Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **00**, 1–29
https://doi.org/10.1093/jrsssc/qlad062

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

Estimating the information content of genetic sequence data

Steinar Thorvaldsen[1] **and Ola Hössjer**[2]

[1]Division of Science, Department of Education, UiT the Arctic University of Norway, Tromsø, Norway
[2]Division of Mathematical Statistics, Department of Mathematics, Stockholm University, Stockholm, Sweden

Address for correspondence: Steinar Thorvaldsen, Division of Science, Department of Education, UiT the Arctic University of Norway, 9037 Tromsø, Norway. Email: steinar.thorvaldsen@uit.no

Abstract

A prominent problem in analysing genetic information has been a lack of mathematical frameworks for doing so. This article offers some new statistical methods to model and analyse information content in proteins, protein families, and their sequences. We discuss how to understand the qualitative aspects of genetic information, how to estimate the quantitative aspects of it, and implement a statistical model where the qualitative genetic function is represented jointly with its probabilistic metric of self-information. The functional information of protein families in the *Cath* and *Pfam* databases are estimated using a method inspired by rejection sampling. Scientific work may place these components of information as one of the fundamental aspects of molecular biology.

Keywords: functional information, mutual information, rejection sampling, self-information

1 Introduction

All known life on our planet is based on genetic sequences stored in DNA, and today a flood of sequence data is available in the nucleotide and amino acid databases. Scientific progress depends crucially on statistical methods and computer algorithms that allow extracting useful information from the sequence data. Mathematics that fits these requirements was created by Claude Shannon with the introduction of information theory (Shannon, 1948), and his theory has been successfully applied to quantify and analyse nucleotide and amino acid sequences (Schneider, 2006; Schneider & Stephens, 1990). As an interdisciplinary domain of study, bioinformatics has unlocked the field of molecular biology through the use of computer science and statistics.

In this way, information has become a central idea of contemporary biology, and there is a common understanding that the informational aspect of life is a key property—maybe the master key property (Godfrey-Smith & Sterelny, 2016; Walker & Davies, 2013). Griffiths (2017) has argued that it is common sense to consider the characteristics of genes and chromosomes as the expression and transmission of information, and he emphasises the current challenge of capitalising on this in strict, scientific terms.

Information theory is founded on probability theory, to the extent that the axiomatics of both theories are formally equivalent (Jizba & Korbel, 2020). This understanding leads to a useful interpretation of the information-theoretic quantities. Even though Shannon's theory was framed plainly to address the problem of communication, modern approaches interpret information quantities as measures of belief-updating in statistical inference, and hence as proper tools to study many complex systems (Mediano et al., 2022).

Received: September 9, 2022. **Revised:** May 31, 2023. **Accepted:** June 23, 2023

Biology has approved the concept of information but has, to a certain extent, avoided the concept of meaning. In qualitative terms, *genetic function* is possibly the best depiction of the 'meaning' of a sequence. This biochemical function is determined by direct empirical experimentation and links information content to functionality (Adami & Nitash, 2022). De Mul (2021) distinguishes between the syntactic aspect of biological sequences and structures, and their semantic aspect. Shannon information is only about syntax, whereas it is life itself that gives meaning to sequences. Functional information is based on the probability that an arbitrary configuration of a system (letter sequence or protein) will obtain a specific function to a specified degree (Hazen et al., 2007). It can be considered as a bridge between Shannon information, which only includes probabilities but not function, and the concept of semantic information, which is very wide and therefore is difficult to quantify in terms of probabilities, although it includes function as a possibility.

It is not easy to capture the general notion of information in a simple definition. We may define information in the broad sense as 'all that which is communicated', and hence, the information within a living cell is much greater than its genetic sequences. There are many heritable structures other than DNA that carry information, such as post-translational protein modifications (phosphorylation, glycosylation, and lipidation) and epigenetic processes (chromatin modifications). All parts of the cell, including the DNA, RNA, and protein molecules, are in steady communication with each other, and thousands of different types of interactions take place. We must acknowledge here that there exist more levels of biological organisation than we address in the present article, and we will only study the presumably most fundamental one. However, extracting the information stored in genetic sequences is a crucial step towards a more comprehensive investigation to detect and decode more facets of biological information.

The study of information in linear genetic sequences is an ongoing topic of research (Koonin, 2016; Popa, Oldenburg & Ebenhöh, 2020). Recently, Adami & Nitash (2022) published a paper based on multivariate correlations. They present a simulation study based on short symbolic sequences, but they intend to extend it to biological data sets with considerably longer sequences.

Despite many years of research, the study of information has suffered from a lack of good framework that could be used to advance theories and guide discussions. There is still a great deal of open conceptual space and much room for new accounts of genetic information. Analysing life's informational properties holds the potential for turning biology into a more quantitative science (Davies & Walker, 2016). Our starting point for such analyses is a *gene family*, that is, a group of closely related genes that encode similar products, usually proteins, but also RNA. In this article, we present, extend, and elaborate on the statistical estimation of the information content in protein sequences, which is quantifiable and amenable to computational assessment. In particular, we develop a method inspired by rejection sampling (Wells et al., 2004) in order to estimate the functional information of a gene family. We will provide an accessible introduction to the framework, examining the merits of the approach in various sequence data of interest. We also discuss some interpretation concerns, while highlighting the benefits of the formalism. The core ideas of the article are briefly explained in Figure 1.

2 Defining and estimating genetic information

In this section, we discuss how to define and estimate genetic information with some of the most important notation, as summarised in Table 1. Genetic information cannot easily be quantified and estimated in a direct and general manner, and the same is true for genetic meaning. However, genetics (as well as many other sciences) makes use of data, based on which various measures of information can be defined and estimated. Such measures are central to how we learn about the properties of subjects and operationalise our theories.

Measures of information involve presumptions or specifications regarding signs, observers, and reference states that require careful consideration of the basic aspects of the system. According to long-term practices, natural entities are divided into qualities and quantities. These are among the fundamental categories of philosophy: Quantities can be measured and are objective, whereas qualities are typically subjective and cannot be measured. However, in the case of genetic information, the type of qualities that appear in this context are typically not subjective but *objective* features of the world because they are the same for all observers (Barbieri, 2016). In addition to quantities (objective and measurable) and qualities (subjective and nonmeasurable), we must

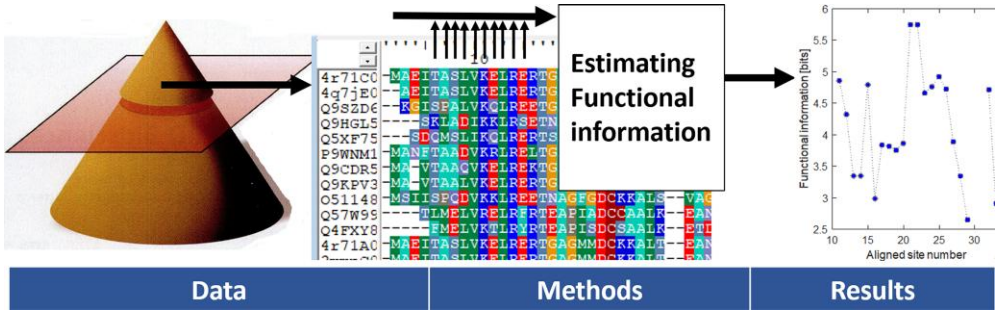

Figure 1. Overview of the core ideas of the article. Data: Imagine a conic stack of protein molecules of all possible finite sequences sorted by a specific activity with the most active at the top (Hazen et al., 2007; Szostak, 2003). A horizontal plane across the stack signifies a given level of activity and defines data—the set of functional sequences. Such aligned sequences may be downloaded from the databases. Methods: In Section 2, we describe various methods for how such sequence alignments may be analysed in different ways, with site specific (vertical) sequences having a crucial role. Results: Section 3 provides an information profile (functional information in bits per site) of aligned amino acid sequences, which highlights sites of particular functional importance in terms of the statistical information that they convey.

therefore recognise the existence of a third type of nominable entity (objective but not directly measurable), presented as *Function* in Table 2. It is sometimes possible though to measure function indirectly, declaring an entity as functional if it satisfies certain measurable properties (such as the outcome of a test for function). When this is the case function serves as a bridge between the concepts of quantities and qualities. On top of such function, sequences have additional (and not easily measurable) properties (meaning) to the linear order of their units, as shown in the table.

The most common representation of information is a linear sequence of symbols. A protein sequence of length L is studied as a discrete and vector-valued random variable $\boldsymbol{X} = (X_1, \ldots, X_L)$, where X_j, $j = 1, \ldots, L$ is the amino acid of site j. Information in proteins has been measured in different ways. As mentioned previously, Shannon's information theory is based on probability theory. In a sequence of characters, the classical Shannon measure of information is a function of the probabilities of the character string. Yockey (1977) initiated the application of information theory to protein sequences by estimating the variability of amino acids at each position in the primary sequence of the cytochrome *c* protein family. He determined the information per amino acid to be 2.953 bits.

2.1 Self-information

What is commonly referred to as *self-information* can be applied to protein sequences. This is a measure of information content or 'surprisal' of what we usually call a letter x_j,

$$I(x_j) \stackrel{\text{def}}{=} \log_2 \frac{1}{p_{x_j}} = -\log_2 (p_{x_j}),$$

where p_{x_j} is the probability of each character, x_j. The use of logs to measure information dates back to Nyquist (1924) and Hartley (1928). By definition, the measure of self-information for an entire sequence of letters is positive and additive if the components of $\boldsymbol{X}$ are independent. This definition states that the surprisal of a symbol corresponds to the amount of self-information carried by that symbol, and the latter has been recognised as an important quantity for the study of information (Dretske, 1981). According to Dretske, self-information reflects the fundamental intuition behind information (Dretske, 1981, p. 529). This measure of heterogeneity, or the intrinsic complexity of data, is denoted in the theory as its self-information. Similar to Shannon entropy, the self-information model only considers the statistical properties of the symbols that form messages. The unit of the self-information is 'bit' because the base of the logarithm in the formula is two.

This probabilistic measure of information does *not* consider the meaning or function of the message. The meaning-free concept of information theory is insufficient to explain the important

Table 1. Summary of notation for genetic data and information. Scalars are denoted as light characters, vectors as bold characters, and matrices as underscored bold characters

Quantity	Description
x	Amino acid ($\in\{1, \ldots, 20\}$)
L	Length of amino acid sequence
j	Site of amino acid sequence ($\in\{1, \ldots, L\}$)
$\boldsymbol{X}$	Amino acid sequence ($=(X_1, \ldots, X_L)$)
$\boldsymbol{p}$	Vector of amino acid probabilities ($=(p_1, \ldots, p_{20})$), corr. to prior distribution of amino acids
$\underline{\boldsymbol{X}}_R$	Reservoir of M_R amino acid sequences of length L ($=(X_{mj}; m=1, \ldots, M_R, j=1, \ldots, L)$)
M	Number of amino acid sequences of a protein family ($M \ll M_R$)
$\underline{\boldsymbol{X}}_f$	Protein family of amino acid sequences of length L ($=(X_{mj}; m=1, \ldots, M, j=1, \ldots, L)$)
$\underline{\boldsymbol{X}}_F$	Family $\underline{\boldsymbol{X}}_f$ of amino acids corresponding to functioning proteins
$\boldsymbol{X}_m$	Amino acid sequence, corresponding to protein m of a protein family ($=(X_{mj}; j=1, \ldots, L)$)
$\boldsymbol{q}_j$	Distribution of amino acids at site j of a protein family ($=(q_{1,j}, \ldots, q_{20,j})$)
$I^{(p)}$	Information content of an amino acid sequence ($=I^{(p)}(\boldsymbol{X})$) or protein family ($=I^{(p)}(\underline{\boldsymbol{X}}_f)$)
$\overline{I^{(p)}}$	Information content per site of an aa sequence ($=\overline{I^{(p)}(\boldsymbol{X})}$) or protein family ($=\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)}$)
$\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)$	Vector of information contents of the amino acids of $\underline{\boldsymbol{X}}_f$ ($=(I^{(p)}(\boldsymbol{X}_m); m=1, \ldots, M)$)

aspects of biology (Jablonka, 2002). It only focuses on some highly relevant statistical characteristics, whereas notions of functional information (defined in Section 2.4) try to include the dimension of functionality.

A protein sequence may be considered as a linear string of symbols produced, consistent with a set of probabilities governing their rate of occurrence. With a Bayesian statistical perspective (Berger, 1993), these are *prior probabilities* that can be deduced directly from the genetic code, as shown in Table 3. This amino acid distribution assigns the same probability to all codons, which is a natural starting point from 'first principles' thinking (cf. Aristotle). It is an instance of the principle of insufficient reason, also referred to as the principle of indifference, applied to the set of 61 nonstop codons.

2.2 Gene sequences

When applying self-information to estimate the information contained in a gene, the possible elements (amino acids) are finite and known, and their probabilities can be computed. This also allows the definition of a standard reference state (baseline, null state) for the estimation of information in genetic systems. As mentioned above, the discrete probability distribution, outlined in the third column of Table 3, is referred to as the baseline *prior distribution* of amino acids based on the assumption that all nonstop codons are equally likely. These probabilities can be put into a probability vector

$$\boldsymbol{p} = (p_1, \ldots, p_{20})$$

that represents a discrete probability distribution, with the 20 entries adding up to 1. Similarly, we define the *self-information vector* of the prior distribution $\boldsymbol{p}$ as

$$\boldsymbol{I}^{(p)} = (I_1^{(p)}, \ldots, I_{20}^{(p)}) = (-\log_2(p_1), \ldots, -\log_2(p_{20})),$$

and it is shown in the fourth column of Table 3. A vector $\boldsymbol{h}^{(p)}$ of Shannon uncertainty values for all amino acids, under the a priori distribution $\boldsymbol{p}$, may likewise be computed as follows:

$$\boldsymbol{h}^{(p)} = (-p_1\log_2(p_1), \ldots, -p_{20}\log_2(p_{20})),$$

Table 2. Five distinct characteristics of protein sequences $\boldsymbol{X}$ (Barbieri, 2016) and their scale of statistical measurement

Property of genes	Scientific framework	Statistical measure level
Probability	Self-information: $I(\boldsymbol{X})$	Metric scale (bits)
Complexity	Algorithmic complexity: length	Metric scale (bits)
Distance	Relative distance: D	Metric scale (bits)
Organic information	Function F in the context of a cell: joint variable $[\underline{\boldsymbol{X}}_R, F]$, functional information	Joint [scale (bits), nominal]
Organic meaning	Cellular and intracellular networks, logistics in a living system	Joint nominals (categories)

Note. A reservoir of such sequence is gathered into a matrix $\underline{\boldsymbol{X}}_R$. Information and meaning are considered non-numerical entities but are objective observables in genetics (due to the concept of function) and hence fundamental nominal data types. The self-information is defined in Section 2.1. Algorithmic complexity was introduced by Kolmogorov (1965).

cf. the rightmost column of Table 3 (by continuity, we define $0 \cdot \log_2(0) = 0$). In particular, the Shannon entropy $H^{(p)}$ is the sum of the elements of $\boldsymbol{h}^{(p)}$, or equivalently, the scalar product $\cdot$ of the a priori distribution and self-information vectors:

$$H^{(p)} = \boldsymbol{p} \cdot \boldsymbol{I}^{(p)} = -\sum_{x=1}^{20} p_x \log_2 p_x = 4.139 \text{ bits},$$

somewhat lower than the theoretical maximal entropy of $-\log_2(1/20) = 4.322$ bits, obtained from a uniform prior distribution with probability 1/20 for all amino acids.

