descending downhill to the starting point. Although the path from $\boldsymbol{p}$ to $\boldsymbol{q}$ is the same as the path from $\boldsymbol{q}$ to $\boldsymbol{p}$, for d_{FI}, the effort of passing through the path is not the same in both directions. The quasi-metric d_{FI} has the additional advantage of quantifying the functional information specified by Szostak, where the asymmetric metric may be interpreted as a greater cost of building up functional information, than dissolving it.

In order to motivate the mountain climbing interpretation of d_{FI}, it follows from (11) that $d_{FI}(\boldsymbol{p},\boldsymbol{q}) = \log_2\left(q_{k_1}/p_{k_1}\right)$ and $d_{FI}(\boldsymbol{q},\boldsymbol{p}) = \log_2\left(p_{k_2}/q_{k_2}\right)$, where k_1 and k_2 are the categories that maximize q_k/p_k and p_k/q_k respectively. It usually costs more to increase the frequency of a category than to decrease it. Indeed, given the sampling mechanism of Section 4.1, a decrease in frequency of a category only requires censoring of this category, whereas an increase in frequency requires that a number of other categories are censored. Since $d_{FI}(\boldsymbol{p},\boldsymbol{q}) > d_{FI}(\boldsymbol{q},\boldsymbol{p})$ means that going from $\boldsymbol{p}$ to $\boldsymbol{q}$ requires a higher, maximal frequency increase than going from $\boldsymbol{q}$ to $\boldsymbol{p}$, we may therefore associate a higher cost (or more mountain climbing) to the former change of frequencies than to the latter.

5.3.2. Functional information with mutations

A drawback of the functional information distance (11) is that $d_{FI}(\boldsymbol{p},\boldsymbol{q}) = \infty$ when $p_k = 0$ and $q_k > 0$ for at least one category k. It is not possible in this case to obtain $\boldsymbol{q}$ from $\boldsymbol{p}$ through rejection sampling alone, since category k will always be absent after censoring some of the elements of the reservoir $\boldsymbol{R}$, when no copies of k are available in $\boldsymbol{R}$ to start with**.** It is possible though to define a version of the FI distance such that a fraction $0\le\varepsilon<1$ of the difference between $\boldsymbol{p}$ and $\boldsymbol{q}$ is due to mutations, whereas the remaining differences between $\boldsymbol{p}$ and $\boldsymbol{q}$ are explained in terms of rejection sampling. If the distribution of each mutated copy is $\boldsymbol{q}$, it follows that the category distribution of the reservoir changes from $\boldsymbol{p}$ to $\boldsymbol{q}_\varepsilon = (1-\varepsilon)\boldsymbol{p} + \varepsilon\boldsymbol{q}$ due to mutations. Hence, the functional information distance required to change the mutated reservoir into $\boldsymbol{q}$ is

$$d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) = d_{FI}(\boldsymbol{q}_\varepsilon,\boldsymbol{q}), \qquad (12)$$

with $d_{FI}^{0} = d_{FI}$ corresponding to the original definition (11) of the FI-distance. Whenever $\varepsilon > 0$, $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$ is always well defined, with a range $[0,-\log_2(\varepsilon))$ when $\boldsymbol{p}$ and $\boldsymbol{q}$ vary independently in $\mathscr{X}$. The following result states that d_{FI}^{ε} is a quasi-metric, and it is proved in Appendix B:

Proposition 2. *Let $\mathscr{X}$ be a multinomial probability space, and assume $0<\varepsilon<1$. The distance $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$ in (12) is always well defined when $\boldsymbol{p},\boldsymbol{q}\in\mathscr{X}$ vary independently, with a range $[0,-\log_2(\varepsilon))$. Moreover, d_{FI}^{ε} is a quasi-metric on $\mathscr{X}$, and the geodesic $l(\boldsymbol{p},\boldsymbol{q})$ between $\boldsymbol{p}$ and $\boldsymbol{q}$ is the same as in Proposition 1.*

We want to apply the functional information distance $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y)$ to investigate how much the observed composition $\boldsymbol{q}_Y$ of categories differs from the prior $\boldsymbol{p}$**.** Recall from (7) that $\boldsymbol{q}_Y$ is a multinomial fraction, based on a sample of size M, that is an estimate of the probability vector $\boldsymbol{q}$. We may therefore view $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y)$ as an estimate of $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$. When M is of the same order as the number of categories K, it is important to adjust for potential bias of this estimate. It is proved in Appendix C that

$$E\left[d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y)\right] = d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) + C_\varepsilon(\boldsymbol{p},\boldsymbol{q})\Big/\sqrt{M} + o\left(1\Big/\sqrt{M}\right) \qquad (13)$$

as $M\to\infty$, with expectation taken with respect to multinomial variations (7) of $\boldsymbol{q}_Y$, and with $C_\varepsilon(\boldsymbol{p},\boldsymbol{q})\ge 0$ a constant defined in Appendix C. Equation (13) suggests that a bias adjusted estimate

$$\widehat{d}_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) = d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y) - C_\varepsilon(\boldsymbol{p},\boldsymbol{q}_Y)\Big/\sqrt{M} \qquad (14)$$

of $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$ might work well. However, (14) is not very useful, due to the fact that $C_\varepsilon(\boldsymbol{p},\boldsymbol{q})$ is a discontinuous function of $\boldsymbol{q}$, with $C_\varepsilon(\boldsymbol{p},\boldsymbol{q}) = 0$ for most values of $\boldsymbol{q}$ (more precisely, $C_\varepsilon(\boldsymbol{p},\boldsymbol{q}) = 0$ for all $\boldsymbol{q}$ such that the maximum of q_k/p_k is attained for a unique category k). For this reason, the bias correction term of (13) will be 0 with a very high probability. Instead of (14) we therefore suggest using parametric bootstrap (Efron and Tibshirani, 1994) in order to defined a bias corrected estimate

$$\widehat{d}_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) = d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y) - \left[E\left(d_{FI}^{\varepsilon}\left(\boldsymbol{p},\boldsymbol{q}_{\boldsymbol{Y}}^{*}\right)\right) - d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y)\right], \qquad (15)$$

of $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$. The expectation in (15) is taken with respect to variations of $\boldsymbol{q}_{\boldsymbol{Y}}^{*}\sim\mathrm{Mult}(M,\boldsymbol{q}_{\boldsymbol{Y}})/M$ when $\boldsymbol{q}_{\boldsymbol{Y}}$ is kept fixed, and it is computed with Monte Carlo simulations.

5.4. Mutual information and Rokhlin measures

Assume that $\boldsymbol{q}$ is generated from $\boldsymbol{p}$ by means of censoring indicator sampling, as described in Section 4.2. Recall that $Y\in\{0,1\}$ can be seen as a binary variable that determines whether an observation is censored or not. After some computations it follows that the mutual information distance (2)–(3) of Section 3.2 satisfies

$$
\begin{aligned}
d_{MI}(\boldsymbol{p},\boldsymbol{q}) &= E[d_{CMI}(\boldsymbol{p},\boldsymbol{q}_Y)] = H(X)+H(Y)-H(X,Y) \\
&= \sum_{k=1}^{K} p_k[r_k \log_2(r_k) + (1-r_k)\log_2(1-r_k)] \\
&\quad - r\log_2(r) - (1-r)\log_2(1-r),
\end{aligned} \tag{16}
$$

where r_k is the probability (6) of retaining category k, whereas $r = \sum_{k=1}^{K} r_k p_k$ is the probability of retaining a randomly chosen category. We interpret (16) as the expected information that the censoring mechanism provides about the frequency distribution of categories.

In Appendix D we prove the following result.

Proposition 3. *Let $\mathscr{X}$ be a multinomial probability space. The distance $d_{MI}(\boldsymbol{p},\boldsymbol{q})$ in* (16) *is always well defined when $\boldsymbol{p},\boldsymbol{q}\in\mathscr{X}$ vary independently. Moreover, d_{MI} satisfies the identity and point equality properties* (*i*) *and* (*ii*) *of* Section 2.3, *and also the triangle inequality* (*iii*), *for triples $\boldsymbol{p},\boldsymbol{q},\boldsymbol{s}$ such that q_k/p_k and s_k/q_k are maximized for the same category $k\in\{1,\ldots,K\}$.*

We conjecture that the triangle inequality of d_{MI} holds more generally, although we have not been able to prove that d_{MI} is a quasi-metric. In spite of several good properties of d_{MI} we regard the mutual information distance as somewhat less relevant than the functional information distance d_{FI}. The reason is that d_{MI} includes the censored as well as the non-censored samples $\boldsymbol{q}_0$ and $\boldsymbol{q}_1$, whereas d_{FI} only involves the more relevant non-censored sample $\boldsymbol{q}_1$.

Another quantity is the Rokhlin measure. From equation (4) it follows that

$$
\begin{aligned}
d_R(\boldsymbol{p},\boldsymbol{q}) &= H(X,Y) - 0.5[H(Y)+H(X)] \\
&= 0.5[r\log_2(r) + (1-r)\log_2(1-r)] - \sum_{k=1}^{K} p_k[0.5\log_2(p_k) \\
&\quad + r_k\log_2(r_k) + (1-r_k)\log_2(1-r_k)].
\end{aligned} \tag{17}
$$

Although this measure is non-negative, it is not appropriate to use it in our context, since $d_R(\boldsymbol{p},\boldsymbol{p}) = 0.5\,H(X) \neq 0$ and thus the identity and point inequality properties (i)-(ii) are violated.

5.5. Distances based on evolutionary waiting times

Consider the Markov model of Section 4.3, where each individual of type k dies at rate λ_k. Let $L = L(0)$ be the size of the sample drawn from the reservoir at time 0, and let

$$
T_{L,\delta}(\boldsymbol{p},\boldsymbol{q}) = \min\{t;\, d_\infty(\boldsymbol{p}_t,\boldsymbol{q}) \le \delta\} \tag{18}
$$

be the waiting time until the supremum norm between the genetic composition $\boldsymbol{p}_t = (p_{t1},\ldots,p_{tK})$ and the targeted probability vector $\boldsymbol{q}$ is at most δ, where $\delta \ge 0$ is a small number. For a general theory of evolutionary waiting times, cf. Durrett and Schmidt (2008), Hössjer et al. (2021) and references therein. Note in particular that $T_{L,\delta}(\boldsymbol{p},\boldsymbol{q}) \in [0,\infty]$ is a random quantity, with $T_{L,\delta}(\boldsymbol{p},\boldsymbol{q}) = \infty$ whenever the set in (18) is empty. However, it is possible to impose conditions on the rates λ_k such that the waiting time is deterministic and finite in the limit of large L and small δ. Indeed, it is shown in Appendix E that

$$
\lim_{\delta\to 0}\lim_{L\to\infty} T_{L,\delta}(\boldsymbol{p},\boldsymbol{q}) = d_{FI}(\boldsymbol{p},\boldsymbol{q}) \tag{19}
$$

for a specific choice of $\lambda_1, \ldots, \lambda_K$, with a weighted average equal to $1/\log_2(e)$. This implies that the functional information distance (11) can be interpreted as an evolutionary waiting time for a large population that starts with a genetic decomposition close to $\boldsymbol{p}$ and then looses individuals according to a Markov death process, with death rates λ_k that are larger for categories k with smaller values of q_k/p_k.

6. Some bioinformatical applications

6.1. Functional information distance

Of all the information-based distances of Section 5 we regard the functional information distance d_{FI} as particularly promising, since it is simple to define, has an intuitive interpretation in terms of functional information, and additionally is a quasi-metric. In this section we will apply d_{FI} to some bioinformatics datasets.

We consider a vector-valued random variable $\boldsymbol{q}_y = \boldsymbol{q} = (q_1,\ldots,q_{20})$ of amino acid probabilities, whose entries correspond to the frequencies by which each of the 20 conventional amino acids in the protein alphabet occur (for simplicity of notation the index of $\boldsymbol{q}_y$ is omitted). Let $\mathscr{X}$ be the multinomial sample space, and as mentioned in Propositions 1 and 2 of Section 5.3 the pair $(\mathscr{X}, d_{FI})$ is a quasi-metric probability space for the amino acid composition of proteins. The elements $\boldsymbol{q}$ of $\mathscr{X}$ are patterns corresponding to the composition of specific proteins, protein domains or families.

In a recent paper Thorvaldsen and Hössjer (2023) demonstrated how an underlying probability distribution of $\mathscr{X}$ may be obtained. A reference distribution $\boldsymbol{p} = (p_1,\ldots,p_{20})$ on the set of amino acids was derived directly from the genetic code, where each amino acid k is assigned a prior probability p_k proportionally to its constituting number n_k of codons (between 1 and 6), with a corresponding *self-information* $I(k) = -\log_2(n_k/61)$ between 5.93 and 3.35 bits. This distribution assigns the same probability to each of the 61 codons of the genetic code (=4^3-3, since 3 of the triplet codons of the genetic code are stop codons). It corresponds to a non-informative *prior distribution* on the set of codons and hence is a natural starting point, from 'first principles' thinking (cf. Aristotle). A uniform distribution on the set of non-stop codons can also be motivated from the Principle of Insufficient Reason (Bernoulli, 2024), to model maximal ignorance about, or maximum entropy for, the codon distribution before any data has been analyzed (Jaynes, 2003). More explicitly we have

$$
\begin{aligned}
\boldsymbol{p} &= (p_1,\ldots,p_{20}) \\
&= (4c, 6c, 2c, 2c, 2c, 2c, 2c, 4c, 2c, 3c, 6c, 2c, c, 2c, 4c, 6c, 4c, c, 2c, 4c),
\end{aligned} \tag{20}
$$

with a constant $c = 1/61$, assuming that the 20 amino acids are listed in standard order: A R N D C Q E G H I L K M F P S T W Y V. We regard $\boldsymbol{p}$ as a prior distribution of a baseline and $\boldsymbol{q}$ as the distribution of amino acid

Table 3
The table shows distance $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ of some protein sequences downloaded from the *Uniprot* database. Column 2 refers to the length M of each protein sequence, which is viewed as a sample with amino acid distribution $\boldsymbol{q} = \boldsymbol{q}_y$. The prior distribution $\boldsymbol{p}$ of (20) is the same for all proteins. The distance $d_{FI}(\boldsymbol{q},\boldsymbol{p})$ is added in column 4.

Protein	Length M (amino acids)	$d_{FI}(\boldsymbol{p},\boldsymbol{q})$	$d_{FI}(\boldsymbol{q},\boldsymbol{p})$
P01308. Insulin HUMAN	110	1.149	1.436
P95469. RecA PARACOCCUS D.	356	1.262	2.545
P00811. Beta-lactamase ECOLI	377	1.071	2.628
Q5T9A4. ATD3B HUMAN	648	1.372	1.087
P05067. A4 HUMAN	770	1.866	1.114
Uniform distribution ($q_k = \tilde{p}_k = 1/20$)	-	*1.609*	*0.976*

Note that $d_{FI}(\boldsymbol{p},\boldsymbol{q}) < d_{FI}(\boldsymbol{q},\boldsymbol{p})$ for a majority of the protein sequences in Table 3, reflecting that every sequence has some very rare amino acids (q_k small). On the other hand, $d_{FI}(\boldsymbol{p},\boldsymbol{q}) > d_{FI}(\boldsymbol{q},\boldsymbol{p})$ when $\boldsymbol{q} = (1/20,\ldots,1/20)$ has a uniform distribution. In this case $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ corresponds to an effort of increasing the frequency of the priori rarest amino acids M and W from 1/61 to 1/20, whereas $d_{FI}(\boldsymbol{q},\boldsymbol{p})$ corresponds to a smaller effort of increasing the frequency of (the apriori most abundant amino acids) R, L, and S, from 1/20 in $\boldsymbol{q}$, to their apriori frequency 6/61 in $\boldsymbol{p}$.

frequencies at which the protein sequence is sampled. Table 3 shows the distance $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ for some typical protein sequences downloaded from Uniprot database. The distance to a uniform distribution $\widetilde{p}_x = 1/\ 20$ is also added (c.f. Durston and Chiu 2007).

An orthologue protein *family* is commonly represented by the *alignment* of its sequences. The vector $\boldsymbol{q}$ may then represent the amino acid composition at a single column or *site* (i.e., *vertical*) along the multiple sequence alignment of the protein family, i.e., a matrix representation with M proteins as rows and a total of L_{proteins} (the length of the proteins) sites as columns. Such large multiple sequence alignments can be downloaded from the databases *Cath* and *Pfam.* Assuming independence between the sites, the functional information will be additive over sites, and an information profile of a protein family may be derived (Thorvaldsen and Hössjer 2023), as shown in Fig. 1.

The asymmetry of the functional information distance is revealed by the fact that $d_{FI}(\boldsymbol{q},\boldsymbol{p}) = \infty$ for a protein family like Fig. 1, for all sites with some lacking amino acids ($q_k\ = 0$). An additional illustration of the asymmetry of d_{FI} appears in Table 4, where the functional information distance is computed, for some sequences $\boldsymbol{p}'$, each one having two compositional modifications of the distribution $\boldsymbol{p}$ in (20).

Note from Table 3 that $d_{FI}(\boldsymbol{p},\boldsymbol{p}')$ increases when $\boldsymbol{p}'$ has higher frequencies of more rare amino acids like W, N and Y, compared to $\boldsymbol{p}$. In particular, note that $d_{FI}(\boldsymbol{p},\boldsymbol{p}') > d_{FI}(\boldsymbol{p}',\boldsymbol{p})$ when $\boldsymbol{p}'$**,** in comparison to $\boldsymbol{p}$, contains more of a rare amino acid and less of an abundant amino acid in (20). On the other hand, $d_{FI}(\boldsymbol{p},\boldsymbol{p}') < d_{FI}(\boldsymbol{p}',\boldsymbol{p})$ when $\boldsymbol{p}'$ contains more of some abundant amino acid and less of some rare amino acid, compared to $\boldsymbol{p}$. When $d_{FI}(\boldsymbol{p},\boldsymbol{p}') > d_{FI}(\boldsymbol{p}',\boldsymbol{p}),$ advancing from $\boldsymbol{p}$ to $\boldsymbol{p}'$ demands a higher, maximal frequency increase $\max_k p'_k/p_k$ than the maximal frequency increase $\max_k p_k/p'_k$ associated with descending from $\boldsymbol{p}'$ to $\boldsymbol{p}$. We may therefore connect a higher cost to the former change of frequencies than to the latter.