Now, consider a second vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{20})$ of amino acid probabilities, whose entries correspond to the frequencies by which the amino acids occur along the amino acid sequences. We regard $\boldsymbol{p}$ as a prior distribution of the null state of a baseline generation and $\boldsymbol{\nu}$ as the distribution of amino acid frequencies of the generation at which the amino acid sequence is sampled. Motivated by problems of random search algorithms, Dembski & Marks (2009a, b) introduced the concept of *active information* which was later applied to population genetics by Díaz-Pachón & Marks (2020). In our context, a change in the frequency of an amino acid x from p_x to ν_x corresponds to active information

$$I_x^+ = \log_2 \frac{\nu_x}{p_x} = I_x^{(p)} - I_x^{(\nu)}.$$

Assume that some exogenous information changed the amino acid frequencies from $\boldsymbol{p}$ to $\boldsymbol{\nu}$. Then, this information was either helpful or detrimental for amino acid x depending on whether $I_x^+ > 0$ or $I_x^+ < 0$, respectively. The *expected active information* per amino acid site for a sequence $\boldsymbol{X} = (X_1, \ldots, X_L)$, whose components are independent of distribution $\boldsymbol{\nu}$, is determined by

$$E(I_{X_i}^+) = \sum_{x=1}^{20} \nu_x I_x^+ = \sum_{x=1}^{20} \nu_x I_x^{(p)} - \sum_{x=1}^{20} \nu_x I_x^{(\nu)} = E_\nu^{(p)} - E_\nu^{(\nu)}.$$

In the last step of the above equation, we introduced the expected self-information (or the cross-entropy of $\boldsymbol{p}$ relative to $\boldsymbol{\nu}$)

$$E_\nu^{(p)} = \boldsymbol{\nu} \cdot \boldsymbol{I}^{(p)} = -\sum_{x=1}^{20} \nu_x \log_2 p_x$$

along a randomly chosen amino acid sequence with amino acid frequencies $\boldsymbol{\nu}$, assuming that these frequencies were obtained from $\boldsymbol{p}$. For instance, $E_p^{(p)} = H^{(p)} = 4.139$ bits if amino acids occur along

Table 3. The prior probabilities and self-information of each amino acid deduced from the codon statistics of the genetic code

Amino acid x	No. of codons coding for x	Probability p_x	Self-information $I(x) = -\log_2(p_x)$ (bits)	Expected surprisal (Shannon uncertainty) $h(x) = -p_x \log_2(p_x)$ (bits/amino acid)
M W	1	1/61	5.93	0.097
N D C Q E H K F Y	2	2/61	4.93	0.162
I	3	3/61	4.35	0.214
A G P T V	4	4/61	3.93	0.258
–	5	–	–	–
R L S	6	6/61	3.35	0.329

Note. A DNA triplet may code for 64 different codons, and 3 of these code for STOP; hence, we divide by 61 to determine the prior probabilities p_x for all amino acids, $x = 1, \ldots, 20$, assuming that all nonstop codons are equally likely a priori. The table makes it possible to attribute an operational measure to the self-information and Shannon uncertainty (entropy) of an individual source letter.

the sequence according to the prior probabilities ($\boldsymbol{\nu} = \boldsymbol{p}$) of Table 3, whereas $E_\nu^{(p)} = 4.514$ bits and $E_\nu^{(\nu)} = H^{(\nu)} = 4.322$ bits if all 20 amino acids are equally likely to occur ($\nu_x = 1/20$). Note also that the expected active information equals the Kullback–Leibler divergence

$$D_{KL}(\boldsymbol{\nu} \,\|\, \boldsymbol{p}) = \sum_{x=1}^{20} \nu_x \log_2 \frac{\nu_x}{p_x} = E_\nu^{(p)} - H^{(\nu)}$$

from $\boldsymbol{\nu}$ to $\boldsymbol{p}$ (Kullback & Leibler, 1951).

Based on the basic considerations above, we can now study any observed sequence $\boldsymbol{X} = (X_1, \ldots, X_L)$ of amino acids with length L, where X_j is the amino acid of site j. It is convenient to introduce the composition vector $\boldsymbol{N}$, according to the occurrences of each of the amino acids in $\boldsymbol{X}$:

$$\boldsymbol{N} = (n_1, \ldots, n_{20}),$$

where n_x is the total number of occurrences of amino acid x in $\boldsymbol{X}$, that is, the number of sites j for which $X_j = x$. This is used to derive a measure of the *information content* based on the self-information of the amino acid in a sequence of length L:

$$I^{(p)}(\boldsymbol{X}) \stackrel{\text{def}}{=} \log_2 \prod_{j=1}^{L} \frac{1}{p_{X_j}} = \sum_{j=1}^{L} \log_2 \frac{1}{p_{X_j}} = -\sum_{j=1}^{L} \log_2 p_{X_j} = -\sum_{x=1}^{20} n_x \log_2 p_x = \boldsymbol{N} \cdot \boldsymbol{I}^{(p)}.$$

We may also compute the *mean* self-information per amino acid of the protein sequence as

$$\overline{I^{(p)}(\boldsymbol{X})} = \frac{1}{L}(\boldsymbol{N} \cdot \boldsymbol{I}^{(p)}).$$

Let us assume that the components of $\boldsymbol{X}$ are independent and identically distributed with marginal distribution $\boldsymbol{\nu}$. The mean self-information can be viewed as a consistent estimator of $E_\nu^{(p)}$ as L tends to infinity. Indeed, by the law of large numbers, the relative frequencies of the amino acids along $\boldsymbol{X}$ converge to the probabilities in $\boldsymbol{\nu}$ as L increases ($n_x/L \approx \nu_x$ for $x = 1, \ldots, 20$), and this implies that the self-information and mean self-information satisfy $I^{(p)}(\boldsymbol{X}) \approx L E_\nu^{(p)}$ and $\overline{I^{(p)}(\boldsymbol{X})} \approx E_\nu^{(p)}$, respectively for large L. When $\boldsymbol{\nu} = \boldsymbol{p}$, the expected self-information $E_p^{(p)} = H^{(p)}$ equals the entropy; the average amount of information per letter generated by the source, and moreover, the mean self-information $\overline{I^{(p)}(\boldsymbol{X})}$ is a consistent estimator of $H^{(p)}$, since $n_x/L \approx p_x$ for $x = 1, \ldots, 20$ when L is

large. In this way, Shannon's theory can deal with the information associated with amino acid sequences, consistent with traditional formalism.

By analogy with Shannon's model of communication, assuming that $\boldsymbol{p}$ represents the initial conditions of codon probabilities in the prior generation, $E_p^{(p)} = H^{(p)}$ can be considered as potential information of the baseline. Then, changes in 'transmitted' forms that occur due to differential replication and elimination processes express information about mutations, natural selection, and random genetic drift. Assuming that amino acid sequence frequencies change from $\boldsymbol{p}$ to $\boldsymbol{v}$ during a certain period of time, there is an associated increase or decrease in the information of variant forms from $H^{(p)}$ to $H^{(v)}$. In theory, one should be able to quantify this change of information for a given population of organisms for a given number of generations and estimate the amount of self-information and genetic variation gained or lost during the evolution of that lineage (cf. Díaz-Pachón & Marks, 2020).

When $\boldsymbol{X}$ is a string that represents the amino acid sequence of the protein that a gene translates to, $I^{(p)}(\boldsymbol{X})$ is the self-information of that gene. For protein families, the self-information can be computed for each sequence of the family. If we let $\underline{\boldsymbol{X}}_f = \{\boldsymbol{X}_m\}$ be the *set f* of all (horizontal) amino acid sequences $\boldsymbol{X}_m$ for the proteins m of a given organism, we obtain the self-information of the proteome.

However, we may also let the vector $\boldsymbol{X}_f$ represent the amino acids at a single *site* (i.e. *vertical*) along a multiple sequence alignment of a protein family. In the next subsection, we will consider matrices $\underline{\boldsymbol{X}}_f$ that not only are interpreted as a set of (horizontal) amino acid sequences but also as collection (vertical) alignments at several sites.

2.3 Multiple sequence alignments

An orthologue protein family f is commonly represented by the *alignment* of its sequences. There are several methods to model and study such alignments. Let L be the length of the alignment of M sequences. This can be represented as a matrix $\underline{\boldsymbol{X}}_f = (X_{mj})$ with M rows and L columns, where X_{mj} refers to a gap or the amino acid of protein m at site j. Whereas only amino acids are used to define basic statistics of $\underline{\boldsymbol{X}}_f$, later on in the text when single summarising measures of information for $\underline{\boldsymbol{X}}_f$ are presented, gaps will also be addressed.

2.3.1 Sample statistics

First, we consider the basic statistics of a multiple sequence alignment. Let L_m, $m = 1, \ldots, M$, be the length of the amino acids in each sequence in the alignment. It is assumed that these sequences exclude gaps so that $L_m \leq L$. The self-information for each sequence in alignment $\underline{\boldsymbol{X}}_f$ may be computed in a straightforward manner, as described in Section 2.2, to obtain the mathematical range, mean, and standard deviation of the information content of the sequences in the alignment (gaps are ignored). To this end, it is convenient to introduce a composition vector $\mathbf{N}_m$ for each protein $m = 1, \ldots, M$, and $\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)$, a vector of length M, whose component m contains the mean self-information content $\overline{I^{(p)}(X_{m1}, \ldots, X_{mL_m})} = \overline{I^{(p)}(\boldsymbol{X}_m)} = \mathbf{N}_{\boldsymbol{m}} \cdot \boldsymbol{I}^{(p)}/L_m$ of all amino acid sequences, that is, of all rows of $\underline{\boldsymbol{X}}_f$. This yields the following statistics computed from the components of $\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)$:

$$\text{range}\,(\underline{\boldsymbol{X}}_f) = [\min(\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)), \quad \max(\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f))],$$

$$\overline{\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{L_m}(\mathbf{N}_{\boldsymbol{m}} \cdot \boldsymbol{I}^{(p)}),$$

$$SD_{\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left(\frac{1}{L_m}(\mathbf{N}_{\boldsymbol{m}} \cdot \boldsymbol{I}^{(p)}) - \overline{\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)}\right)^2}.$$

2.3.2 The site-tolerant model

Whereas $\boldsymbol{I}^{(p)}(\underline{\boldsymbol{X}}_f)$ consists of the mean self-information contents of all M amino acid sequences of the protein family, it is also of interest to find a single information content measure $I^{(p)}(\boldsymbol{X}_f)$ of this

family that takes alignment and dependencies between the sequences into account. To this end, we introduce a binary matrix $\underline{\boldsymbol{b}} = (b_{xj})$, where $x = 1, 2, \ldots, 20, j = 1, 2, \ldots, L$, and b_{xj} equals 1 or 0 depending on whether amino acid x is present or not at site j, among the sequences with no gap at this site. As mentioned above, the alignment of M sequences can be represented as matrix $\underline{\boldsymbol{X}}_f = (X_{mj})$ with M rows and L columns, where X_{mj} refers to a gap or the amino acid of protein m at site j. Thus, $b_{xj} = 1$ if at least one of the amino acids of Column j in $\underline{\boldsymbol{X}}_f$ equals x. The additive laws of probabilities are applied to calculate the combined information for all sites in the alignment $\underline{\boldsymbol{X}}_f$:

$$I^{(p)}(\underline{\boldsymbol{X}}_f) = -\sum_{j=1}^{L} \log_2(\boldsymbol{p} \cdot \boldsymbol{b}_j) = -\log_2 \prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{b}_j,$$

where $\boldsymbol{b}_j = (b_{1j}, \ldots, b_{20,j})$ is a row vector of length 20, whose transpose $\boldsymbol{b}_j^T$ corresponds to Column j of matrix $\underline{\boldsymbol{b}}$. The gapped sites are considered to carry no information because amino acids are observed to be redundant at these sites in one or more of the sequences in the alignment. To incorporate gapped sites into the definition of $I^{(p)}(\underline{\boldsymbol{X}}_f)$, we set $\boldsymbol{b}_j = \boldsymbol{1}$ for these sites. Note in particular that $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{b}_j$ is the probability that a randomly chosen amino acid sequence of length L (under the assumed independence between sites and amino acid distribution $\boldsymbol{p}$ per site) at each site agrees with at least some members of the protein family at that site. Sites where all amino acids are represented (as well as gapped sites) carry no information because $b_{xj} = \boldsymbol{1}$ for all x and hence $\boldsymbol{p} \cdot \boldsymbol{b}_j = \boldsymbol{p} \cdot \boldsymbol{1} = 1$, where $\boldsymbol{1}$ is a row vector of length 20 with 1 at all positions. Entirely conserved sites ($b_{xj} = 1$ for some $x = X_j$ and $b_{xj} = 0$ for all $x \neq X_j$) contain the self-information $-\log_2 p_{X_j}$ of the conserved amino acid, inferring that the proteins have a common construct and that all of them maintain the amino acid $X_j = X_{mj}$ of this construct.