In a recent paper (Thorvaldsen and Hössjer, 2023) we have compared the results, estimated by the Functional information model and the Conditional mutual information model, for nearly 50 protein families downloaded from the *Cath* and the *Pfam* databases. The *Cath* data are based on shared 3D structure and sequence similarity, and they showed the strongest correlations.

The Functional information of a 3-dimensional categorical distribution may be visualized, and in Fig. 2 we present an illustrative example. In the figure all instances of compositional data are represented as points in the probability simplex, for the specific case $\boldsymbol{p} = (0.50, 0.242, 0.258)$, located in the dark blue area. The information distance increases as indicated by the color bar.[1]

6.2. Quantifying physiochemical properties of amino acids

The amino acids also have many different physicochemical properties (Gromiha et al. 1999), and the amino acid alphabet can be reduced to data based on the physicochemical properties of each acid. All property scales assign specific numerical values to each of the 20 amino acids (e,g., its polarity, hydrophobicity or volume). This vector may be normalized and directly transformed to a *probability* vector $\boldsymbol{v}$ where the 20 entries add up to 1.

Furthermore, it is straightforward to define a normalized scalar dot product, σ, between two probability vectors $\boldsymbol{v} = (v_1, \ldots, v_{20})$ and $\boldsymbol{p} = (p_1, \ldots, p_{20}),$ that represent normalized physicochemical properties and prior probabilities of all amino acids. This quantity σ is the scalar product divided by the product of the Euclidean norms of the two vectors:

$$\sigma(\boldsymbol{p},\boldsymbol{v}) = \frac{\boldsymbol{v} \bullet \boldsymbol{p}}{\|\boldsymbol{v}\| \bullet \|\boldsymbol{p}\|}$$

This normalized scalar product of two probability vectors measures their *similarity*, with σ =*1* being most similar (the vectors are identical, corresponding to a scenario where the property v_k of an amino acid k is proportional to its prior abundance p_k), whereas σ =*0* indicates that none of the amino acids in one vector occurs in the other (orthogonality, corresponding to a scenario where only the apriori absent amino acids, $p_k\ = 0$, can have nonzero values $v_k > 0$ of the property of interest). The similarity measure is related but not equivalent to the correlation coefficient between the sequences in $\boldsymbol{v}$ and $\boldsymbol{p}$.

It can be seen that the similarity measure satisfies $\sigma(a\boldsymbol{p}, b\boldsymbol{v}) = \sigma(\boldsymbol{p},\boldsymbol{v})$ for any non-negative constants $a > 0, b > 0$. This implies that $\sigma(\boldsymbol{p},\boldsymbol{v})$ is not dependent on normalizing the physiochemical property vector $\boldsymbol{v}$ to a probability vector. On the other hand, the Hellinger distance of Section 2.3 between $\boldsymbol{v}$ and $\boldsymbol{p}$ can be expressed in terms of the similarity measure as

$$d_2(\boldsymbol{p},\boldsymbol{v}) = \left(1 - \sigma\left(\sqrt{\boldsymbol{p}}\,,\sqrt{\boldsymbol{v}}\right)\right)^{1/2},$$

with $\sqrt{\boldsymbol{p}} = \left(\sqrt{p_1}\,, \ldots, \sqrt{p_{20}}\right)$ and $\sqrt{\boldsymbol{v}}\ = \left(\sqrt{v_1}\,, \ldots, \sqrt{v_{20}}\right)$, only when both of $\boldsymbol{v}$ and $\boldsymbol{p}$ are probability vectors. The Hellinger distance can be seen as a probabilistic analogue of Euclidean distance.

Since $\boldsymbol{v}$ and $\boldsymbol{p}$ refer to different quantities (a vector of normalized values of a physiochemical property and a distribution respectively), it is not possible to obtain one of these two vectors from the other by means of a sampling procedure. For this reason, we use the symmetric measures ρ and d_2 to quantify similarity and distance respectively, between $\boldsymbol{v}$ and $\boldsymbol{p}$**.** By application of these vector space techniques, the physicochemical properties that are most similar (closest) and most orthogonal (most distant) to distribution (20) are quantified and listed in Table 5. One may hypothesize that properties (like compressibility) with large similarity scores σ are larger for more abundant amino acids in order for the protein to fold well.

7. Concluding remarks

7.1. Quasi-metrics

The multinomial probability space, and the functional information (*FI*) distance estimated by rejection sampling in the present paper constitutes a quasi-metric space for multicategory data. It measures information in a landscape of mountains, where it takes more effort to go up than down. This corresponds with an intuitive understanding of information in a meaningful way, where a larger effort, or more information, is obtained when initially a large collection of categories is given, with a composition according to prior expectations, and then a smaller fraction of categories remains after censoring, with the empirically observed composition of categories. The model has the advantage of quantifying functional information specified by Szostak (2003). Together with an underlying prior distribution derived from the genetic code, the new model has been applied with success on real data from proteins and large multiple sequence alignments.

7.2. Impact of mutations

The *FI* quasi-metric incorporates selection as the mechanism which causes the allelic composition of sampled data to be different than prior expectations, assuming that alleles with higher rejection probabilities are less fit. However, in Section 5.3.2 we added, as a second mechanism of the sampling procedure, a fraction ε of mutations in order to obtain a more robust functional information distance d_{FI}^{ε}.

An important aspect in the definition (12) of d_{FI}^{ε} is the assumption that all mutated categories are distributed as $\boldsymbol{q}$**.** This corresponds to a mutational distribution that for a given value of the mutation probability ε minimizes the functional information associated with transforming $\boldsymbol{p}$ to $\boldsymbol{q}$**.** An alternative approach would be to introduce a mutational matrix $\Pi = (\pi_{kl})$ of order K, with π_{kl} the probability that the

[1] This specific case $\boldsymbol{p} = (0.50, 0.242, 0.258)$ is based on a distribution from Premier League football (home wins, draws, away wins).

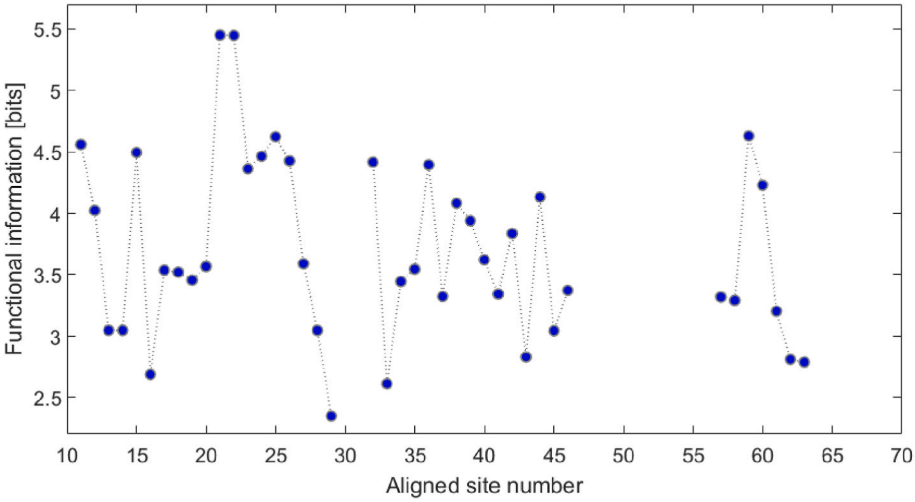

Fig. 1. The information profile shows how information values $d_{FI}(\boldsymbol{p}, \boldsymbol{q})$ vary along the sites of a protein alignment, with $M = 2295$ proteins in the alignment. At each site $d_{FI}(\boldsymbol{p}, \boldsymbol{q})$ quantifies the distance between the proteins at this site. The site dependence of $d_{FI}(\boldsymbol{p}, \boldsymbol{q})$ is due to the fact that the amino acids of the proteins at each particular site is viewed as a sample y, with a site-dependent, observed amino acid distribution $\boldsymbol{q} = \boldsymbol{q}_y$, whereas $\boldsymbol{p}$ in (20) is fixed. The peak at sites 21 and 22 is dominated by amino acid M with prior probability $p_k = 1/61$, in spite of the fact that M is not fully conserved at any of these two sites ($q_k < 1$ and hence $d_{FI}(\boldsymbol{p}, \boldsymbol{q}) < \log_2(61) = 5.93$). A minor correction factor from Appendix B in Thorvaldsen and Hössjer (2023) is also included. Sites containing gaps are unfilled.

Table 4
The table illustrates asymmetries of $d_{FI}(\boldsymbol{p}, \boldsymbol{p}')$ versus $d_{FI}(\boldsymbol{p}', \boldsymbol{p})$ based on various choices of $\boldsymbol{p}'$, each one having two compensating modifications of $\boldsymbol{p}$ in (20). Note in particular that $d_{FI}(\boldsymbol{p}, \boldsymbol{p}')$ is larger than, equal to, or less than $d_{FI}(\boldsymbol{p}', \boldsymbol{p})$ for the top, middle (rows 8-10), and bottom rows respectively.

$\boldsymbol{p}$’ modifications (amino acids)	$d_{FI}(\boldsymbol{p}, \boldsymbol{p}')$	$d_{FI}(\boldsymbol{p}', \boldsymbol{p})$
R: $6c \rightarrow 4c$, W : $1c \rightarrow 3c$	1.5850	0.5850
R: $6c \rightarrow 5c$, W : $1c \rightarrow 2c$	1	0.2630
A: $4\ c \rightarrow 3c$, W : $1c \rightarrow 2c$	1	0.4150
I: $3\ c \rightarrow 2c$, W : $1c \rightarrow 2c$	1	0.5850
A: $4\ c \rightarrow 3c$, N : $2c \rightarrow 3c$	0.5850	0.4150
I : $3c \rightarrow 4c$, R: $6\ c \rightarrow 5c$	0.4150	0.2630
A: $4\ c \rightarrow 5c$, R : $6c \rightarrow 5c$	0.3219	0.2630
Y : $2c \rightarrow 1c$, W: $1\ c \rightarrow 2c$	1	1
Y: $2\ c \rightarrow 3c$, I : $3c \rightarrow 2c$	0.5850	0.5850
I: $3\ c \rightarrow 4c$, A : $4c \rightarrow 3c$	0.4150	0.4150
A: $4\ c \rightarrow 6c$, I : $3c \rightarrow 1c$	0.5850	1.5850
A: $4\ c \rightarrow 5c$, N : $2c \rightarrow 1c$	0.3219	1
R: $6c \rightarrow 7c$, D : $2c \rightarrow 1c$	0.2224	1
A: $4\ c \rightarrow 5c$, I : $3c \rightarrow 2c$	0.3219	0.5850
R: $6c \rightarrow 7c$, A : $4c \rightarrow 3c$	0.2224	0.4150

mutated version of category k is l. The corresponding functional information distance would be

$$d_{FI}^{\varepsilon,\Pi}(\boldsymbol{p}, \boldsymbol{q}) = d_{FI}((1-\varepsilon)\boldsymbol{p} + \varepsilon\boldsymbol{p}\Pi, \boldsymbol{q}).$$

It can be shown however that the distance $d_{FI}^{\varepsilon,\Pi}(\boldsymbol{p}, \boldsymbol{q})$ does not satisfy the triangle inequality and hence does not quality as a quasi metric. It differs from (12) in replacing the target dependent mutational distribution $\boldsymbol{q}$ with the target-independent mutational distribution $\boldsymbol{p}\Pi$**.** Examples of mutational matrices Π for the set of $K = 61$ non-stop codons appear in Thorvaldsen (2016).

7.3. Impact of genetic drift

In this article we included two forces of genetic change, a sampling mechanism with selection and mutations, in order to define a functional information distance $d_{FI}(\boldsymbol{p}, \boldsymbol{q})$ between a prior distribution $\boldsymbol{p}$ of categories or alleles and their observed empirical distribution $\boldsymbol{q}$**.** In Sections 4.3 and 5.5 we added a dynamic perspective to the sampling mechanism, with allelic distribution $\boldsymbol{p}_t$ at time $t \geq 0$, $\boldsymbol{p} = \boldsymbol{p}_0$ the allelic distribution at time 0 and $\boldsymbol{q} = \boldsymbol{p}_T$ the allelic distribution after some waiting time T. Since strong selection (and no mutations) was assumed in Section 5.5, $\boldsymbol{p}_t$ was, in the limit of large populations, a deterministic function of t, and this made it possible to characterize $d_{FI}(\boldsymbol{p}, \boldsymbol{q})$ as a nonrandom waiting time.

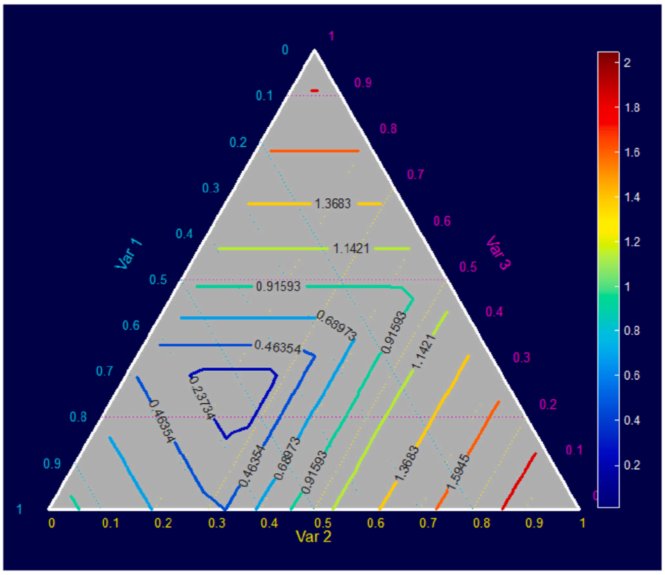

Fig. 2. An example that depicts compositional data and its representation in the probability simplex for the case $\boldsymbol{p} = (0.50, 0.242, 0.258)$, cf. Theune (2023).

A possible extension of our work is to include genetic drift and regard $\boldsymbol{p}_t$ as a stochastic process, such as the composition of a K-allelic Wright-Fisher or Moran population whose size $L(t)$ possibly varies with time t (Crow and Kimura, 1970; Durrett, 2008). Note that the functional information distance $d_{FI}(\boldsymbol{p}, \boldsymbol{p}_t)$, with $\boldsymbol{p} = \boldsymbol{p}_0$, will also be a stochastic process when genetic drift is accounted for. It is of interest then to study how $d_{FI}(\boldsymbol{p}, \boldsymbol{p}_t)$ evolves over time due to the forces of selection, mutations and genetic drift. In particular, the impact of genetic drift will not vanish in the diffusion limit of large populations if selection coefficients and mutation probabilities are inversely proportional to population size. If additionally, the mutational matrix Π between the K alleles is irreducible, it is known that under mild conditions $\boldsymbol{p}_t$ converges weakly as $t \rightarrow \infty$

Table 5
The table shows the similarity $\sigma(\mathbf{p},\mathbf{v})$ and the Hellinger distance $d_2(\mathbf{p},\mathbf{v})$ between $\mathbf{p}$ in (20) and some of the physicochemical properties $\mathbf{v}$ of the amino acids. The properties with the highest (lowest) detected similarity (Hellinger distance) scores are shown in the upper part of the table, whereas the corresponding lowest (highest) scores are presented in the lower part.

Property	Similarity	Hellinger distance
Unfolding entropy change (Oobatake and Ooi, 1993)	0.893	0.219
Compressibility (Iqbal and Verrall, 1988)	0.885	0.222
Unfolding entropy changes of chain (Oobatake and Ooi, 1993)	0.881	0.225
Isoelectric point (Zimmerman et al., 1968)	0.855	0.241
Average flexibility indices (Bhaskaran and Ponnuswamy, 1988)	0.851	0.248
Equilibrium constant (Zimmerman et al., 1968)	0.850	0.255
Self-information ($-\log_2(\mathbf{p})$)	0.508	0.492
Metabolic costs (Akashi and Gojobori, 2002)	0.484	0.500
Polarity (Zimmerman et al., 1968)	0.439	0.613

to a limiting distribution Q on the $(K-1)$-dimensional simplex $\mathscr{X}$ (Wright, 1949; Shiga, 1981; Barbour et al., 2000). This implies that the functional information distance $d_{FI}(\mathbf{p},\mathbf{p}_t)$ converges weakly to $d_{FI}(\mathbf{p},\mathbf{q})$ as $t\rightarrow\infty$, where $\mathbf{q}$ is a random variable with distribution Q. It is of interest then to study how the distribution of $d_{FI}(\mathbf{p},\mathbf{q})$ depends on $\mathbf{p}$, genetic drift, mutation probabilities and selection coefficients. If $\mathbf{p}$ is far away from the region where Q puts most of its probability mass, we expect that a large amount of functional information is needed in order to change the allele frequency vector from $\mathbf{p}$ to the equilibrium distribution Q.

7.4. *Symmetric similarity/distance measures and ordinary metrics*

We did not use a quasi-metric approach for quantifying similarity/distance between qualitatively different variables, such as between the distribution $\mathbf{p}$ and physiochemical properties $\mathbf{v}$ of amino acids. The reason is that quasi-metrics are closely related to paths in a landscape with mountains and valleys, and there is no clearly defined path in such a landscape, in terms of a sampling mechanism with selection and mutations, from $\mathbf{p}$ to $\mathbf{v}$. For this reason, we used more robust and symmetric measures in order to quantify similarity/distance between vectors of different origin.