One assumption of this site-tolerant model is that all mutations that are tolerated individually at a site are also simultaneously acceptable. This is truly not the case, and the consequence is that $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{b}_j$ is considerably larger than the fraction of acceptable amino acid sequences. Equivalently, the above-computed information $I^{(p)}(\underline{\boldsymbol{X}}_f)$ is considerably lower than the real amount of information. Studies of prokaryotic genomes in GenBank have shown that 98% of sites cannot accept an amino acid substitution at any given moment, but a large majority of all sites may be permitted to alter when other compensatory changes occur (Povolotskaya & Kondrashov, 2010). A single amino acid substitution is often deleterious owing to its one-sided effect on protein structure, expression, or function, as native protein structures are only marginally stable with small values of Gibbs free energy of unfolding. The same study also showed that at least 90% of the sites in any prokaryotic protein can accept a substitution given the correct combination of amino acids at other sites. This illustrates a dependency between a protein family's amino acids at different loci, and it calls for a more restrictive model to define the information content $I^{(p)}(\underline{\boldsymbol{X}}_f)$ of a protein family.

2.3.3 Site-distribution model

With the goal of defining a more restrictive model, with a larger $I^{(p)}(\underline{\boldsymbol{X}}_f)$, we will extend the site-tolerant model of the previous section. In the extension, we will not only register which amino acids are present at various sites of the protein family $\underline{\boldsymbol{X}}_f$ but also take the amino acid frequencies into account. We define a probability matrix $\underline{\boldsymbol{q}} = (q_{xj})$, where $x = 1, 2, \ldots, 20$, $j = 1, 2, \ldots, L$, and the entries of q_{xj} correspond to the frequencies at which each amino acid x occurs at each site j along the alignment, among those of the M sequences of $\underline{\boldsymbol{X}}_f$ that have no gap at site j. Let $\underline{\boldsymbol{r}} = (r_{xj})$ be another matrix whose component at row x and Column j is

$$r_{xj} = \begin{cases} 1, & \text{at a site } j \text{ with gaps for all } M \text{ proteins,} \\ \dfrac{q_{xj}/p_x}{\max(q_{1j}/p_1, \ldots, q_{20,j}/p_{20})}, & \text{at a site } j \text{ with no gaps.} \end{cases} \quad (1)$$

A more general formula, which also come to sites j for which some but not all of the M amino acid sequences have gaps, is provided in Appendix D. We interpret (1) as an instance of rejection sampling (Wells et al., 2004), with proposal distribution $\boldsymbol{p}$ and a target distribution $\boldsymbol{q}_j = (q_{1j}, \ldots, q_{20,j})$ whose transpose $\boldsymbol{q}_j^T$ corresponds to column number j of $\underline{\boldsymbol{q}}$. This sampling procedure is repeated independently for all sites $j = 1, \ldots, L$. More specifically, we will hypothetically assume that the M sequences of the protein family have been obtained through a sampling procedure with censoring (or rejection). Amino acid sequences are generated independently between sequences and sites from a large reservoir $\underline{\boldsymbol{X}}_R$ of amino acids with distribution $\boldsymbol{p}$, and an amino acid x at site j is retained (not censored) with probability r_{xj}, independently between sites and sequences. A noncensored amino acid at site j is either visible (no gap) or not visible (a gap). This sampling procedure is continued until eventually a protein family $\underline{\boldsymbol{X}}_f$ with M sequences of L noncensored sites is obtained. When M is large, due to the definition of r_{xj} in (1), the noncensored amino acids at site j with no gaps occur in proportions $\boldsymbol{q}_j$, and at a site j with gaps only they occur in proportions $\boldsymbol{p}$ (although in the latter case these M sampled amino acids are only seen as M gaps). At a site j with some gaps the M noncensored amino acids at this site occur in proportions between $\boldsymbol{q}_j$ and $\boldsymbol{p}$ (see Appendix D). The probability of retaining a randomly picked amino acid at site j is $\boldsymbol{p} \cdot \boldsymbol{r}_j$, where $\boldsymbol{r}_j^T$ refers to column number j of $\underline{\boldsymbol{r}}$, whereas the probability of retaining all amino acids of a whole sequence of length L is $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j$. Consequently, only a fraction $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j$ of the sampled sequences from $\underline{\boldsymbol{X}}_R$ are retained after censoring. The smaller this number is, the more similar the sequences of the protein family are since censoring tends to eliminate differences between amino acids x at each site j and retain those x for which r_{xj} is large. The censoring mechanism also gives rise to a self-information

$$I^{(p)}(\underline{\boldsymbol{X}}_f) = -\sum_{j=1}^{L} \log_2(\boldsymbol{p} \cdot \boldsymbol{r}_j) = -\log_2 \prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j, \tag{2}$$

defined as the number of bits of information obtained from the noncensoring probability of a sequence from the protein family. We will interpret $I^{(p)}(\boldsymbol{X}_f)$ as the amount information infused when sampling from the reservoir, or the amount of information that the observed family of M noncensored amino acid sequences represent. In addition, it will be seen in Section 2.4 that (2) corresponds to functional information if the noncensored sequences are defined as functional, whereas the censored sequences are nonfunctional. Note also that $\boldsymbol{r}_j$ equals $\boldsymbol{b}_j$ in the special case when all amino acids at site j with nonzero frequencies ($q_{xj} > 0$) are retained with certainty ($r_{xj} = 1$). The gapped sites, or those with gaps above a threshold, are considered to carry no information ($\boldsymbol{p} \cdot \boldsymbol{r}_j = \boldsymbol{p} \cdot 1 = 1$), and the same is true for sites where the amino acid distribution is the same as the prior distribution ($\boldsymbol{q} = \boldsymbol{p}$). Entirely conserved sites ($q_{xj} = r_{xj} = 1$ for some x and $r_{yj} = 0$ for all other amino acids y) contain the self-information $-\log_2 p_{X_j}$ of the conserved amino acid of this site between 3.346 and 5.931 bits (Table 3).

For both the site-tolerant and site-distribution models, the mean value of self-information per site of sequence alignment is determined by

$$\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)} = \frac{1}{L} I^{(p)}(\underline{\boldsymbol{X}}_f),$$

where L is the total number of sites of the alignment. In particular, when the protein family consists of a single sequence of length L such that $M = 1$ and $\underline{\boldsymbol{X}}_f = \boldsymbol{X}$, then $I^{(p)}(\underline{\boldsymbol{X}}_f)$ and $\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)}$ reduce to $I^{(p)}(\boldsymbol{X})$ and $\overline{I^{(p)}(\boldsymbol{X})}$, respectively.

In the appendices, we will derive a number of mathematical properties of $I^{(p)}(\underline{\boldsymbol{X}}_f)$. In Appendix A, we verify that $I^{(p)}(\underline{\boldsymbol{X}}_f)$ is a measure of information that satisfies the triangle inequality. This is based on the observation that $I^{(p)}(\underline{\boldsymbol{X}}_f)$ can be viewed as a distance between the prior amino acid distribution $\underline{\boldsymbol{p}} = (\boldsymbol{p}^T, \ldots, \boldsymbol{p}^T) = (p_{xj} = p_x)$, and the observed amino acid distribution $\underline{\boldsymbol{q}} = (\boldsymbol{q}_1^T, \ldots, \boldsymbol{q}_L^T) = (q_{xj})$ of $\underline{\boldsymbol{X}}_f$. Although this distance is not a metric (it is not symmetric), it still satisfies the triangle inequality. In the mathematical literature such spaces are often named a

quasi-metric space (Khamsi, 2015), or a 'mountainous' space, since the effort of climbing up to the top of a mountain is not the same as descending back to the starting point.

In Appendix B, we notice that the above interpretation of $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j$ as a noncensoring probability is only exact in the limit when the number M of aligned sequences is large. For this reason, $\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)}$ will not be zero but positive (of the order $1/\sqrt{M}$), for a family of M sequences that are randomly generated from the prior distribution. It is however possible to derive a correction term of $\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)}$, so that randomly generated protein families will have an expected self-information of 0 per site if this correction term is subtracted from $\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)}$.

The site-distribution model may be generalised to account for dependencies between the amino acids at different loci. In Appendix C, we generalise $I^{(p)}(\underline{\boldsymbol{X}}_f)$ to protein families that allow for dependencies between sites, so that the amino acids of each protein are draw from a Markov chain.

2.4 Functional information

It is beneficial to characterise semantic information that is represented in biology as *functional information*, as it comes closer to expressing the intuitive sense of the word information than mere Shannon's reduced uncertainty or combinatorial uncertainty. There is no general agreement on the measurement of functional information, and this is an ongoing discussion. A measure of *functional information* is required to account for all possible sequences that could carry out an equivalent biochemical function, such as a protein's ability to react or bind with a specific molecule. For any observed phenotypic function F, we denote the set of amino acid sequences with this phenotype as $\underline{\boldsymbol{X}}_F$. To simplify, we focus on discrete phenotypes (e.g. reaction or not, function or not) rather than on quantitative phenotypes (e.g. the catalytic constant), thus assuming that all amino acid sequences with a specific phenotype F correspond to functioning proteins.

In 2003, Jack Szostak published a short paper in *Nature*, pointing out that the meaning or functionality of a message is vital in molecular biology (Szostak, 2003). Because classical information theory does not distinguish between functionality and nonfunctionality, Szostak introduced the need for a new measure of information, which he called *functional information*. Some years later, he and three other colleagues defined *functional information* in terms of a gene string as $-\log_2$ of the 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 a molecule 'with greater fitness than any given degree of function'; in other words,

$$I_F(\underline{\boldsymbol{X}}_F) = -\log_2 \frac{\#(\underline{\boldsymbol{X}}_F)}{\#(\underline{\boldsymbol{X}}_R)}, \tag{3}$$

where $\underline{\boldsymbol{X}}_F$ is a matrix whose rows are different sequences of length L that meet or exceed the required level of function within a cell, and $\#(\underline{\boldsymbol{X}}_F)$ is the number of such sequences, that is, the number of rows of this matrix. Similarly, $\underline{\boldsymbol{X}}_R$ is a matrix whose rows consist of a reservoir of all possible sequences, both functional and nonfunctional, of the same length L, whereas $\#(\underline{\boldsymbol{X}}_R)$ is the number of such sequences. Consequently, $\underline{\boldsymbol{X}}_F$ includes only a small fraction of the rows of $\underline{\boldsymbol{X}}_R$, and the ratio $\#(\underline{\boldsymbol{X}}_F)/\#(\underline{\boldsymbol{X}}_R)$ represents the probability of a functional sequence within the larger set of all possible sequences. In the present setting, this type of idea enables us to distinguish between functionally significant and random information to isolate the former in terms of functional measures. The justification for this is that the information that makes no difference contributes nothing of operative value. Therefore, we seek to isolate and quantify only the effective information content of the system.

Many genotype sequences usually form the same family of phenotypes and $\#(\underline{\boldsymbol{X}}_F)$ is an unknown number for most proteins. Then, unfortunately, we are unable to use the definition by itself to calculate the probability or the functional information required to code for a functional protein. Some restrictions must be introduced.

An *orthologue* protein *family* carrying a specific function is commonly represented by the alignment $\underline{\boldsymbol{X}}_f$ of its sequences. Regarding $\underline{\boldsymbol{X}}_F = \underline{\boldsymbol{X}}_f$ as our protein family, we may approximate the fraction $\#(\underline{\boldsymbol{X}}_F)/\#(\underline{\boldsymbol{X}}_R)$ of functional genomic sequences with the noncensoring probability $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j$ that was introduced in Section 2.3. The rationale for this approximation is our previous assumption that the sequences of the protein family are randomly generated (with amino acids chosen

independently between sites with distribution $\boldsymbol{p}$) and censored (with $\boldsymbol{r}_j$ containing the fraction of noncensored sequences for all amino acids at site j). In the present context, we may interpret noncensored as functional, so that $\prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j$ is the fraction of the randomly generated amino acid sequences that are functional. By taking the $-\log_2$ of both terms, we thus approximate the functional information by $I^{(p)}(\underline{\boldsymbol{X}}_f)$ in (2).

2.5 Mutual information

The common measure of *mutual information* from the information theory may also be applied to aligned sequences. Mutual information is calculated between two variables, and measures the *expected* reduction in Shannon uncertainty, H, for one variable $\underline{\boldsymbol{X}}$ given *a randomly chosen* value of the other variable $\underline{\boldsymbol{X}}_f$. Here, we will use a conditional version of mutual information which corresponds to the *observed* change in H for $\underline{\boldsymbol{X}}$ given an *observed* value of $\underline{\boldsymbol{X}}_f$. Let $\underline{\boldsymbol{X}}_f$ be an aligned family of proteins of length L, $\underline{\boldsymbol{X}}_R$ is the ground state, a reservoir of all possible amino acid sequences of the same length, whereas $\underline{\boldsymbol{X}}$ has the same dimensionality $M \times L$ as $\underline{\boldsymbol{X}}_f$, with a distribution that corresponds to randomly sampling M out of all M_R rows of $\underline{\boldsymbol{X}}_R$ with no censoring. The mutual information quantifies the average information in $\underline{\boldsymbol{X}}_f$ about $\underline{\boldsymbol{X}}$, when $\underline{\boldsymbol{X}}_f$ is randomly sampled from the reservoir (with censoring, and with noncensoring probabilities $\boldsymbol{r}$ as in (1)), whereas the conditional version of mutual information refers to the information that the observed $\underline{\boldsymbol{X}}_f$ (also obtained with censoring) provides about $\underline{\boldsymbol{X}}$. The conditional mutual information between these two variables can be stated more formally as:

$$I(\underline{\boldsymbol{X}}_f) = \Delta H(\underline{\boldsymbol{X}}; \underline{\boldsymbol{X}}_f) = H(\underline{\boldsymbol{X}}) - H(\underline{\boldsymbol{X}}|\underline{\boldsymbol{X}}_f),$$

where $H(\underline{\boldsymbol{X}})$ is the total entropy of the L columns of $\underline{\boldsymbol{X}}$, each one with components distributed according to the prior $\boldsymbol{p}$, whereas $H(\underline{\boldsymbol{X}}|\underline{\boldsymbol{X}}_f)$ is the total entropy of the L columns of $\underline{\boldsymbol{X}}$, given that each one of them conforms with the corresponding column of $\underline{\boldsymbol{X}}_f$, or equivalently, a sequence that is chosen from a posterior distribution of $\underline{\boldsymbol{X}}$ given data from $\underline{\boldsymbol{X}}_f$.