7.5. *Practical use of the FI quasi-metric*

We believe that the *FI* quasi-metric can also be applied to other types of biomolecules than proteins, in order to quantify how much the composition of their building blocks (categories) differs from prior expectations. In this context, it is of interest to apply the functional information quasi-metric to structures having different types of building blocks such as amino acids, carbohydrates, and lipids.

As mentioned in Section 2.2, we hypothesize that the FI quasi-metric can be used to pinpoint, from empirical data, alleles with a selective advantage, with higher frequencies compared to prior expectations. We believe that this opens up for finding functional reasons for the fitness gain of these alleles, when their role in the cell is understood at a molecular level.

Data accessibility

The Matlab code is available in the next version of the *DeltaProt* toolbox (Thorvaldsen et al. 2010). https://doi.org/10.1186/1471-2105-11-573.

CRediT authorship contribution statement

Steinar Thorvaldsen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Ola Hössjer:** Writing – review & editing, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

Conflicts of interest: None declared.

Data availability

Data will be made available on request.

Acknowledgement

We thank Professor Boris Kruglikov for valuable comments and references concerning the Rokhlin metric. We are also grateful to two anonymous reviewers, whose comments considerably improved the manuscript.

Appendix A. Quasi-metric properties of the Functional information distance

In order to verify Proposition 1, that the functional information distance $d_{FI}(\mathbf{p},\mathbf{q})$ in (11), between the multinomial vectors $\mathbf{p} = (p_1,\ldots,p_K)$ and $\mathbf{q} = (q_1,\ldots,q_K)$, is a quasi-metric, we need to verify the requirements of a quasi-metric in Section 2.3. This includes showing that the range is nonnegative, as well as establishing the three properties (i) identity, (ii) point equality, and (iii) triangle equality. We will also verify that the symmetry condition (iv) fails, from which it follows that d_{FI} is not a metric.

Range

We have that $d_{FI}(\mathbf{p},\mathbf{q}) = [0, -\log_2(\min(p_k)\ )] \subset [0,\infty)$ for any choice of $\mathbf{q} \in \mathscr{X}$, as long as $\min(p_k) > 0$.

Identity

Since $\mathbf{r} = \mathbf{1}$ when $\mathbf{p} = \mathbf{q}$, $d_{FI}(\mathbf{p},\mathbf{p}) = -\log_2(\mathbf{p}\bullet\mathbf{1}) = -\log_2\left(\sum_{k=1}^{K} p_k\right) = -\log_2(1) = 0$ follows.

Triangle inequality

Recall from equation (11) that

$$d_{FI}(\boldsymbol{p},\boldsymbol{q}) = -\log_2(\boldsymbol{p}\bullet\boldsymbol{r}) = \log_2 \max\left(\frac{q_1}{p_1},\dots,\frac{q_K}{p_K}\right) - \log_2\sum_{k=1}^{K} q_k = \max\log_2\left(\frac{q_1}{p_1},\dots,\frac{q_K}{p_K}\right),$$

where the second equality follows from the definition of $\boldsymbol{r} = (r_1,\dots,r_K)$ in (6), whereas in the third equality we make use of $\sum^{K} q_k = 1$. Now assume that the output of the first sampling mechanism, that generated $\boldsymbol{q}$, is the input of a second sampling mechanism that generates a new $\boldsymbol{s}$. That is, the observed amino acid distribution $\boldsymbol{s} = (s_k)$ is obtained from a pool of amino acids with distribution $\boldsymbol{q}$, corresponding to a distance $d_{FI}(\boldsymbol{q},\boldsymbol{s})$. On the other hand, a combined sampling procedure, with a pool of amino acids with frequencies $\boldsymbol{p}$, and observed frequencies $\boldsymbol{s}$, corresponds with $d_{FI}(\boldsymbol{p},\boldsymbol{s})$. In order to demonstrate the triangle inequality for d_{FI} we have to prove

$$d_{FI}(\boldsymbol{p},\boldsymbol{s}) \le d_{FI}(\boldsymbol{p},\boldsymbol{q}) + d_{FI}(\boldsymbol{q},\boldsymbol{s}).$$

But this follows by taking the base 2 logarithm of the inequality

$$\max\left(\frac{s_1}{p_1},\dots,\frac{s_K}{p_K}\right) \le \max\left(\frac{q_1}{p_1},\dots,\frac{q_K}{p_K}\right) \times \max\left(\frac{s_1}{q_1},\dots,\frac{s_K}{q_K}\right).$$

Hence, this measure of functional information satisfies the triangle inequality since all components p_k and q_k of $\boldsymbol{p}$ and $\boldsymbol{q}$ are non-negative.

Point equality.

Suppose by contraposition that $\boldsymbol{p} \ne \boldsymbol{q}$. Since $\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} q_k = 1$, it follows that at least one of the elements of $\boldsymbol{r}$, say r_k, is strictly less than 1. Since also $p_k > 0$ by assumption, it follows that $\boldsymbol{p}\bullet\boldsymbol{r} \le \boldsymbol{p}\bullet\boldsymbol{1} - p_k(1-r_k) < 1$. Consequently $d_{FI}(\boldsymbol{p},\boldsymbol{q}) = -\log_2(\boldsymbol{p}\bullet\boldsymbol{r}) > 0$.

Symmetry.

It is easily seen that the measure of functional information is not symmetric and thus does not qualify as a regular metric. As simple example with $K = 2$ is $\boldsymbol{p} = (1/3, 2/3)$ and $\boldsymbol{q} = (1/2, 1/2)$. It can be shown that $d_{FI}(\boldsymbol{p},\boldsymbol{q}) = -\log_2(2/3) = 0.585$ whereas $d_{FI}(\boldsymbol{q},\boldsymbol{p}) = -\log_2(3/4) = 0.415$.

We end this appendix by motivating why $l(\boldsymbol{p},\boldsymbol{q})$, defined in Proposition 1, is a geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$. It can be seen that the triangle inequality holds with equality, for any three points along $l(\boldsymbol{p},\boldsymbol{q})$. From this it follows that the length $L[l(\boldsymbol{p},\boldsymbol{q})]$ of $l(\boldsymbol{p},\boldsymbol{q})$, in the sense of differential geometry, is $d(\boldsymbol{p},\boldsymbol{q})$, whereas the length $L[\gamma]$ of any other path γ between $\boldsymbol{p}$ and $\boldsymbol{q}$ exceeds $d(\boldsymbol{p},\boldsymbol{q})$. Since $l(\boldsymbol{p},\boldsymbol{q})$ is the path γ from $\boldsymbol{p}$ to $\boldsymbol{q}$ that minimizes $L[\gamma]$, it is a geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$.

Appendix B. Quasi-metric properties of the Functional information distance with mutations

In order to verify Proposition 2, we start by rewriting the functional information distance with mutations as

$$d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) = \log_2 \max_k \frac{q_k}{(1-\varepsilon)p_k + \varepsilon q_k} = \log_2 \frac{\max_k(q_k/p_k)}{1+\varepsilon[\max_k(q_k/p_k)-1]} = f(d_{FI}(\boldsymbol{p},\boldsymbol{q})),$$

where the function $f : [0,\infty) \to [0, -\log_2(\varepsilon))$ is defined through

$$f(d) = \log_2 g(2^d),$$

with $g : [1,\infty) \to [1,\varepsilon^{-1})$ is the strictly increasing function

$$g(z) = \frac{z}{1+\varepsilon(z-1)}.$$

After some computations it can be seen that the first two derivatives of f satisfy

$$f'(d) = \frac{2^d g'(2^d)}{g(2^d)} = \frac{1-\varepsilon}{1+\varepsilon(2^d-1)}$$

and

$$f''(d) = -\log(2)\bullet\frac{2^d\varepsilon(1-\varepsilon)}{\left[1+\varepsilon(2^d-1)\right]^2} < 0$$

respectively. From this it follows that f is a strictly increasing and strictly concave function on $[0,\infty)$ with $f(0) = 0$, $f'(0) = 1-\varepsilon$ and $f(\infty) = \lim_{d\to\infty} f(d) = -\log_2(\varepsilon)$.

It follows from (12) and the fact that $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ can be made arbitrarily large as $\boldsymbol{p}$ and $\boldsymbol{q}$ vary independently in $\mathscr{X}$, that the range of $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$ is $f(\infty) = -log_2(\varepsilon)$. Identity (i) and point equality (ii) of d_{FI}^{ε} follow from the identity and point equality properties of d_{FI}, and the fact that $f(0) = 0$ and $f(d) > 0$ for $d > 0$ respectively. Finally, the triangle inequality (iii) of d_{FI}^{ε} follows from

$$d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{s}) = f(d_{FI}(\boldsymbol{p},\boldsymbol{s})) \le f(d_{FI}(\boldsymbol{p},\boldsymbol{q}) + d_{FI}(\boldsymbol{q},\boldsymbol{s})) \le f(d_{FI}(\boldsymbol{p},\boldsymbol{q})) + f(d_{FI}(\boldsymbol{q},\boldsymbol{s})) = d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) + d_{FI}^{\varepsilon}(\boldsymbol{q},\boldsymbol{s}),$$

where in the first and fourth steps we invoked the definition (12) of d_{FI}^{ε}, in the second step we made use of the triangle inequality of d_{FI}, and in the third step we used that f is a strictly concave function on $[0,\infty)$ with $f(0) = 0$.