The difference at each site j, between the entropy of Column j of $\underline{\boldsymbol{X}}$ and the conditional entropy of Column j of $\underline{\boldsymbol{X}}$, given Column j of $\underline{\boldsymbol{X}}_f$, along the L alignments is of major importance. Durston et al. (2007) computed these two uncertainties using amino acid frequencies calculated at each aligned site. The sum of the contributions to the difference between the entropy and conditional entropy at each position of the alignment leads to $\Delta H(\underline{\boldsymbol{X}}; \underline{\boldsymbol{X}}_f)$. To calculate the site uncertainties, Durston et al. (2007) determined the proportion of each amino acid at each site in the dataset using d_{xj}/M, where d_{xj} is the total number of occurrences of a specific amino acid $x = 1, \ldots, 20$ at site j and M denotes the number of sequences in the alignment. Using $\tilde{p}_x = 1/20$ as prior for amino acid x at site j (rather than the prior probabilities p_x of Table 3), the corresponding posterior probability based on data from the protein family is d_{xj}/M. Thus, $\tilde{p}_x = 1/20$ and d_{xj}/M, $x = 1, \ldots, 20$, are the probabilities used to calculate the contribution of site j to $H(\underline{\boldsymbol{X}})$ and $H(\underline{\boldsymbol{X}}|\underline{\boldsymbol{X}}_f)$, respectively. Summing over all sites, we obtain the conditional mutual information

$$\begin{aligned} I(\underline{\boldsymbol{X}}_f) = \Delta H(\underline{\boldsymbol{X}};\ \underline{\boldsymbol{X}}_f) = H(\underline{\boldsymbol{X}}) - H(\underline{\boldsymbol{X}}|\underline{\boldsymbol{X}}_f) &= -\sum_{j=1}^{L}\sum_{x=1}^{20}\frac{1}{20} log_2 \frac{1}{20} + \sum_{j=1}^{L}\sum_{x=1}^{20}\frac{d_{xj}}{M}\log_2\frac{d_{xj}}{M} \\ = L \cdot \log_2 20 + \sum_{j=1}^{L}\sum_{x=1}^{20} q_{xj}\log_2 q_{xj} &= L \cdot \log_2 20 + \sum_{j=1}^{L} \boldsymbol{q}_j \cdot \log_2 \boldsymbol{q}_j \end{aligned} \tag{4}$$

of the aligned family, where in the fourth step, we used $q_{xj} = d_{xj}/M$. Because the authors assumed a uniform prior distribution of the 20 amino acids, they obtained $H(\underline{\boldsymbol{X}}) = L \cdot 4.322$ bits. The extreme contributions to $I(\underline{\boldsymbol{X}}_f)$ at a site j occur when either one amino acid is completely conserved ($d_{xj} = M$ for some x, with a per-site contribution 4.322 to $I(\underline{\boldsymbol{X}}_f)$) or when all 20 amino acids occur with the same frequency ($d_{xj} = M/20$ for all x, with a per-site contribution 0 to $I(\underline{\boldsymbol{X}}_f)$).

Durston et al. (2007) applied this method to determine the lower bound of bits to 35 protein families from the Pfam database to estimate the information of some important biological

functions. Pfam (https://pfam.xfam.org/browse) is a widely used repository of protein family HMMs. Twelve examples with more than 500 bits were found. The highest value reported was for the protein family Flu PB2, with 2,416 bits.

It is possible to generalise conditional mutual information to arbitrary prior distributions. Using a similar argument as in (4), we obtain

$$I(\underline{\boldsymbol{X}}_f) = H(\underline{\boldsymbol{X}}) - H(\underline{\boldsymbol{X}}|\underline{\boldsymbol{X}}_f) = -L \cdot \boldsymbol{p} \cdot \log_2 \boldsymbol{p} + \sum_{j=1}^{L} \boldsymbol{q}_j \cdot \log_2 \boldsymbol{q}_j, \tag{5}$$

where in our setting the prior distribution $\boldsymbol{p}$ will be chosen from Table 3 and the posterior probabilities of the second term are obtained from Bayes' rule as

$$\frac{p_x r_{xj}}{\sum_{y=1}^{20} p_y r_{yj}} = q_{xj}$$

at a site with no gaps, making use of the fact that the observed protein family was obtained through sampling with censoring, so that the noncensoring probabilities r_{xj} (cf. (1)) appear in the likelihood. By applying the distribution $\boldsymbol{p}$ from Table 3, we obtain $H(\underline{\boldsymbol{X}}) = L \cdot 4.139$ bits, which is somewhat less than Durston assessed from his model with a uniform prior.

Another possibility is to apply the expected active information to the protein family as

$$I^+(\underline{\boldsymbol{X}}_f) = \sum_{j=1}^{L} E(I^+_{X_j}) = E_q^{(p)}(\underline{\boldsymbol{X}}_f) - E_q^{(q)}(\underline{\boldsymbol{X}}_f) = -\sum_{j=1}^{L} \boldsymbol{q}_j \cdot \log_2 \boldsymbol{p} + \sum_{j=1}^{L} \boldsymbol{q}_j \cdot \log_2 \boldsymbol{q}_j, \tag{6}$$

where X_j is a randomly chosen amino acid from Column j of $\underline{\boldsymbol{X}}_f$. It follows from Section 2.2 that (6) equals the total Kullback–Leibler divergence $\sum_{j=1}^{L} D_{KL}(\boldsymbol{q}_j||\boldsymbol{p})$ between the prior and posterior distributions $\boldsymbol{p}$ and $\boldsymbol{q}_j$, summed over all loci.

2.6 Methodological overview

In Table 4, we summarise some of the properties of the various methods for measuring information at sites of amino acid alignments. Methods (2) and (6) have the advantage of being nonnegative and (2) additionally approximates the functional information in Equation (3). Equation (6) does not qualify as a distance measure, because it is asymmetric and path-dependent, and in contrast to (2) it does not satisfy the triangle inequality (Cover & Thomas, 2006). On the other hand, Method (5) has the conditional mutual information interpretation. It needs no correction in order for random sequences to have approximately zero information, whereas the derivations of Appendix A reveal that such a correction is needed for (2) (and by a similar type of argument, a correction term of (6) is obtained as well).

3 Results

It is difficult to experimentally determine the connection between functional information and activity because of the extreme scarcity of functional sequences in populations of random sequences. Since there are multiple sequences with a given expression, the corresponding functional information will always be lower than the measure of the information needed to specify any particular sequence. It is important to notice that functional information is not a property of any one molecule, but of the collection of all optional sequences classified by activity.

Proteins generally have one or more functional regions termed *domains*. A protein domain is a region of the protein sequence that folds independently of the rest to form a distinct structural unit. *CATH* is a hierarchical classification of protein domain *structures* (Figure 2), which clusters proteins at four major levels, abbreviated by the letters CATH (Sillitoe et al., 2021). In this hierarchy, the domains are classified at the Class (C) level (1 = all alpha, 2 = all beta, 3 = mixture of alpha and beta, 4 = few secondary structures), Architecture (A) level (information on the gross secondary

structure arrangement in three-dimensional space), Topology/fold (T) level (information on how the secondary structure elements is connected and numbers of secondary structures), and Homologous superfamily (H) level that clusters domains with highly similar sequences and functions. By combining protein structure and sequence, the CATH resource provides comprehensive structure-based domain family assignments to millions of protein sequences. Table 5 shows the functional information of five selected alignments with different architectures from each CATH class, except Class 4, which contains only one architecture. More specifically, Table 5 shows the functional information and mutual information for the selected protein domains determined by the methods described in the present article (equations (2) and (5), respectively).

The same table also presents four randomly generated alignments analysed by the methods described in the present article (equations (2) and (5)). For a randomly generated protein family with M amino acid sequences, it follows from the Central Limit Theorem that the information content density $I^{(p)}(\underline{\boldsymbol{X}}_f)/\bar{L}$ obtained from (2) is always positive and of the order $1/\sqrt{M}$ (see Appendix A for more details), whereas the corresponding information content density obtained from (5) approximately equals 0.

Four of the CATH domains in Table 5 have a functional information density below 2 bits/site. The main reason is that these alignments contain gaps in 30% or more of their sites, and the gaped sites are estimated with zero information. If we skip the gap sites, the density of the remaining sites is in line with that of the other domains.

Pfam is the largest database of functional families and domains, and it is represented by sequence alignments (Mistry et al., 2021; Wang et al., 2021). It was trained on a representative set of aligned sequences that are known to belong to the unit (the 'seed' alignment). From this database, it is possible to retrieve a FASTA file of all known sequences of a domain and align them. We generated alignments of 30 selected functional domains from Pfam and studied them using the framework developed in this study. Table 6 presents the results calculated from the sequence data, jointly with the defined functionality variable (cf. Durston et al., 2007).

The scatter plot in Figure 3 compares the information density estimated by Models A and B in Table 5 (CATH data) and Table 6 (Pfam data). The CATH data show higher correlations than the Pfam data.

To demonstrate further how this approach can be visualised, a sample plot of functional information along the aligned sequence of the first domain of Table 5 is shown in Figure 4, making use of equation (2) (Model A). From the plot, one can observe which sites have higher measured functional information and possibly play a critical role in either the structure or the binding site of that protein domain. If the structure of the protein domain is known, one may also generate similar 3D plots.

A few protein groups have also been studied experimentally in labs to estimate the prevalence of biological functions (Axe, 2004; Ferrada & Wagner, 2010; Kozulic & Leisola, 2015). By studying one of the subdomains of a typical Class A β-lactamase of $L = 153$ amino acids, Axe found a prevalence of performing this function *via any fold* in the range between 10^{-53} and 10^{-77}. Hence, it follows from (3) that the functional information satisfies

$$176 \text{ bits} = -\log_2(10^{-53}) \leq I_{\beta\text{-lactam}}(\underline{\boldsymbol{X}}_{\beta\text{-lactam}}) \leq -\log_2(10^{-77}) = 256 \text{ bits},$$

which corresponds to bits of information per site between 1.151 and 1.672. This is an estimate of the overall prevalence of adopting functional folds by supporting a working active site, not restricted to specific domains and families as the examples in Tables 5 and 6. Hence, the number of bits per site is lower than in the tables.

Only a tiny fraction of the total sequence space conforms to the specific structural and functional characteristics of a particular protein or protein family. This sparseness of functional protein sequences in the sequence space, and the ruggedness of the protein energy landscape with minor amounts of Gibbs free energy of unfolding, are emphasised by observations from studies of prokaryotic genomes (Povolotskaya & Kondrashov, 2010).

Even a library of 100-residue proteins with the mass of the Earth itself (5.98×10^{24} kg) would comprise at most 3.3×10^{47} different sequences (Taylor et al., 2001). Therefore, only a combinatorial approach that couples modular design with mutagenesis and selection is appropriate to

Table 4. Various models for measuring bits of information $I(\underline{\boldsymbol{X}}_f)$ at a single site $j = 1$ ($L = 1$) in an alignment $\underline{\boldsymbol{X}}_f$ of M amino acids with distribution $\boldsymbol{q} = \boldsymbol{q}_j$

Model	Range of values [min, max] when q_j varies	Value for conserved amino acid x ($q_{xj} = 1$)	Value for distribution $q_j = p$
(2)	[0, 5.931]	$-\log_2 p_x$	0
(4)	[0, 4.322]	$\log_2(20) = 4.322$	$0.183 = 4.322 - 4.139$
(5)	[−0.183, 4.139]	4.139	0
(6)	[0, 5.931]	$-\log_2 p_x$	0

Note. The first column refers to equations numbered in the article, whereas Columns 2–4 display properties of $I(\underline{\boldsymbol{X}}_f)$ when $\boldsymbol{q}_j$ varies, with null distribution $\boldsymbol{p}$ as in Table 3. The max information of 5.931 bits in Models (2) and (6) occurs if either an amino acid M or W is conserved. Models (2) and (5) are elaborated further in Tables 5 and 6 as Models A and B.

successfully introduce known or new catalytic activity. Current biotechnological approaches for developing functional de novo proteins include rational design, computational optimisation, and selection from combinatorial libraries (Smith & Hecht, 2011).

4 Discussion and concluding remarks

In the study above, we have shown how variants of the Shannon information measure can be applied to a variety of molecular sequence data sets. We put forward a series of very specific analyses constrained by a priori assumptions about the underlying probability distributions of sequences from which sample data have been obtained, before all nonfunctional sequences are censored. In the study, we also assumed that aligned sequences from the CATH and Pfam databases could be assigned the same functionality, and a measure of functional information in bits for each site, based on rejection sampling, was computed from their aligned sequences. The results for the analysed protein domains and families, as well as some arrays of randomly generated sequences, are shown in Tables 5 and 6, and in Figures 3 and 4.

The *functional information,* estimated by rejection sampling (Model A), reveals similar results as the conditional *mutual information* results (Model B) for the CATH data. However, results from the Pfam data are more diverse. Both the CATH and Pfam databases are populated by sequences taken from the genome database *UniprotKB/Swiss-Prot*, who store several identical copies of a sequence if they are detected in different species. Multiple copies of sequences imply that the corresponding amino acids will look more conserved. This may explain some of the observed variation between Methods A and B for the Pfam alignments, suggesting that the multiple copies have a larger effect on Model B. Indeed, it follows from the third column of Table 4 that whereas Model B always gives the same values at conserved sites, the values of Model A at such sites will depend on the rareness of the conserved amino acids, with common conserved variants leading to a smaller functional information. Consequently, if a majority of the multiple copies of the Pfam alignments correspond to common amino acid variants, if follows that the conditional mutual information should be larger than the functional information. To elaborate further on this, one might screen the alignments downloaded from the databases for multiple copies of identical sequences to represent each unique sequence only once, regardless of whether they are from different species or not.

There is a complex relation between protein sequence, structure, and function. Protein function is directly related to the resulting 3D structure of the sequence. This structure–function correlation is high, implying that 3D conservation is more important than the primary amino acid sequence (Hvidsten et al., 2009; Sousounis et al., 2012). Function and structure may be conserved in the context of large sequence differences. Hence, the CATH data based on structure will be more relevant than the Pfam data when estimating information related to function. However, the CATH homologous superfamily (H) also introduces sequence similarity into the data and therefore obscures the pure structure–function signal.

There are some limitations to the approach of the present study. Genes and proteins may have similar biochemical functions, without any noticeable sequence similarity, as mentioned in the

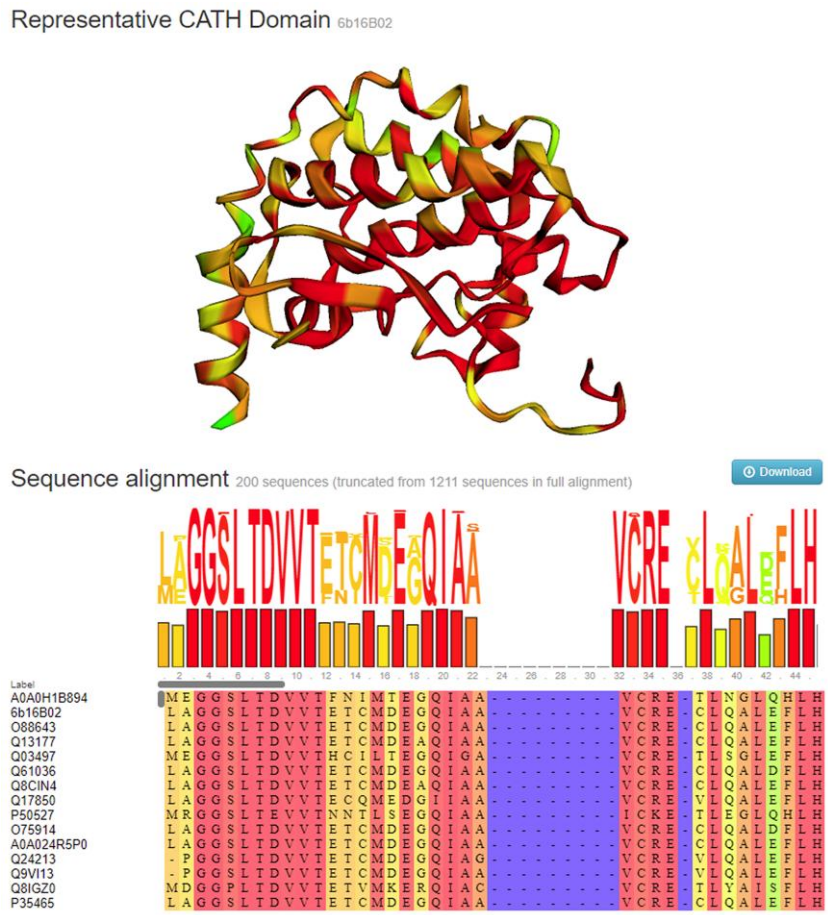

Figure 2. Structure and sequence alignment of the nonspecific serine/threonine protein *kinase*. From CATH superfamily 1.10.510.10 at http://www.cathdb.info/browse/.

introduction. These *isoenzymes* vary in sequence but catalyse the same reaction (Guzzi et al., 2012). Methods used in the present study only work for aligned orthologues and must handle isoenzymes as separate groups. The information content of these groups may be compared.

The existence of isoenzymes may explain why studies of functional information via any 3D fold (Section 2.4) provides estimates of 1.15–1.67 bits per site, while our results based on orthologue sequences are higher (mainly between 2.37 and 4.40 bits per site, see Method A of Tables 5 and 6).

The calculation of functional information for proteins employs all recognised sequences of a domain family out of the sequence space that displays any degree of that family's biofunction. It does not deal with the *degree* of functionality (e.g. the reaction rate) of any one protein in that family. However, to the best of our knowledge our self-information approach (2) comes closer to an information measure of biofunction than any other measures in the literature, and we conjecture that degree of functionality can be incorporated into the noncensoring probabilities (1).

Another premise of our computational models of Sections 2.3 and 2.5 is that functionally equivalent amino acids at each site are independent of those at any other site. This is not valid if there is a linkage between sites that are close to each other in 3D, such as a salt-bridge configuration. Consequently, the computed amount of information, under the assumption of independence, could be lower than the real amount, and the protein families contain more bits of information and represent a much smaller subset among random sequences. In order to account for such dependencies between sites, we developed a more general estimate of functional information in Appendix C. However, the effect of these site-dependencies is to some extent balanced by

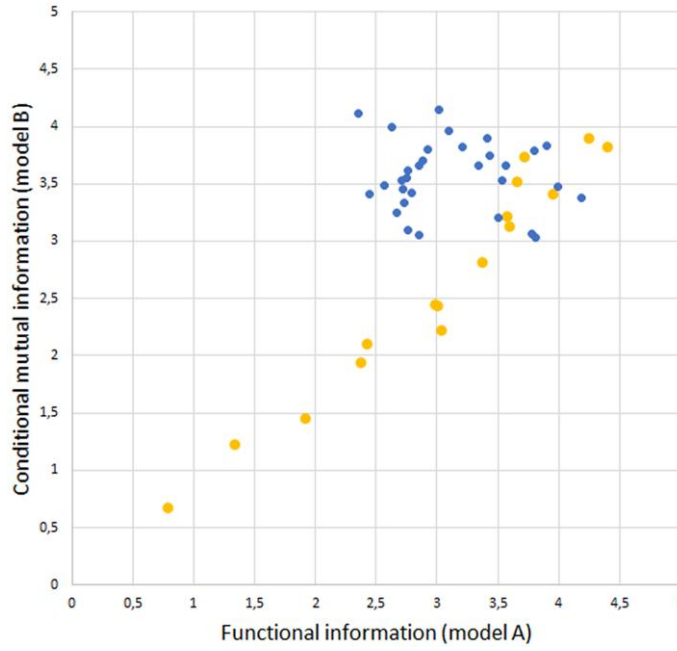

Figure 3. A scatter plot comparison of information density (information per site) estimated by the functional information model (a) ($= I^{(\boldsymbol{p})}(\underline{\boldsymbol{X}}_{\boldsymbol{f}})/\bar{L}$) and the conditional mutual information model (b) ($= I(\underline{\boldsymbol{X}}_{\boldsymbol{f}})/\bar{L}$) for a number of protein families $\underline{\boldsymbol{X}}_{\boldsymbol{f}}$ of average sequence length $\bar{L}$. The big yellow dots are the protein domains/families from the CATH database (Table 5), and the smaller blue dots are the protein families from Pfam (Table 6).

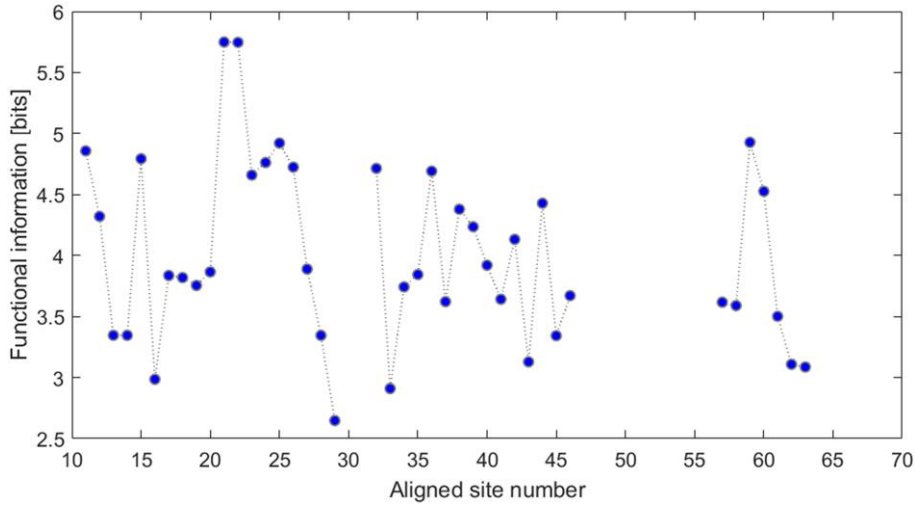

Figure 4. The information profile shows how information values vary between sites for the domain *elongation factor ts* (entry 4r71C01 of CATH superfamily 1.10.8.10). There are $M = 2295$ sequences in the full sequence alignment, and only two sites are fully conserved (amino acid L at site 13 and amino acid R at site 14). Despite of the fact that these two sites are conserved, information is not largest there, because these amino acids are more likely a priori (probability 6/61). The peak at sites 21 and 22 is dominated by amino acid M (probability 1/61), in spite of the fact that M is not fully conserved at these two sites. Sites containing gaps are unfilled.

Table 5. Evaluation of the functional information for 16 CATH protein domains $\underline{\boldsymbol{X}}_f$, with rows $\boldsymbol{X}_m$ of lengths L_m for $m = 1, \ldots, M$, where M is the number of proteins of each family/domain

		Sample statistics (Section 2.3.1)				Functional and mutual information (Sections 2.4 and 2.5)				
CATH	Number of Seq. M	Mean length $\bar{L}$ (aa)	Conserved sites	Mean of $I(X_m)$ (SD) (bits)	Density $\overline{I^{(p)}(\underline{X}_f)}$ (SD) (bits/aa)	Sites used L	Model A: $I^{(p)}(\underline{X}_f)$ (bits)	Model A: density $I^{(p)}(\underline{X}_f)/\bar{L}$ (bits/site)	Model B: $I(\underline{X}_f)$ (bits)	Model B: density $I(\underline{X}_f)/\bar{L}$ (bits/site)
1.10.8.10	2,295	55.0	2	240 (5)	4.35 (0.07)	41	164	2.98	135	2.45
1.20.5.10	578	44.3	2	193 (4)	4.36 (0.05)	35	135	3.04	98	2.22
1.25.10.10	951	232.6	106	988 (9)	4.25 (0.01)	206	851	3.66	820	3.52
1.40.20.10	11	284.4	258	1,173 (38)	4.12 (0.01)	258	1,059	3.72	1,063	3.74
1.50.10.10	356	470.6	53	2,040 (44)	4.33 (0.02)	284	1,143	2.43	987	2.10
2.10.10.10	190	50.6	19	228 (10)	4.49 (0.03)	46	200	3.95	173	3.41
2.20.20.101	24	42.0	27	191 (2)	4.55 (0.06)	41	185	4.40	160	3.82
2.30.18.10	333	53.6	9	234 (11)	4.36 (0.03)	39	161	3.01	131	2.44
2.40.10.104	581	104.4	1	446 (20)	4.27 (0.07)	*24*	82	0.79	70	0.67
2.50.20.10	296	182.2	71	792 (4)	4.35 (0.02)	181	774	4.25	711	3.90
3.10.10.10	972	130.0	19	572 (14)	4.40 (0.01)	*40*	175	1.34	159	1.23
3.15.10.10	234	176.7	0	747 (21)	4.23 (0.06)	108	338	1.92	256	1.45
3.20.10.10	197	173.4	51	750 (11)	4.33 (0.01)	147	620	3.58	559	3.22
3.30.9.10	938	156.5	6	663 (12)	4.24 (0.07)	107	371	2.37	303	1.94
3.40.5.10	384	56.1	8	248 (14)	4.42 (0.03)	48	189	3.37	158	2.82
4.10.70.10	63	67.9	18	296 (17)	4.36 (0.04)	60	245	3.60	213	3.13
Ran. 1 (prior distribution)	100	100	0	413 (8)	4.13 (0.08)	100	101	1.01	15	0.15
Ran. 2 (prior distribution)	1,000	100	0	414 (7)	4.14 (0.07)	100	42	0.42	1	0.01
Ran. 3 (prior distribution)	1,000	10,000	0	41,392 (71)	4.14 (0.01)	10,000	4,022	0.40	137	0.01
	10,000	1,000	0	4,139 (22)	4.14 (0.02)	1,000	136	0.14	1.32	0.00

(*continued*)

Table 5. Continued

CATH	Number of Seq. M	Sample statistics (Section 2.3.1)				Functional and mutual information (Sections 2.4 and 2.5)				
		Mean length $\bar{L}$ (aa)	Conserved sites	Mean of $I(X_m)$ (SD) (bits)	Density $\overline{I^{(p)}(\underline{X}_f)}$ (SD) (bits/aa)	Sites used L	Model A: $I^{(p)}(\underline{X}_f)$ (bits)	Model A: density $I^{(p)}(\underline{X}_f)/\bar{L}$ (bits/site)	Model B: $I(\underline{X}_f)$ (bits)	Model B: density $I(\underline{X}_f)/\bar{L}$ (bits/site)
Ran. 4 (prior distribution)										
Prior sequences	61	100	0	413.9	4.14	100	0	0	0	0

Note. The table contains for each family its name (Column 1), the number M of sequences analysed for each domain/family (Column 2), its mean sequence length $\bar{L} = \sum_{m=1}^{M} L_m/M$ (Column 3), its mean Information content $\sum_{m=1}^{M} I^{(p)}(\mathbf{X}_m)/M$ (bits, Column 5) and mean information density (bits/aa, Column 6, cf. Section 2.3) with standard deviations (SD) included. In Columns 8 and 10, approximations of the *functional information* in bits are calculated based on the models in equation (2), corresponding to Model A, and equation (5), corresponding to the *conditional mutual information* of Model B, respectively. Column 7 displays the number of nongaped sites in the alignment applied in these estimates, as sites with one or more gaps are considered to carry zero information. The density of functional or mutual information, defined as the functional or mutual information/length (bits/amino acid), is also shown for Models A and B in Columns 9 and 11, respectively. For comparison, results are also shown for four randomly generated amino acid samples, for which the components of $\underline{\mathbf{X}}_f$ are drawn randomly from $\boldsymbol{p}$, whereas the last row corresponds to a family of $M = 61$ amino acid sequences where the frequencies of each column of $\underline{\mathbf{X}}_f$ perfectly match the prior distribution ($\boldsymbol{q}_j = \boldsymbol{p}$, for $j = 1, \ldots, L$). The functional information values of Method A can be interpreted as a measure of the shift in functional uncertainty required to specify any functional sequence that falls into the given domain. Alignments are obtained from http://www.cathdb.info/browse/(version 3.3).