In order to prove that $l(\boldsymbol{p},\boldsymbol{q})$, defined in Proposition 1, is a geodesic from $\boldsymbol{p}$ to $\boldsymbol{q}$ under d_{FI}^{ε}, we let $L^{\varepsilon}[\gamma]$ refer to the length of a path γ from $\boldsymbol{p}$ to $\boldsymbol{q}$ under d_{FI}^{ε}, whereas $L[\gamma]$ is the length of such a path under d_{FI}. Since

$$L^{\varepsilon}[\gamma] = f'(0)L[\gamma] = (1-\varepsilon)L[\gamma],$$

it follows that $L^{\varepsilon}[\gamma]$ must be minimized by the same path from $\boldsymbol{p}$ to $\boldsymbol{q}$ as $L[\gamma]$. We already know from the proof of Proposition 1 that $l(\boldsymbol{p},\boldsymbol{q})$ is the path from $\boldsymbol{p}$ to $\boldsymbol{q}$ that minimizes $L[\gamma]$. Hence, $l(\boldsymbol{p},\boldsymbol{q})$ must also be the path from $\boldsymbol{p}$ to $\boldsymbol{q}$ that minimizes $L^{\varepsilon}[\gamma]$. By definition, $l(\boldsymbol{p},\boldsymbol{q})$ must therefore be the geodesic between $\boldsymbol{p}$ to $\boldsymbol{q}$ under d_{FI}^{ε}.

Appendix C. Bias of Functional information distance with mutations for small samples

In this appendix we will verify (13), the bias formula for the functional information distance d_{FI}^{ε} of equation (12). This will be done in two steps, where firstly we prove the bias formula

$$E[d_{FI}(\boldsymbol{p},\boldsymbol{q}_Y)] = d_{FI}(\boldsymbol{p},\boldsymbol{q}) + C_0(\boldsymbol{p},\boldsymbol{q}) \Big/ \sqrt{M} + o\left(1 \Big/ \sqrt{M}\right) \tag{21}$$

for the functional information distance (11) without mutations, as the sample size $M \to \infty$. This corresponds to the special case $\varepsilon = 0$ of (13). The general case (13) with mutations ($0 < \varepsilon < 1$) will then be derived as a corollary of (21).

In order to prove (21), it is convenient to first rewrite the ratios between the coordinates of $\boldsymbol{q}_Y$ and $\boldsymbol{p}$ as

$$\frac{q_{kY}}{p_k} = \frac{q_k}{p_k} + \frac{\delta_k}{\sqrt{M}}. \tag{22}$$

Due to the Central Limit Theorem and the covariance matrix of the multinomial distribution, for large M, the vector $\delta = (\delta_1, \ldots, \delta_K)$ approximately has a multivariate normal distribution with expected value $0 = (0, \ldots, 0)$ and a covariance matrix $\Sigma = (\Sigma_{kl})$ whose diagonal elements are given by $\Sigma_{kk} = q_k(1-q_k)/p_k^2$, whereas the off-diagonal elements are $\Sigma_{kl} = -q_k q_l/(p_k p_l)$. Because of the form of the covariance matrix Σ, it is possible to rewrite the coordinates of δ as

$$\delta_k \approx \frac{\sqrt{q_k}Z_k}{p_k} - \frac{q_k \sum_{l=1}^{K} \sqrt{q_l} Z_l}{p_k},$$

where $Z_1, \ldots, Z_K$ are independent and identically distributed random variables with a standard normal distribution $N(0,1)$. When M is large, it follows from (22) and a Taylor expansion of $d_{FI}(\boldsymbol{p},\boldsymbol{q}_Y)$ around $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ that

$$d_{FI}(\boldsymbol{p},\boldsymbol{q}_Y) \approx d_{FI}(\boldsymbol{p},\boldsymbol{q}) + \frac{1}{\log(2) \bullet 2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})}} \bullet \frac{1}{\sqrt{M}} \bullet \max_{k \in J} \delta_k, \tag{23}$$

where

$$J = \left\{ k; \frac{q_k}{p_k} = \max_{1 \le l \le K} \frac{q_l}{p_l} = 2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})} \right\}$$

is the set of indexes k that maximize the ratios between the components of $\boldsymbol{q}$ and $\boldsymbol{p}$. Using the fact that $q_k/p_k = 2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})}$ for all $k \in J$, we find that

$$\max_{k \in J} \delta_k = \max_{k \in J} \frac{\sqrt{q_k} Z_k}{p_k} - 2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})} \sum_{l=1}^{K} \sqrt{q_l} Z_l. \tag{24}$$

Taking expectation in (23), and using (24), it follows that (21) holds with

$$C_0(\boldsymbol{p},\boldsymbol{q}) = \frac{E\left[\max_{k \in J}\left(\sqrt{q_k}\, Z_k / p_k\right)\right]}{\log(2) \bullet 2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})}} = \frac{-\int_{-\infty}^{0} F(t)dt + \int_{0}^{\infty} [1 - F(t)]dt}{\log(2) \bullet 2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})}}, \tag{25}$$

and

$$F(t) = \prod_{k \in J} \Phi\left(p_k t \Big/ \sqrt{q_k}\right)$$

the distribution function of $\max_{k \in J}\left(\sqrt{q_k}\, Z_k / p_k\right)$, whereas Φ is the distribution function of a standard normal $N(0,1)$ random variable. In particular, it follows from (25) that $C_0(\boldsymbol{p},\boldsymbol{q}) = 0$ if and only if $|J| = 1$, that is, when q_k/p_k has a unique maximizer. The reason is that F in this case has a symmetric normal distribution, so that the sum of the two integrals of the numerator of (25), vanishes.

In order to derive (13) from (21), we use the characterization of $d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q})$, in the first displayed equation of Appendix C, in terms of the function f. A first order Taylor expansion of f around the point $d_{FI}(\boldsymbol{p},\boldsymbol{q})$ leads to

$$E\left[d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}_Y)\right] = E[f(d_{FI}(\boldsymbol{p},\boldsymbol{q}_Y))] \approx f(d_{FI}(\boldsymbol{p},\boldsymbol{q})) + f'(d_{FI}(\boldsymbol{p},\boldsymbol{q})) \frac{C_0(\boldsymbol{p},\boldsymbol{q})}{\sqrt{M}} = d_{FI}^{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) + \frac{C_{\varepsilon}(\boldsymbol{p},\boldsymbol{q})}{\sqrt{M}},$$

where

$$C_{\varepsilon}(\boldsymbol{p},\boldsymbol{q}) = f'(d_{FI}(\boldsymbol{p},\boldsymbol{q}))C_0(\boldsymbol{p},\boldsymbol{q}) = \frac{1-\varepsilon}{1 + \varepsilon\left(2^{d_{FI}(\boldsymbol{p},\boldsymbol{q})} - 1\right)} \bullet C_0(\boldsymbol{p},\boldsymbol{q}). \tag{26}$$

Note in particular that $C_\varepsilon(\boldsymbol{p},\boldsymbol{q}) = 0$ if and only if $C_0(\boldsymbol{p},\boldsymbol{q}) = 0$, that is, if and only if $|J| = 1$. Whenever $C_0(\boldsymbol{p},\boldsymbol{q}) > 0$, we have that $C_\varepsilon(\boldsymbol{p},\boldsymbol{q}) > 0$ is a strictly decreasing function of ε that approaches 0 as $\varepsilon \to 1$. Hence, the larger the fraction of mutations, the less important bias correction of $d^{\varepsilon}_{FI}(\boldsymbol{p},\boldsymbol{q}_Y)$ is.