Table 6. Evaluation of the functional information for 30 Pfam protein domains/families $\underline{\boldsymbol{X}}_f$, with rows $\boldsymbol{X}_m$ of lengths L_m for $m = 1, \ldots, M$, where M is the number of proteins of each domain or family

		Sample statistics (Section 2.3.1)			Functional and mutual information (Sections 2.4 and 2.5)				
Name	Number of Seq. M	Mean length $\bar{L}$ (aa)	Mean of $I(X_m)$ (*SD*) (bits)	Density $\overline{I^{(p)}(\underline{X}_f)}$ (SD) (bits/aa)	Sites used L	Model A: $I^{(p)}(\underline{X}_f)$ (bits)	Model A: density $I^{(p)}(\underline{X}_f)/\bar{L}$ (bits/site)	Model B: $I(\underline{X}_f)$ (bits)	Model B: density $I(\underline{X}_f)/\bar{L}$ (bits/site)
ANKH	427	292.0	1,232 (343)	4.22 (0.05)	290	1,165	3.99	1,013	3.47
HTH 8	36,806	41.2	172 (11)	4.20 (0.12)	40	117	2.85	126	3.05
HTH 7	4,162	43.5	181 (17)	4.17 (0.12)	43	122	2.79	148	3.42
HTH 5	24,116	47	192 (11)	4.11 (0.12)	47	127	2.92	178	3.80
HTH 11	19,477	55	225 (17)	4.12 (0.13)	54	149	2.72	189	3.45
HTH 3	85,856	53.9	225(17)	4.17 (0.12)	55	148	2.75	191	3.55
Insulin	277	70.7	300 (167)	4.24 (0.10)	71	214	3.02	293	4.15
Ubiquitin	47,663	70.4	300 (31)	4.26 (0.08)	71	242	3.43	264	3.75
Kringle domain	1,083	78.0	343 (27)	4.39 (0.07)	78	260	3.34	286	3.66
VPR	38	89.7	388 (25)	4.32 (0.07)	89	339	3.78	274	3.06
RVP	1,921	97.0	411 (61)	4.24 (0.09)	97	259	2.67	315	3.25
Acyl-Coa dh N-dom	82,288	112.9	480 (45)	4.25 (0.09)	110	305.9	2.71	398.7	3.53
MMR HSR1	96,703	126.1	531 (101)	4.21 (0.08)	125	331.3	2.63	503.0	3.99
FtsH	9,331	100.5	429 (56)	4.26 (0.08)	100	246	2.45	343	3.41
Ribosomal S7	11,957	146.5	622 (79)	4.25 (0.07)	147	499	3.41	571	3.90
P53 DNA domain	1,799	181.6	775 (136)	4.27 (0.03)	183	708	3.90	695	3.83
Vif	32	192.0	832 (11)	4.33 (0.03)	192	731	3.81	581	3.03
SRP54	22,629	194.7	824 (102)	4.23 (0.06)	194	625	3.21	743	3.82
Ribosomal S2	14,921	158.7	681 (278)	4.28 (0.06)	160	561	3.53	607	3.53
Viral helicase1	1,224	198.2	839 (292)	4.23 (0.08)	197	466	2.35	814	4.11

(*continued*)

Table 6. Continued

Name	Number of Seq. M	Sample statistics (Section 2.3.1)			Functional and mutual information (Sections 2.4 and 2.5)				
		Mean length $\bar{L}$ (aa)	Mean of $I(X_m)$ (SD) (bits)	Density $\overline{I^{(p)}(\underline{X}_f)}$ (SD) (bits/aa)	Sites used L	Model A: $I^{(p)}(\underline{X}_f)$ (bits)	Model A: density $I^{(p)}(\underline{X}_f)/\bar{L}$ (bits/site)	Model B: $I(\underline{X}_f)$ (bits)	Model B: density $I(\underline{X}_f)/\bar{L}$ (bits/site)
Beta-lactamase 2	7,057	202.7	851 (174)	4.20 (0.10)	201	521.1	2.57	707.7	3.49
RecA	9,621	248.8	1,049 (190)	4.22 (0.05)	251	945.9	3.80	943.2	3.79
tRNA-synt 1b	25,665	274.5	1,178 (224)	4.29 (0.07)	275	782	2.85	1,005	3.66
SecY	11,699	335.6	1,410 (258)	4.20 (0.04)	334	1,039	3.10	1,329	3.96
EPSP Synthase	1,971	392.6	1,643 (301)	4.19 (0.07)	395	1,129.4	2.88	1,453.4	3.70
FTHFS	6,142	529.3	2,247 (582)	4.24 (0.05)	530	1,884.7	3.56	1,935.8	3.66
DctM	24,837	403.3	1,673 (312)	4.15 (0.05)	402	1,114.7	2.76	1,458.8	3.62
Corona S2	63	39.9	178 (19)	4.45 (0.07)	39	140	3.50	128	3.20
Flu PB2	55	628.6	2,691 (852)	4.28 (0.03)	665	2,631	4.19	2,128	3.38
Usher	1,786	448.8	1,876 (656)	4.17 (0.07)	451	1,239.6	2.76	13,901	3.10
ACR Tran	5,473	811.5	3,392 (1,360)	4.18 (0.06)	811	2,215.1	2.73	2,709.0	3.34

Note. The notation is the same as in Table 5. Column 6 displays the number of sites in the alignment used in these estimates, as sites with gaps above a threshold (60%–95%) are considered to carry no information. Thresholds are adjusted so that the number of sites matches the average number of amino acids in the sequences (Column 3). Alignments are obtained from https://pfam.xfam.org/browse (version 34.0).

the fact that not all organisms that ever lived are represented in the dataset, indicating the measures of information of Sections 2.3 and 2.5 could still be used as approximations of the actual functional information.

Another quantity that affects all measures of information studied in this article is the prior deduced from the genetic code. We assumed that amino acids are independent a priori over sites and sequences, with a distribution $\boldsymbol{p}$ obtained from Table 3 under the assumption that all nonstop codons are equally likely a priori. This assumption was relaxed in Appendices A and C by allowing for a site-dependent a priori distribution $\boldsymbol{p}_j$, still assuming a priori independence between sites. Such an extended model will typically lower the functional and conditional mutual information, at least if $\boldsymbol{p}_j$ is closer than $\boldsymbol{p}$ to $\boldsymbol{q}_j$ for most sites j. A further extension is to allow for a priori dependencies between sites, and this might further lower the functional and conditional mutual information. Regardless of which type of prior that is used, it is of interest to estimate it from the data, and thereby adapt it to the coding region of the studied protein family $\underline{\boldsymbol{X}}_f$. For instance, for the simplest model with a priori independence over sites and sequences, the prior distribution might be estimated empirically from $\underline{\boldsymbol{X}}_f$ as $\hat{\boldsymbol{p}} = \sum_{j=1}^{L} \boldsymbol{q}_j / L$. This will typically lower the functional and conditional mutual information, compared to using an a priori fixed prior $\boldsymbol{p}$.

Although the same functionality may not be applicable to individual sites, both site significance and independence between sites are assumed in Sections 2.3 and 2.5 when estimating the site information and summing it up for the entire alignment, whereas short-range Markov process dependencies are treated in Appendix C. More general scenarios may be considered for observed long-range dependencies between segmented regions in the alignment based on conserved as well as interdependent patterns (Durston et al., 2012).

The ability to measure information in bits at each *site* of a protein domain may be applied to locate key functional components of a gene. Durston et al. (2007, 2012) observed in their analysis of *ubiquitin*, that six of the seven sites with the largest number of bits per site were clustered around the binding site. Surprisingly, the binding site itself was poorly conserved, although it had a relatively high bits/site information.

We may also perform *simulation studies* of our models using the notion of expected self-information (or cross-entropy) $E_{\boldsymbol{v}}^{(\boldsymbol{p})}$, as introduced in Section 2.2, based on the prior distribution $\boldsymbol{p}$ of Table 3 and another given distribution $\boldsymbol{v}$. The random sequence $\boldsymbol{X}$ of maximal length L is then sampled from $\boldsymbol{v}$, whereas its self-information is computed from $\boldsymbol{p}$. We recall from Section 2.2 that the Kullback–Leibler divergence between $\boldsymbol{v}$ and $\boldsymbol{p}$ tells how much $E_{\boldsymbol{v}}^{(\boldsymbol{p})}$ exceeds the entropy $H^{(\boldsymbol{v})} = E_{\boldsymbol{v}}^{(\boldsymbol{v})}$ of $\boldsymbol{v}$. On the other hand, in Appendix A, we showed that our estimate of functional information corresponds to replacing the Kullback–Leibler divergence by another distance that satisfies the triangle inequality. The relation between these two measures of information, and the other models of Table 4, may be further elaborated by simulation studies and analysing more alignments to find out more on how accurate the methods are regarding sequence similarity, and provide some similarity threshold above which the methods can be expected to no longer work.

Genes contain instructions, which are a type of effective procedural information, such as algorithms. In this sense, genes represent special types of respectable informational entities, which are in themselves instructions or algorithms. This interpretation of genetic information is compatible with Shannon's probabilistic theory of information but is less demanding than a full semantic interpretation. However, the basic dichotomy between syntactic and semantic information requires better coupling and coherent understanding to develop an integrated and more global theory of information for genetic systems. The explicit consideration of *functional information* in the present study bridges the syntactic and semantic information concepts and leads to operational statements and empirical testable hypotheses about the role of information in genetic systems. This can help in the challenging task of discovering environmental adaptation or genetic innovation by quantifying changes in information. Functional information profiling of the proteins is a significant step in understanding the role of gene sequences in the context of the full genetic repertoire of an organism.

Given its restricted nature, probability-based self-information (or expected active information), functional information, and conditional mutual information are important for measuring information also in other contexts than presented here. Their ability to quantify genetic information has the same utility as any other quantitative method. It allows us to analyse and compare systems from the perspective of their informativeness. Mutational drift, emerging pathogenic viral, and

microbial strains, can be evaluated quantitatively as well as in a qualitative way. In principle, it may be possible to estimate the values of functional information in bits for all proteins and protein sites in a virus or cell if all the translated genes are known.

There exist other formalisms than information theory to study the information content in genetic sequences. Formally, the *Kolmogorov algorithmic information*, or complexity, of a string $\boldsymbol{X}$ of bits is the length of the shortest computer program that generates string $\boldsymbol{X}$ and stops. Kolmogorov complexity is related to the compression of data. A nontrivial string may be algorithmically incompressible and requires an algorithm or instruction set of complexity, such as the system it describes. Numerous programs will generate $\boldsymbol{X}$, but the Kolmogorov measure of complexity discredits from the fact that it is algorithmically unknowable, as there is no general method to compute it (Cover & Thomas, 2006). Nevertheless, it is possible to derive upper bounds for the Kolmogorov complexity, and accordingly, it is bounded without being computed exactly. Viruses and archaea have a higher relative DNA complexity than bacteria and eukaryotes (Pratas & Pinho, 2017). Several protein compression methods have been proposed in the literature. For a review, we refer to Hosseini et al. (2016). Interestingly, the Kolmogorov complexity may be estimated from its output frequency distribution (Soler-Toscano et al., 2014), just as the measures of information presented in this article. It is an interesting topic of further research to compare (upper bounds of) Kolmogorov complexity with other measures of information such as functional information.

A grand unified theory of biological information may be intangible, perhaps even fundamentally inconceivable, given the uncontained use of the term. There may be no direct quantifiable framework for mathematical biology in the same manner as well-established mathematical physics (Chaitin 1979). There is much more information present in a biological system than can be counted by plain and straightforward observation; therefore, its quantification by counting of alleles within genes (or of amino acids within proteins) amounts to gross bias by discarding. There is a global biological complexity within a set of hierarchical levels from DNA to ecology (Farnsworth et al., 2012; Griffiths, 2017). The flow and accumulation of information in ecological systems is information processing, which integrates information in multiple forms (O'Connor et al., 2019). Extracting knowledge from different information levels can be fruitful for sequence data analysis. Meta-information like temperature, salinity, pH, pressure, etc. may be implemented. The present formalism may be extended to study time-series of biological sequences, as 'sequences of sequences'. Developing methods for observing, quantifying, and tracing information remains the target of research efforts across disciplines. We hope that the present contribution can help to serve as the first steps to obtain the common ground needed to build a general, more satisfactory theory of biological information.

In biology, information has both a probabilistic and qualitative dimension over an observable dataset. Through the representation and use of information and its pragmatic assessment, there is substantial justification for considering biology within such an informational platform. In this article, we have presented some advances in terms of quantifying genetic information as a joint variable of function and sequence data, and to estimate the corresponding functional information through rejection sampling. Both structure-based and sequence-based domains, which are available in CATH and Pfam databases respectively, are promising representatives of functional regions within proteins with quantifiable information content. Information is a conceptual key to a proper understanding of reality that reveals itself as we do science. Despite the large amount of evidence that information plays a vital role in genetic systems, quantitative information processing does not yet feature much in the genetic principles at the centre of mainstream theories and textbooks. This absence is problematic and disconnects genetics from other scientific disciplines. Scientific and philosophical thinking and work should rather, with increasing confidence, place information 'as one of three elemental components of existence (along with space/time and energy/matter)' (Atmar 2001).

The present models are implemented in MATLAB, and the routines can be downloaded from the next version of the *DeltaProt* toolbox (Thorvaldsen et al., 2010).

Acknowledgments

The authors thank Professor Peter Øhrstrøm for helpful discussions during the preparation of the article. We also greatly appreciate the comments from the editor, and all reviewers involved, who helped improving the quality of the article.

Appendix A

Suppose $\underline{\boldsymbol{X}}_f = (X_{mj})$ is a protein family, whose amino acids are assumed a priori to be independent with locus-dependent distributions $X_{mj} \sim \boldsymbol{p}_j = (p_{1j}, \ldots, p_{20,j})$ having all its elements > 0. This is slightly more general than the framework of Section 2, where $\boldsymbol{p}_1 = \cdots = \boldsymbol{p}_L = \boldsymbol{p}$ was assumed. Let also $\boldsymbol{q}_j = (q_{1j}, \ldots, q_{20,j})$ contain the observed amino acid frequencies of $\underline{\boldsymbol{X}}_f$ at site j. Define matrices $\underline{\boldsymbol{p}} = (\boldsymbol{p}_1^T, \ldots, \boldsymbol{p}_L^T) = (p_{xj})$ and $\underline{\boldsymbol{q}} = (\boldsymbol{q}_1^T, \ldots, \boldsymbol{q}_L^T) = (q_{xj})$ of dimension $20 \times L$ that contain the prior and observed amino acid distributions of $\underline{\boldsymbol{X}}_f$ at all loci, respectively. For simplicity, consider only non-gapped sites j and assume that $\underline{\boldsymbol{q}}$ is obtained from a pool of amino acids with distribution $\underline{\boldsymbol{p}}$ through a censoring mechanism. We generalise the definition of (1) of the noncensoring probabilities to

$$r_{xj} = \frac{q_{xj}/p_{xj}}{\max(q_{1j}/p_{1j}, \ldots, q_{20,j}/p_{20,j})},$$

so that the prior probabilities p_{xj} are locus-dependent. The self-information (2) of $\underline{\boldsymbol{X}}_f$ similarly generalises to

$$I^{(p)}(\underline{\boldsymbol{X}}_f) = -\sum_{j=1}^{L} \log_2(\boldsymbol{p}_j \cdot \boldsymbol{r}_j) = D(\underline{\boldsymbol{q}}, \underline{\boldsymbol{p}}) = \sum_{j=1}^{L} D(\boldsymbol{q}_j, \boldsymbol{p}_j), \tag{7}$$

where

$$D(\boldsymbol{q}_j, \boldsymbol{p}_j) = -\log_2(\boldsymbol{p}_j \cdot \boldsymbol{r}_j) = \log_2 \max\left(\frac{q_{1j}}{p_{1j}}, \ldots, \frac{q_{20,j}}{p_{20,j}}\right) - \log_2 \sum_{x=1}^{20} q_{xj}$$

$$= \max \log_2 \left(\frac{q_{1j}}{p_{1j}}, \ldots, \frac{q_{20,j}}{p_{20,j}}\right).$$

Note that $D(\boldsymbol{q}_j, \boldsymbol{p}_j)$ differs from the Kullback–Leibler divergence $D_{\mathrm{KL}}(\boldsymbol{q}_j || \boldsymbol{p}_j)$ in that a weighted summation of all $\log_2(q_{xj}/p_{xj})$ for $x = 1, \ldots, 20$, with weights q_{xj}, is replaced by a maximum operation. Moreover, the information measure (7) differs from (6) in that the distance D replaces the Kullback–Leibler divergence D_{KL}.