Appendix D. Quasi-metric properties of the Mutual information distance

In order to verify that d_{MI} satisfies the conditions of Proposition 3 we proceed as in Appendix A, and show that (0) its range is nonnegative, as well as establishing the (i) identity, (ii) point equality, and (iii) parts of the triangle equality properties of Section 2.3. To this end, we rewrite the definition of d_{MI} in (16) as

$$d_{MI}(\boldsymbol{p},\boldsymbol{q}) = \sum_{k=1}^{K} p_k[\rho(r) - \rho(r_k)],$$

where $\rho(x) = -x\log_2(x) - (1-x)\log_2(1-x)$ is non-negative and strictly concave on (0,1), whereas $r = \sum_{k=1}^{K} p_k r_k$. The concavity of ρ and Jensen's Inequality imply that $d_{MI}(\boldsymbol{p},\boldsymbol{q}) \geq 0$, with equality if and only if $r_k = r$ for $k = 1,\ldots,K$, which in turn, because of (6), requires $r = 1$ and $\boldsymbol{p} = \boldsymbol{q}$. Thus we have verified (0), (i), and (ii). In order to prove the triangle inequality for d_{MI} under the conditions of Proposition 3, we have to demonstrate that

$$d_{MI}(\boldsymbol{p},\boldsymbol{s}) \leq d_{MI}(\boldsymbol{p},\boldsymbol{q}) + d_{MI}(\boldsymbol{q},\boldsymbol{s}). \tag{27}$$

As in Appendix A we assume that the output of the first sampling mechanism (with non-rejection probabilities $r_k = rq_k/p_k$) that generated $\boldsymbol{q}$, is the input of a second sampling mechanism (with non-rejection probabilities $v_k = vs_k/q_k$, where $v = \sum_{k=1}^{K} q_k v_k$) that generates a new amino acid frequency vector $\boldsymbol{s}$. That is, the observed amino acid distribution $\boldsymbol{s} = (s_k)$ is obtained from a pool of amino acids with distribution $\boldsymbol{q}$, corresponding to a distance $d_{MI}(\boldsymbol{q},\boldsymbol{s})$. On the other hand, a combined sampling procedure, with a pool of amino acids with frequencies $\boldsymbol{p}$, and observed frequencies $\boldsymbol{s}$, corresponds to non-rejection probabilities $w_k = ws_k/p_k$, where $w = \sum_{k=1}^{K} p_k w_k$, and a distance $d_{MI}(\boldsymbol{p},\boldsymbol{s})$. Note in particular that the combined non-rejection probabilities satisfy

$$w_k = \frac{w}{rv} r_k v_k \geq r_k v_k, \tag{28}$$

for all $k = 1,\ldots,K$, with equality if and only of if

$$r_{k_0} = v_{k_0} = 1 \tag{29}$$

for some $k_0 \in \{1,\ldots,K\}$. As stated in Proposition 3, we will prove the triangle inequality (27) when (29) holds. Assume $X \sim \mathrm{Cat}(\boldsymbol{p})$, and let $Y, Z, U \in \{0,1\}$ be non-censoring indicators for the first, second and combined sampling mechanism respectively. Note also that $U = YZ$ whenever (29) holds, since (29) implies $P(Y = 1|X = k) = r_k$, $P(Z = 1|X = k, Y = 1) = v_k$ and $P(U = 1|X = k) = w_k = r_k v_k$. Without loss of generality, we assume that $Z = 0$ whenever $Y = 0$. It follows from (16) that

$$\begin{aligned} d_{MI}(\boldsymbol{p},\boldsymbol{s}) = H(YZ) - E[H(YZ|X)] \leq H(Y,Z) - E[H(Y,Z|X)] = H(Y) + E[H(Z|Y)] - E[H(Y|X)] - E[H(Z|X,Y)] = d_{MI}(\boldsymbol{p},\boldsymbol{q}) + P(Y=1)d_{MI}(\boldsymbol{q},\boldsymbol{s}) \\ \leq d_{MI}(\boldsymbol{p},\boldsymbol{q}) + d_{MI}(\boldsymbol{q},\boldsymbol{s}), \end{aligned}$$

so that the triangle inequality holds whenever (29) is satisfied. In the general case, we will compare

$$d_{MI}(\boldsymbol{p},\boldsymbol{s}) = \sum_{k=1}^{K} p_k[\rho(w) - \rho(w_k)],$$

with

$$\widetilde{d}_{MI}(\boldsymbol{p},\boldsymbol{s}) = \sum_{k=1}^{K} p_k[\rho(rv) - \rho(r_k v_k)] = \sum_{k=1}^{K} p_k[\rho(cw) - \rho(cw_k)],$$

where $rv = r\sum_{k=1}^{K} q_k v_k = \sum_{k=1}^{K} p_k r_k v_k$ and $c = r_k v_k / w_k \in (0,1]$. Note that (29) is equivalent to $c = 1$. It follows from the proof of the triangle inequality, when (29) holds, that

$$\widetilde{d}_{MI}(\boldsymbol{p},\boldsymbol{s}) \leq d_{MI}(\boldsymbol{p},\boldsymbol{q}) + d_{MI}(\boldsymbol{q},\boldsymbol{s}) - (1-r)d_{MI}(q,\boldsymbol{s}),$$

since $r = P(Y = 1)$. Note also that

$$d_{MI}(\boldsymbol{p},\boldsymbol{s}) - \widetilde{d}_{MI}(\boldsymbol{p},\boldsymbol{s}) = \sum_{k=1}^{K} p_k[g(w) - g(w_k)] \geq 0,$$

where the last step follows from Jensen's Inequality, since

$$g(x) = \rho(x) - \rho(cx)$$

is a strictly concave function of x. Putting things together we find that

$$d_{MI}(\boldsymbol{p},\boldsymbol{s}) \le d_{MI}(\boldsymbol{p},\boldsymbol{q}) + d_{MI}(\boldsymbol{q},\boldsymbol{s}) - (1-r)d_{MI}(\boldsymbol{q},\boldsymbol{s}) + \sum_{k=1}^{K} p_k[g(w)-g(w_k)]. \tag{30}$$

Equation (30) does not prove the triangle inequality for d_{MI}, since the right-hand side might exceed $d_{MI}(\boldsymbol{p},\boldsymbol{q}) + d_{MI}(\boldsymbol{q},\boldsymbol{s})$. However, the right-hand size of (30) is often below $d_{MI}(\boldsymbol{p},\boldsymbol{q}) + d_{MI}(\boldsymbol{q},\boldsymbol{s})$ when c is close to 1. Since some of the estimates that lead to (30) are conservative, we believe the triangle inequality holds much more generally than (29).

In order verify asymmetry of d_{MI} we use the same example as in Appendix A, with $K = 2$, $\boldsymbol{p} = (1/3, 2/3)$ and $\boldsymbol{q} = (1/2, 1/2)$. It can be shown that

$$d_{MI}(\boldsymbol{p},\boldsymbol{q}) = \rho\left(\frac{2}{3}\right) - \frac{2}{3}\rho\left(\frac{1}{2}\right) = 0.252.$$

whereas

$$d_{MI}(\boldsymbol{q},\boldsymbol{p}) = \rho\left(\frac{3}{4}\right) - \frac{1}{2}\rho\left(\frac{1}{2}\right) = 0.311,$$

Note in particular that the inequality between the two path lengths goes in the direction $d_{MI}(\boldsymbol{q},\boldsymbol{p}) > d_{MI}(\boldsymbol{p},\boldsymbol{q})$, whereas the opposite inequality $d_{FI}(\boldsymbol{q},\boldsymbol{p}) < d_{FI}(\boldsymbol{p},\boldsymbol{q})$ holds for the functional information distance. The reason is that whereas d_{FI} only takes the non-censored subsample into account, d_{MI} includes the censored as well as the non-censored subsamples. As mentioned in Section 5.4, we consider the latter mutual information approach less appealing, and this is due to the fact that $d_{MI}(\boldsymbol{q},\boldsymbol{p})$ is larger than $d_{MI}(\boldsymbol{p},\boldsymbol{q})$, in spite of the fact that the path from $\boldsymbol{p}$ to $\boldsymbol{q}$ requires a larger increase of frequencies of categories.

Appendix E. Conditions under which the evolutionary waiting times equals the functional information distance

In order to verify (19), notice that the genetic composition $\boldsymbol{p}_t = (p_{t1}, \ldots, p_{tK})$ at time t has components

$$\lim_{L\to\infty} p_{tk} = \frac{p_k e^{-\lambda_k t}}{\sum_{j=1}^{K} p_j e^{-\lambda_j t}} \tag{31}$$

in the limit of large L. Suppose the components of the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ of death rates is chosen as functions $\lambda_k = -\log(r_k)/C$ of the non-censoring probabilities r_k in (6) and some constant $C > 0$ (to be defined below). It follows from (6) and (31) that

$$\lim_{\delta\to 0}\lim_{L\to\infty} T_{L,\delta}(\boldsymbol{p},\boldsymbol{q}) = C. \tag{32}$$

In particular, with

$$C = d_{FI}(\boldsymbol{p},\boldsymbol{q}) \tag{33}$$

the functional information distance between $\boldsymbol{p}$ and $\boldsymbol{q}$, the limiting evolutionary waiting time

$$\lim_{\delta\to 0}\lim_{L\to\infty} T_{L,\delta}(\boldsymbol{p},\boldsymbol{q}) = d_{FI}(\boldsymbol{p},\boldsymbol{q}) \tag{34}$$

agrees with (31). We will show that (33) provides a convenient normalization of the death intensities λ_k. To this end we introduce the function $f(x) = e^{-Cx}$ and notice that

$$\sum_{k=1}^{K} p_k f(\gamma_k) = \sum_{k=1}^{K} p_k r_K = 2^{-C} = e^{-C/\log_2(e)} = f\left(\frac{1}{\log_2(e)}\right). \tag{35}$$

Equation (35) states that $1/\log_2(e)$ is a type of weighted average of the elements $\lambda_1, \ldots, \lambda_K$ of $\boldsymbol{\lambda}$.

References

Akashi, H., Gojobori, T., 2002. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. PNAS 99 (6), 3695–3700. https://doi.org/10.1073/pnas.062526999.