Now assume that the output of the first sampling mechanism, that generated $\underline{\boldsymbol{X}}_f$, is the input of a second sampling mechanism that generates a new protein family $\underline{\boldsymbol{Y}}_f = (Y_{mj})$ of dimension $M \times L$. That is, the observed amino acid distribution $\underline{\boldsymbol{s}} = (s_{xj})$ of $\underline{\boldsymbol{Y}}_f$ is obtained from a pool of amino acids with distribution $\underline{\boldsymbol{q}}$, corresponding to a self-information

$$I^{(q)}(\underline{\boldsymbol{Y}}_f) = D(\underline{\boldsymbol{s}}, \underline{\boldsymbol{q}}) = \sum_{j=1}^{L} D(\boldsymbol{s}_j, \boldsymbol{q}_j).$$

On the other hand, the self-information of $\underline{\boldsymbol{Y}}_f$ for a combined sampling procedure, with a pool of amino acids with frequencies $\underline{\boldsymbol{p}}$, and observed frequencies $\underline{\boldsymbol{s}}$, is

$$I^{(p)}(\underline{\boldsymbol{Y}}_f) = D(\underline{\boldsymbol{s}}, \underline{\boldsymbol{p}}) \leq I^{(p)}(\underline{\boldsymbol{X}}_f) + I^{(q)}(\underline{\boldsymbol{Y}}_f) = D(\underline{\boldsymbol{q}}, \underline{\boldsymbol{p}}) + D(\underline{\boldsymbol{s}}, \underline{\boldsymbol{q}}).$$

In order to prove this triangle inequality, since D is additive over loci it suffices to prove the corresponding locus-wise triangle inequality

$$D(\boldsymbol{s}_j, \boldsymbol{p}_j) \leq D(\boldsymbol{q}_j, \boldsymbol{p}_j) + D(\boldsymbol{s}_j, \boldsymbol{q}_j)$$

for $j = 1, \ldots, L$. But, this follows from the fact that

$$\max\left(\frac{s_{1j}}{p_{1j}}, \ldots, \frac{s_{20,j}}{p_{20,j}}\right) \leq \max\left(\frac{q_{1j}}{p_{1j}}, \ldots, \frac{q_{20,j}}{p_{20,j}}\right) \times \max\left(\frac{s_{1j}}{q_{1j}}, \ldots, \frac{s_{20,j}}{q_{20,j}}\right)$$

Appendix B

In this appendix, we will analyse the per site information $\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)} = I^{(p)}(\underline{\boldsymbol{X}}_f)/L$ of a family $\underline{\boldsymbol{X}}_f = (X_{mj})$ of M amino acid sequences of length L, where $I^{(p)}(\underline{\boldsymbol{X}}_f)$ is given by equation (2). We will assume that the actual amino acid distributions and sampling probabilities at site j are $\overline{\boldsymbol{q}}_j = (\overline{q}_{1j}, \ldots, \overline{q}_{20,j})$ and $\overline{\boldsymbol{r}}_j = (\overline{r}_{1j}, \ldots, \overline{r}_{20,j})$, respectively, with $\overline{r}_{xj}$ obtained from $\boldsymbol{p} = (p_1, \ldots, p_{20})$ and $\overline{\boldsymbol{q}}_j$ in the same way as r_{xj} was obtained from $\boldsymbol{p}$ and $\boldsymbol{q}_j$ in (1). In more detail, we assume that all amino acids $X_{1j}, \ldots, X_{Mj}$ at site j are drawn randomly and independently from $\overline{\boldsymbol{q}}_j$ so that the observed amino acid frequencies at site j have a standardised multinomial distribution

$$\boldsymbol{q}_j = (q_{1j}, \ldots, q_{20,j}) \sim \text{Mult}(M, \overline{\boldsymbol{q}}_j)/M.$$

This implies in particular that the limit of the per site information, as $M \to \infty$, is

$$\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)} = -\frac{1}{L}\sum_{j=1}^{L} \log_2(\boldsymbol{p} \cdot \overline{\boldsymbol{r}}_j), \tag{8}$$

whereas for finite M the observed per-site information will differ from (8) by a random amount of order $1/\sqrt{M}$.

In the remaining part of Appendix B, we will make this argument more precise when the protein family is randomly generated from the prior distribution, i.e. when $\overline{\boldsymbol{q}}_j = \boldsymbol{p}$ and $\overline{\boldsymbol{r}}_j = \boldsymbol{1}$ at all sites j, so that the $M \to \infty$ limit of the per site information in (7) is $\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)} = 0$. Consequently, all X_{mj} are independent and identically distributed, with a multinomial Mult(1, $\boldsymbol{p}$) distribution, whereas $\boldsymbol{p} = (p_1, \ldots, p_{20})$ is the vector of amino acid frequencies of Table 3. To this end, we will first consider a protein family of length $L = 1$. For simplicity of notation, we omit locus index j, so that $q_{x1} = q_x$ refers to the fraction of all amino acids $X_{11}, \ldots, X_{M1}$ that equal x. From, this it follows that the empirical amino acid frequencies

$$\boldsymbol{q} = (q_1, \ldots, q_{20}) \sim \text{Mult}(M, \boldsymbol{p})/M$$

have a standardised multinomial distribution. For large M, the empirical frequencies in $\boldsymbol{q}$ will approximate those in $\boldsymbol{p}$. It is convenient to introduce the random variables $\delta_x = \sqrt{M}(q_x - p_x)/p_x$, which quantify how much the empirical frequencies q_x differ from the sampling probabilities p_x on a standardised scale. When M is large, it follows from the multivariate Central Limit Theorem that approximately

$$\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{20}) \sim N(0, \Sigma),$$

where the covariance matrix of the multivariate normal distribution has diagonal elements $\Sigma_{xx} = (1 - p_x)/p_x$ and off-diagonal elements $\Sigma_{xy} = -1$. This implies that

$$\delta_x \approx \frac{Z_x}{\sqrt{p_x}} - \sum_{y=1}^{20} \sqrt{p_y} Z_y,$$

where $Z_1, \ldots, Z_{20}$ are independent standard normal random variables and $\approx$ refers to approximate equivalence in distribution. Assume that the site of interest is un-gapped. The noncensoring probabilities of (1) can be expressed as

$$r_x = \frac{1 + \delta_x/\sqrt{M}}{1 + \max(\delta_1/\sqrt{M}, \ldots, \delta_{20}/\sqrt{M})} \approx 1 - \frac{1}{\sqrt{M}}(\max(\delta_1, \ldots, \delta_{20}) - \delta_x),$$

where the last approximation is valid for large M. From this, it follows that

$$I^{(p)}(\underline{\boldsymbol{X}}_f) = -\log_2\left(\sum_{x=1}^{20} p_x r_x\right) \approx \frac{\max(\delta_1, \ldots, \delta_{20}) - \sum_{x=1}^{20} p_x \delta_x}{\log(2)\sqrt{M}} = \frac{\max(\delta_1, \ldots, \delta_{20})}{\log(2)\sqrt{M}}.$$

Expressing δ_x in terms of the standard normal variables Z_x, we find that

$$I^{(p)}(\underline{\boldsymbol{X}}_f) \approx \frac{\max\left(Z_1/\sqrt{p_1}, \ldots, Z_{20}/\sqrt{p_{20}}\right) - \sum_{x=1}^{20} \sqrt{p_x} Z_x}{\log(2)\sqrt{M}}. \tag{9}$$

Now consider a randomly generated family $\underline{\boldsymbol{X}}_f$ of M amino acid sequences of length L. Then,

$$\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)} = \frac{1}{L}\sum_{j=1}^{L} I_j,$$

where $I_1, \ldots, I_L$ are independent random variables that represent information at loci $j = 1, \ldots, L$, all having the same distribution as in (9). When L is large, it follows from the Law of Large Numbers that

$$\overline{I^{(p)}(\underline{\boldsymbol{X}}_f)} \approx \frac{C}{\sqrt{M}}, \tag{10}$$

with

$$C = \frac{E\left[\max\left(Z_1/\sqrt{p_1}, \ldots, Z_{20}/\sqrt{p_{20}}\right)\right]}{\log(2)}. \tag{11}$$

The cumulative distribution function of the random variable within the expectation of (11) is

$$F(t) = \prod_{x=1}^{20} \Phi\left(\sqrt{p_x} t\right),$$

where Φ is the standard normal distribution function. Consequently,

$$C = \frac{-\int_{-\infty}^{0} F(t)dt + \int_{0}^{\infty} [1 - F(t)]dt}{\log(2)}. \tag{12}$$

The integral of (12) can be evaluated numerically. For the amino acid distribution of Table 3, we find that $C = 14.34$. Inserting this value into (10), the number of bits per site is obtained for protein families of various size (cf. Table 7). It can be seen that the agreement between the analytical approximations of Table 7 and the corresponding simulation results of Table 5 improves the larger M is.

Appendix C

In this appendix, we will generalise (2) and define a notion $I^{(p)}(\underline{\boldsymbol{X}}_f)$ of self-information of a protein family $\underline{\boldsymbol{X}}_f = (X_{mj})$ of M aligned amino acid sequences that accounts for dependencies between sites. That is, for each protein $m = 1, \ldots, M$, we will allow for dependencies between the amino acids of the sequence $\boldsymbol{X}_m = (X_{m1}, \ldots, X_{mL})$. For amino acid $x = 1, \ldots, 20$ and locus $j = 1, \ldots, L$, we let d_{xj} refer to the number of $\{X_{mj}\}_{m=1}^{M}$ that equal x, whereas for any pair of amino acids $x, y = 1, \ldots, 20$ and locus $j = 2, \ldots, L$ we let d_{xyj} be the number of $\{(X_{m,j-1}, X_{mj})\}_{m=1}^{M}$ that equal (x, y). Let also $q_{yj} = d_{yj}/M$ and $q_{xyj} = d_{xyj}/d_{x,j-1}$ refer to the amino acid frequencies, and conditional amino acid frequencies at site j (if $d_{x,j-1} = 0$, we put $q_{xyj} = 1$).

As in Section 2, we will assume that the amino acids of $\underline{\boldsymbol{X}}_f$ are drawn with censoring from a large reservoir $\underline{\boldsymbol{X}}_R$. Random sampling from this reservoir (without censoring) corresponds to drawing

amino acids independently between proteins and sites, with amino acid distribution $\boldsymbol{p}_j = (p_{1j}, \ldots, p_{20,j})$ at site j. Randomly drawn amino acid sequences of length L (proteins), are censored independently between sequences, but not between the sites of each sequence. For simplicity of exposition, we only consider un-gapped sites j. The censoring mechanism of each amino sequence of length L is defined as a Markov process, such that the probability r_{x1} of not censoring amino acid $X_{m1} = x$ at locus $j = 1$ is given by the lower part of (1), whereas the probability of not censoring $X_{mj} = y$ at locus $j \in \{2, \ldots, L\}$, given that $X_{m,j-1} = x$, is

$$r_{xyj} = \frac{q_{xyj}/p_{yj}}{\max(q_{x1j}/p_{1j}, \ldots, q_{x,20,j}/p_{20,j})}.$$

Let $\boldsymbol{r}_1 = (r_{11}, \ldots, r_{20,1})$ contain the noncensoring probabilities at site $j = 1$ and define the vector $\boldsymbol{p}_1 \odot \boldsymbol{r}_1 = (p_{11}r_{11}, \ldots, p_{20,1}r_{20,1})$ of elementwise products of prior probabilities and noncensoring probabilities at this site. For each site $j = 2, \ldots, L$, we let $\boldsymbol{P}_j = (p_{xyj} = p_{yj})_{x,y=1}^{20}$ and $\boldsymbol{R}_j = (r_{xyj})_{x,y=1}^{20}$ be square matrices of order 20 that contain the a priori transition probabilities and noncensoring probabilities, respectively, between $j-1$ and j. Let also $\boldsymbol{P}_j \odot \boldsymbol{R}_j = (p_{yj}r_{xyj})_{x,y=1}^{20}$ be a matrix of the same size that contains the elementwise products of these two matrices. The self-information of $\underline{\boldsymbol{X}}_f$ is then minus the base 2 logarithm of the noncensoring probability of a randomly chosen amino acid sequence of length L, that is,

$$I^{(p)}(\underline{\boldsymbol{X}}_f) = -\log_2\left[\boldsymbol{p}_1 \odot \boldsymbol{r}_1\left(\prod_{j=2}^{L}\boldsymbol{P}_j \odot \boldsymbol{R}_j\right)\boldsymbol{1}\right] = D(\underline{\boldsymbol{q}}, \underline{\boldsymbol{p}}), \tag{13}$$

where $\boldsymbol{1}$ is a column vector of 20 ones, $\underline{\boldsymbol{p}} = (\boldsymbol{p}_1^T, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_L)$, $\underline{\boldsymbol{q}} = (\boldsymbol{q}_1^T, \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_L)$, and $\boldsymbol{Q}_j = (q_{xyj})_{x,y=1}^{20}$ is the matrix of transition probabilities between pairs of amino acids of sites $j-1$ and j. It can be seen that (13) reduces to (7) when the noncensoring probabilities are independent between sites ($r_{xyj} = r_{yj}$ for $j = 2, \ldots, L$). As in Appendix A, it is possible to prove that the distance measure D between the a priori Markov process $\underline{\boldsymbol{p}}$ and the observed Markov process $\underline{\boldsymbol{q}}$, satisfies the triangle inequality.