Basener, W., Cordova, S., Hössjer, O., Sanford, J., 2021. Dynamical systems and fitness maximization in evolutionary biology. In: Sriraman, B. (Ed.), Handbook of the Mathematics of the Arts and Sciences. Springer, Cham. https://doi.org/10.1007/978-3-319-70658-0_121-1.

Barbour, A., Ethier, S.N., Griffiths, R.C., 2000. A transition function expansion for a diffusion model with selection. Ann. Appl. Probab. 10 (1), 123–162. https://doi.org/10.1214/aoap/1019737667.

Behrens, S., Vingron, M., 2010. Studying evolution of promoter sequences: a waiting time problem. J. Comput. Biol. 17 (12), 1591–1606. https://doi.org/10.1089/cmb.2010.0084.

Bernoulli, J., 1713. Ars conjectandi. Thurneysen Brothers. Basel.

Bhaskaran, R., Ponnuswamy, P.K., 1988. Amino acid scale: Average flexibility index. Int. J. Pept. Protein. Res. 32, 242–255. https://doi.org/10.1111/j.1399-3011.1988.tb01258.x.

Burgin, M., 2010. Theory of Information: Fundamentality, Diversity and Unification. World Scientific, Singapore.

Cobzas, S., 2013. Functional Analysis in Asymmetric Normed Space. Springer. https://doi.org/10.1007/978-3-0348-0478-3.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory, second ed. Wiley.

Crow, J.F., Kimura, M., 1970. An Introduction to Population Genetics Theory. The Blackburn Press, Cadwell, New Jersey.

Dembski, W.A., Marks II, R.J., 2009a. Bernoulli's principle of insufficient reason and conservation of information in computer search. In: Proc. Of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2647–2652. https://doi.org/10.1109/ICSMC.2009.5346119. San Antonio, TX.

Dembski, W.A., Marks II, R.J., 2009b. Conservation of information in search: measuring the cost of success. IEEE Transactions on Systems, Man and Cybernetics A, Systems & Humans 5 (5), 1051–1061. https://doi.org/10.1109/TSMCA.2009.2025027.

Díaz-Pachón, D.A., Marks, R.J., 2020. Active information requirements for fixation on the Wright-Fisher model of population genetics. Biocomplexity (4), 1–6. https://doi.org/10.5048/BIO-C.2020.4.

Díaz-Pachón, D.A., Hössjer, O., 2022. Assessing, testing and estimating the amount of fine-tuning by means of active information. Entropy 24, 1323. https://doi.org/10.3390/e24101323.

Durrett, R., 2008. Probability Models for DNA Sequence Evolution. Springer, New York.

Durrett, R., Schmidt, D., 2008. Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. Genetics 180, 1501–1509. https://doi.org/10.1534/genetics.107.082610.

Durrett, R., Schmidt, D., Schweinsberg, J., 2009. A waiting time problem arising from the study of multi-stage carinogenesis. Ann. Appl. Probab. 19 (2), 676–718. https://doi.org/10.1214/08-AAP559.

Durston, K.K., Chiu, D.K.Y., Abel, D.L., Trevors, J.T., 2007. Measuring the functional sequence complexity of proteins. Theor. Biol. Med. Model. 4, 47. https://doi.org/10.1186/1742-4682-4-47.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman and Hall, New York.

Ewens, W.J., 2004. Mathematical population genetics I. In: Theoretical Introduction. Springer, New York.

Fano, R.M., 1961. Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, MA.

Godfrey-Smith, P., Sterelny, K., 2016. Biological Information. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/information-biological/. (Accessed 21 August 2021).

Gromiha, M.M., Oobatake, M., Sarai, A., 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys. Chem. 82 (1), 51–67. https://doi.org/10.1016/S0301-4622(99)00103-9.

Griffiths, P.E., 2017. Genetic, epigenetic and exogenetic information in development and evolution. Interface Focus 7 (5), 20160152. https://doi.org/10.1098/rsfs.2016.0152.

Hazen, R.M., Griffin, P.L., Carothers, J.M., et al., 2007. Functional information and the emergence of biocomplexity. Proceedings of the National Academy of Sciences of the USA 104 (Suppl. 1), 8574–8581. https://doi.org/10.1073/pnas.0701744104.

Hössjer, O., Bechly, G., Gauger, A., 2021. On the waiting time until coordinated mutations get fixed in regulatory sequences. J. Theor. Biol. 524, 110657 https://doi.org/10.1016/j.jtbi.2021.110657.

Iqbal, M., Verrall, R.E., 1988. Implications of protein folding. Additivity schemes for volumes and compressibilities. J. Biol. Chem. 263 (9), 4159–4165.

Jaynes, T., 2003. Probability Theory: the Logic of Science. Cambridge University Press, Cambridge.

Khamsi, M.A., 2015. Generalized metric spaces: a survey. J. Fixed Point Theory Appl. 17, 455–475. https://doi.org/10.1007/s11784-015-0232-5.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86. https://doi.org/10.1214/aoms/1177729694.

Lewontin, R., 2003. Four Complications in Understanding the Evolutionary Process. Institute Bulletin (Santa Fe Institute), Santa Fe. https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/publication/2016/10/31/winter2003v18n1.pdf.

Logan, R.K., 2012. What is information?: why is it relativistic and what is its relationship to materiality, meaning and organization. Information 3 (1), 68–91. https://doi.org/10.3390/info3010068.

Oobatake, M., Ooi, T., 1993. Hydration and heat stability effects on protein unfolding. Prog Biophys Mol Biol 59, 237–284. https://doi.org/10.3390/info3010068.

Rokhlin, V.A., 1967. Lectures on the entropy theory of measure preserving transformations. Russ. Math. Surv. 22, 1–52. https://rene.ruhr/files/Rokhlin1967_Entropy.pdf.

Sanford, J., Brewer, W., Smith, F., Baumgardner, J., 2015. The waiting time problem in a model hominin population. Theor. Biol. Med. Model. 12, 18. https://doi.org/10.1186/s12976-015-0016-z.

Schneider, T.D., 2006. Claude Shannon: biologist. The founder of information theory used biology to formulate the channel capacity. IEEE Eng. Med. Biol. Mag. 25 (1), 30–33. https://doi.org/10.1109/MEMB.2006.1578661.

Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18 (20), 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 623–656, 379-423.

Shiga, T., 1981. Diffusion processes in population genetics. J. Math. Kyoto Univ. 21, 133–151.

Srivastava, P., Khare, M., 1999. Conditional entropy and Rokhlin metric. Math. Slovaca 49 (4), 433–441. http://dml.cz/dmlcz/133065.

Stojmirovic, A., 2005. Quasi-metrics: Similarities and Searches: Aspects of Geometry and Protein Datasets. Victoria University of Wellington, New Zeeland. https://doi.org/10.48550/arXiv.0810.5407. PhD Thesis.

Szostak, J., 2003. Functional information: molecular messages. Nature 423, 689. https://doi.org/10.1038/423689a.

Theune, U., 2023. Ternary plots. In: MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/7210-ternary-plots. (Accessed 6 November 2023).

Thorvaldsen, S., 2016. A mutation model from first principles of the genetic code. IEEE ACM Trans. Comput. Biol. Bioinf 13, 878–886. https://doi.org/10.1109/TCBB.2015.2489641.

Thorvaldsen, S., Flå, T., Willassen, N.P., 2010. Deltaprot: a software toolbox for comparative genomics. BMC Bioinf. 11 (1), 573. https://doi.org/10.1186/1471-2105-11-573.

Thorvaldsen, S., Hössjer, O., 2023. Estimating the information content of genetic sequence data. J. Roy. Stat. Soc. C Appl. Stat. 72 (5), 1310–1338. https://doi.org/10.1093/jrsssc/qlad062.

Thorvaldsen, S., Øhrstrøm, P., Hössjer, O., 2024. The representation, quantification, and nature of genetic information. Synthese 204 (15). https://doi.org/10.1007/s11229-024-04613-z.

Walker, S.I., Davies, P.C.W., 2013. The algorithmic origins of life. J. R. Soc. Interface 10 (79), 20120869. https://doi.org/10.1098/rsif.2012.0869.

Wells, M.T., Casella, G., Robert, C.P., 2004. Generalized Accept-Reject sampling schemes. Institute of Mathematical Statistics Lecture Notes. A Festschrift for Herman Rubin 45, 342–347.

Wilson, W.A., 1931. On quasi-metric spaces. Am. J. Math. 53 (3), 675–684. https://doi.org/10.2307/2371174.

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones, D.F. (Ed.), Proceedings of the Sixth International Congress of Genetics. Brooklyn Botanic Gardens, Brooklyn NY, pp. 356–366.

Wright, S., 1949. Adaptation and selection. In: Jepson, G.L., Mayr, E., Simpson, G.G. (Eds.), Genetics, Paleontology, and Evolution. Princeton Univ. Press, pp. 365–389.

Zimmerman, J.M., Eliezer, N., Simha, R., 1968. The characterization of amino acid sequences in proteins by statistical methods. J. Theor. Biol. 21 (2), 170–201. https://doi.org/10.1016/0022-5193(68)90069-6.