The above Markov process characterisation is only exact in the limit of many amino acid sequences ($M \to \infty$). More generally, we may assume that the M amino acid sequences of $\underline{\boldsymbol{X}}_f$ of length L are drawn independently from a Markov process with initial distribution $\bar{\boldsymbol{q}}_1 = (\bar{q}_{11}, \ldots, \bar{q}_{20,1})$ at locus $j = 1$ and transition matrices $\bar{\boldsymbol{Q}}_j = (\bar{q}_{xyj})_{x,y=1}^{20}$ between loci $j-1$ and j for $j = 2, \ldots, L$. Let $\bar{\boldsymbol{q}} = (\bar{\boldsymbol{q}}_1^T, \bar{\boldsymbol{Q}}_1, \ldots, \bar{\boldsymbol{Q}}_L)$. When $M \to \infty$, the self-information of (13) converges to $D(\bar{\boldsymbol{q}}, \underline{\boldsymbol{p}})$. In fact, (13) can be interpreted as an estimate of the limiting ($M = \infty$) self-information $D(\bar{\boldsymbol{q}}, \bar{\boldsymbol{p}})$.

Appendix D

In this appendix, we will extend formula (1) and define a rejection sampling algorithm that generates a family $\underline{\boldsymbol{X}}_f$ of M amino acid sequences of length L for more general gap scenarios than in Section 2.3. Recall from Section 2.3 that the ML elements of $\underline{\boldsymbol{X}}_f$ are sampled independently from a large reservoir of amino acids with distribution $\boldsymbol{p} = (p_1, \ldots, p_{20})$, with r_{xj} the probability of retaining a sampled amino acid x at site j. Let $0 \le g_j \le 1$ represent the proportion of sequences in $\underline{\boldsymbol{X}}_f$ with a gap at j, so that Mg_j sequences have gap at this position, whereas the other $M(1-g_j)$ sequences do not. Whereas formula (1) treats the two special cases of a site with gaps only ($g_j = 1$) or no gaps ($g_j = 0$), we will now consider the general case $0 \le g_j \le 1$. When $g_j < 1$, let $\boldsymbol{q}_j = (q_{1j}, \ldots, q_{20,j})$ represent the amino acid distribution at site j among the $M(1-g_j)$ sequences with no gap at this position, whereas $\boldsymbol{r}_j^{\text{ng}} = (r_{1j}^{\text{ng}}, \ldots, r_{20,j}^{\text{ng}})$ refers to the vector of noncensoring probabilities

$$r_{xj}^{\text{ng}} = \frac{q_{xj}/p_x}{\max(q_{1j}/p_1, \ldots, q_{20,j}/p_{20})}, \tag{14}$$

Table 7. Analytical approximation of the information per site for a randomly generated family $\underline{\boldsymbol{X}}_f$ of M amino acid sequences of length L

M	$\overline{I^{(p)}(\underline{X}_f)}$
100	1.43
1,000	0.453
10,000	0.143
100,000	0.0453

Note. The approximation is independent of L, and it is most accurate when M and L are both large.

so that the corresponding rejection sampling algorithm generates $\boldsymbol{q}_j$, with ng an acronym of no gap. The probability of retaining a randomly sampled amino acid, with noncensoring probabilities (14), is $\boldsymbol{p} \cdot \boldsymbol{r}_j^{\mathrm{ng}}$. The sought for probability of retaining a sampled amino acid x at a site j with a fraction g_j of gaps, and amino acid distribution $\boldsymbol{q}_j$ among the nongapped amino acids, is

$$r_{xj} = \left(g_j + \frac{(1-g_j)r_{xj}^{\mathrm{ng}}}{\boldsymbol{p} \cdot \boldsymbol{r}_j^{\mathrm{ng}}}\right) \Big/ \left(g_j + \frac{1-g_j}{\boldsymbol{p} \cdot \boldsymbol{r}_j^{\mathrm{ng}}}\right). \tag{15}$$

Formula (15) implies that a sampled amino acid from the reservoir represents a gap with probability

$$g_j' = \frac{g_j}{g_j + \dfrac{1-g_j}{\boldsymbol{p} \cdot \boldsymbol{r}_j^{\mathrm{ng}}}}$$

and a nongap with probability $1 - g_j'$. If a sampled amino acid x represents a gap, it is retained with probability 1 but not visible as x (it is seen as a gap), whereas if it does not represent a gap it is retained with probability r_{xj}^{ng} and visible as x. Since nongapped amino acids are retained more seldomly, when M is large the fraction of gapped and nongapped amino acids at j are close to g_j and $1 - g_j$ rather than g_j' and $1 - g_j'$. Moreover, when M is large the amino acid distribution among the M noncensored amino acids at j is close to $\boldsymbol{q}_j$.

Conflicts of interest: None declared.

Funding

No additional funding from external private sources, nor from specific research grants has been provided for completion of this work.

Data availability

The sequence data used in this work are publicly available in a Dryad repository at, https://doi.org/10.5061/dryad.h44j0zpn6.

References

Adami, C., & Nitash, C. G. (2022). Emergence of functional information from multivariate correlations. *Philosophical Transactions of the Royal Society A*, *380*(2227), 3802021025020210250. https://doi.org/10.1098/rsta.2021.0250

Atmar, W. (2001). A profoundly repeated pattern. *Bulletin of the Ecological Society of America*, *82*(3), 208–211. https://www.jstor.org/stable/20168572

Axe, D. D. (2004). Estimating the prevalence of protein sequences adopting functional enzyme folds. *Journal of Molecular Biology*, *341*(5), 1295–1315. https:// doi.org/10.1016/j.jmb.2004.06.058

Barbieri, M. (2016). What is information? *Philosophical Transactions of the Royal Society A*, *374*(2063), 20150060. https://doi.org/10.1098/rsta.2015.0060

Berger, J. O. (1993). *Statistical decision theory and Bayesian analysis, 3rd corrected printing*. Springer.
Chaitin, G. J. (1979). Toward a mathematical definition of 'life'. In R. D. Levine, & M. Tribus (Eds.), *The maximum entropy formalism* (pp. 477–498). MIT Press.
Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.
Davies, P. C. W., & Walker, S. I. (2016). The hidden simplicity of biology. *Reports on Progress in Physics*, *79*(10), 102601. https://doi.org/10.1088/0034-4885/79/10/102601
Dembski, W.A. and Marks R.J. II. (2009a). Bernoulli's principle of insufficient reason and conservation of information in computer search. In *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*, San Antonio, TX (pp. 2647–2652). https://doi.org/10.1109/ICSMC.2009.5346119
Dembski, W. A., & Marks, R. J. II (2009b). Conservation of information in search: Measuring the cost of success. *IEEE Transactions on Systems, Man and Cybernetics Part A Systems and Humans*, *39*(5), 1051–1061. https://doi.org/10.1109/TSMCA.2009.2025027
de Mul, J. (2021). The living sign. Reading noble from a biosemiotic perspective. *Biosemiotics*, *14*(1), 107–113. https://doi.org/10.1007/s12304-021-09426-y
Díaz-Pachón, D. A., & Marks, R. J. (2020). Active information requirements for fixation on the Wright-Fisher model of population genetics. *BIO-Complexity*, *2020*(4), 1–6. https://doi.org/10.5048/BIO-C.2020.4
Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.
Durston, K. K., Chiu, D. K., Wong, A. K., & Li, G. C. (2012). Statistical discovery of site inter-dependencies in sub-molecular hierarchical protein structuring. *EURASIP Journal on Bioinformatics and Systems Biology*, *8*(2012), 1–18. https://doi.org/10.1186/1687-4153-2012-8
Durston, K. K., Chiu, D. K. Y., Abel, D. L., & Trevors, J. T. (2007). Measuring the functional sequence complexity of proteins. *Theoretical Biology and Medical Modelling*, *4*(1), 47. https://doi.org/10.1186/1742-4682-4-47
Farnsworth, K. D., Lyashevska, O., & Fung, T. (2012). Functional complexity: The source of value in biodiversity. *Ecological complexity*, *11*, 46–52. https://doi.org/10.1016/j.ecocom.2012.02.001
Ferrada, E., & Wagner, A. (2010). Evolutionary innovations and the organization of protein functions in genotype space. *PLoS One*, *5*(11), e14172. https://doi.org/10.1371/journal.pone.0014172
Godfrey-Smith, P. and Sterelny, K. (2016). Biological information. In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/information-biological/. Date accessed August 21, 2021.
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
Guzzi, P. H., Mina, M., Cannataro, M., & Guerra, C. (2012). Semantic similarity analysis of protein data: Assessment with biological features and issues. *Briefings in Bioinformatics*, *13*(5), 569–585. https://doi.org/10.1093/bib/bbr066
Hartley, R. V. L. (1928). Transmission of information. *The Bell System Technical Journal*, *7*(3), 535–563. https://doi.org/10.1002/j.1538-7305.1928.tb01236.x
Hazen, R. M., Griffin, P. L., Carothers, J. M., & Szostak J. W. (2007). Functional information and the emergence of biocomplexity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(suppl_1), 8574–8581. https://doi.org/10.1073/pnas.0701744104
Hosseini, M., Pratas, D., & Pinho, A. J. (2016). A survey on data compression methods for biological sequences. *Information*, *7*(4), 56. https://doi.org/10.3390/info7040056
Hvidsten, T. R., Lægreid, A., Kryshtafovych, A., Andersson, G., Fidelis, K., & Komorowski, J. (2009). A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLoS One*, *4*(7), e6266. https://doi.org/10.1371/journal.pone.0006266
Jablonka, E. (2002). Information: Its interpretation, its inheritance and its sharing. *Philosophy of Science*, *69*(4), 578–605. https://doi.org/10.1086/344621
Jizba, P., & Korbel, J. (2020). When Shannon and Khinchin meet Shore and Johnson: Equivalence of information theory and statistical inference axiomatics. *Physical Review E*, *101*(4), 042126. https://doi.org/10.1103/PhysRevE.101.042126
Khamsi, M. A. (2015). Generalized metric spaces: A survey. *Journal of Fixed Point Theory and Applications*, *17*(3), 455–475. https://doi.org/10.1007/s11784-015-0232-5
Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*(1), 3–11. https://doi.org/10.1080/00207166808803030
Koonin, E. V. (2016). The meaning of biological information. *Philosophical Transactions of the Royal Society A*, *374*(2063), 20150065. https://doi.org/10.1098/rsta.2015.0065
Kozulic, B., & Leisola, M. (2015). Have scientists already been able to surpass the capabilities of evolution? *viXra Biochemistry*, *1504*, 0130. http://vixra.org/bioch/1504. Date accessed August 21, 2021.
Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694
Mediano, P. A. M., Rosas, F. E., Luppi, A. I., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2022). Greater than the parts: A review of the information decomposition approach to causal

emergence. *Philosophical Transactions of the Royal Society A*, *380*(2227), 20210246. https://doi.org/10.1098/rsta.2021.0246

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Nyquist, H. (1924). Certain factors affecting telegraph speed. *The Bell System Technical Journal*, *3*(2), 324–346. https://doi.org/10.1002/j.1538-7305.1924.tb01361.x

O'Connor, M. I., Pennell, M. W., Altermatt, F., Matthews, B., Melián, C. J., & Gonzalez, A. (2019). Principles of ecology revisited: Integrating information and ecological theories for a more unified science. *Frontiers in Ecology and Evolution*, *7*, 219. https://doi.org/10.3389/fevo.2019.00219

Popa, O., Oldenburg, E., & Ebenhöh, O. (2020). From sequence to information. *Philosophical Transactions Royal Society B*, *375*(1814), 20190448. https://doi.org/10.1098/rstb.2019.0448

Povolotskaya, I., & Kondrashov, F. (2010). Sequence space and the ongoing expansion of the protein universe. *Nature*, *465*(7300), 922–926. https://doi.org/10.1038/nature09105

Pratas, D., & Pinho, A. J. (2017) On the approximation of the Kolmogorov complexity for DNA sequences. In L. Alexandre, J. Salvador Sánchez, & J. Rodrigues (Eds.), *Pattern recognition and image analysis. IbPRIA 2017. Lecture notes in computer science* (pp. 259–266). Springer. https://doi.org/10.1007/978-3-319-58838-4_29

Schneider, T. D. (2006). Claude Shannon: Biologist. The founder of information theory used biology to formulate the channel capacity. *IEEE Engineering in Medicine and Biology Magazine*, *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 Research*, *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*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, *49*(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

Smith, B. A., & Hecht, M. H. (2011). Novel proteins: From fold to function. *Current Opinion in Chemical Biology*, *15*(3), 421–426. https://doi.org/10.1016/j.cbpa.2011.03.006

Soler-Toscano, F., Zenil, H., Delahaye, J.-P., & Gauvrit, N. (2014). Calculating Kolmogorov complexity from the output frequency distributions of small turing machines. *PLoS One*, *9*(5), e96223. https://doi.org/10.1371/journal.pone.0096223

Sousounis, K., Haney, C. E., Cao, J., Sunchu, B., & Tsonis, P. A. (2012). Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Human Genomics*, *6*(1), 10. https://doi.org/10.1186/1479-7364-6-10

Szostak, J. (2003). Functional information: Molecular messages. *Nature*, *423*(6941), 689. https://doi.org/10.1038/423689a

Taylor, S. V., Walter, K. U., Kast, P., & Hilvert, D. (2001). Searching sequence space for protein catalysts. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10596–10601. https://doi.org/10.1073/pnas.191159298

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

Walker, S. I., & Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of the Royal Society Interface*, *10*(79), 20120869. https://doi.org/10.1098/rsif.2012.0869

Wang, Y., Zhang, H., Zhong, H., & Xue, Z. (2021). Protein domain identification methods and online resources. *Computational and Structural Biotechnology Journal*, *19*, 1145–1153. https://doi.org/10.1016/j.csbj.2021.01.041

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

Yockey, H. P. (1977). On the information content of cytochrome. *Journal of Theoretical Biology*, *67*(3), 345–376. https://doi.org/10.1016/0022-5193(77)90043-1