Synthese (2024) 204:15
https://doi.org/10.1007/s11229-024-04613-z

ORIGINAL RESEARCH

The representation, quantification, and nature of genetic information

Steinar Thorvaldsen[1] · **Peter Øhrstrøm**[2] · **Ola Hössjer**[3]

Received: 17 December 2022 / Accepted: 30 April 2024 / Published online: 27 June 2024

Abstract

Current genetics studies often refer to notions from information science. The purpose of this paper is to summarize and structure the different notions of information used in biology, as a step towards developing a taxonomy of information. Within this framework we propose an extension of Floridi's conceptual model of information. We also make use of the concept of specified information and show that functional information and many other notions of information are either special cases of, or are closely related to, specified information. Since functionality of the proteins that genes code serves as an external and independent specification, this makes it possible to define genetic information in a way that includes semantic aspects. In particular, we discuss how to understand the qualitative aspects of genetic information, how to measure its quantitative aspects, and how variants of Shannon's information measure can be applied to molecular sequence data of protein families. While a mathematical framework may not be able to incorporate all that is included within biological information, some aspects of it allow for statistical modelling. This is especially true if we restrict our focus on the discipline of genetics. The concept of genetic information is still disputed because it attributes semantic traits to what seems to be regular biochemical entities. Some researchers maintain that the use of information in biology is just metaphorical and may even be misleading. We argue that the foundation of the metaphorical view is relatively weak given the current findings in bioinformatics and show that the present understanding of genetics fits well into the context of the modern philosophy of information. The paper concludes that informational concepts have robust scientific applications at the level of genes.

✉ Steinar Thorvaldsen
steinar.thorvaldsen@uit.no

1 Department of Education, Division of Science, UiT the Arctic University of Norway, Tromsø, Norway

2 Department of Communication and Psychology, Aalborg University, Aalborg, Denmark

3 Department of Mathematics, Division of Mathematical Statistics, Stockholm University, Stockholm, Sweden

Keywords Floridi · Self-information · Natural information · Algorithms · Instructional information · Functional information · Specified information

1 Introduction

1.1 The use of information in biology and genetics

The concept of information is important within a number of scientific disciplines. The founder of cybernetics, Norbert Wiener, even argued that information was something more essential than matter and energy. He put it in this way: 'Information is information, not matter or energy. No materialism which does not admit this can survive at the present day' (Wiener, 1961, p. 132). The short slogan 'It from Bit' by physicist Wheeler (1990) also points out that the ultimate physical reality (It) is information-based (Bit).

In particular, it is well known that the notion of information is relevant in biology. The analysis of tree-rings is an interesting example. The study of the patterns of these concentric circles may be said to follow certain syntactic rules, and the semantic interpretation of them may give rise to factual information not only about the ages of the trees but also about the weather during their history. In this paper we shall concentrate on the genetics of life, and we shall show that the usefulness and relevance of information studies in biology is even more obvious when we focus on genetics.

One of the rather surprising discoveries in biology is that genetic information is organised in a way that resembles conventional text, while the cell operates in a way that resembles modern technology. Life displays a deep unity at the biochemical level, despite its huge diversity expressed at the morphological level. With a few exceptions, all forms of life use DNA as their genetic material, and proteins are constructed from the same 20 amino acids as their building blocks. RNA bridges the two worlds through the genetic code, where one symbolic language translates into another with the aid of a sophisticated universal apparatus. Bioinformatics has unlocked informational aspects of molecular biology through the use of computers and statistics. Terms such as 'code', 'language', 'transcription', 'messenger', 'information', 'library' and 'motor' have proved useful in describing and understanding biology at its molecular level. Notions of information are embedded within the description of the cell machinery. Bioinformatics is the application of mathematical and informational techniques to perform biological exploration, usually by developing computer programs, mathematical models, or both. In this way, life is treated as a close partnership between genes and mathematics (Stewart, 1999). One major aspect of bioinformatics is data mining and the analysis of data assembled by numerous genome projects. As an interdisciplinary domain of study, bioinformatics splices biology, computer science, information engineering, statistics, and mathematics to study and interpret these biological data.

Historically, the concept of information has been a central issue within biology since the discovery of the coding structure of DNA in 1953 by Francis Crick and James Watson, and the idea that DNA serves as a computer program is not new. John von Neumann was among the first who attempted to formalise this (von Neumann, 1961), and other theoretical researchers followed (Chaitin, 1979; Yockey, 1974). Von Neumann even drew much of his inspiration from genetics when designing his famous

von Neumann architecture for an electronic digital computer (von Neumann et al., 1987).

In this way, information has become a major notion of present biology, and there is a common understanding that the informational aspect of life is a key property. Some researchers have even suggested that it might be seen as the master key property (Godfrey-Smith & Sterelny, 2016; Walker & Davies, 2013). If so, life should be studied as fundamentally related to information processing and communication. Such an analysis of life's informational properties and contents holds the potential for turning biology into a more quantitative science (Davies & Walker, 2016).

After the discovery of the genetic code (Crick, 1958), it is clear that most notions of genetic information can either be formulated in terms of nucleic acid sequences or amino acid (protein) sequences. In particular, amino acid sequence data often provide thousands of examples of similar, but different, protein sequences that convey the same meaning, in the sense that they encode essentially the same structure and function. A *gene family* is a group of closely related genes that encode similar products, usually proteins, but also RNA. More radical are examples of protein molecules with entirely different sequences and structures but similar biochemical functions. Thus, different molecular structures may be functionally equivalent. Such examples pose crucial questions regarding the nature of the information contained in genetic sequences. How can we best define and quantify the information content of protein sequences, when there is no one-to-one correspondence between these sequences and their function/meaning? There is still a great deal of open conceptual space and much room for new accounts of biological information.

1.2 Aim and contents of the article

This paper aims to present a promising new line of enquiry to bring coherence to the domain of biology by focusing on *information* as a unifying and computable concept. We make use of the scientific literature from statistics, molecular biology, cybernetics, and biosemiotics in order to seek a common theme across diverse fields. Our synthesis is intended to provide some common conceptual ground for further scientific exploration of the role of information in genetic systems, with a particular focus on gene families. Our hope is that such a synthesis is a step towards formulating a taxonomy of information for genetics and molecular biology. A crucial part of our analysis is the role of semantics and meaning for information of biological systems. Some researchers maintain that the use of information in biology is just metaphorical and may even be misleading, whereas others argue that the role of information in biology is more profound. This discussion is possibly one of the deep issues in modern biology.

Our article naturally splits into three parts. In the first part (Sects. 2 and 3) we analyse and try to structure the different meanings of information. In Sect. 2 we review some major concepts of information theory, with specified information a core concept. Then in Sect. 3 we present Floridi's conceptual model of information (Floridi, 2010) and explain how this model unifies several of the information theoretic notions of Sect. 2. In the second part (Sect. 4 and Appendix A) we present methods of quantifying

information of gene families in ways that incorporate semantic aspects of meaning, based on the foundation laid in Sects. 2–3. Finally, in the third part (Sect. 5) we provide a discussion with concluding remarks. In particular, in this discussion we argue that genetic information is not only a metaphor, but that it is a genuine, essential, and non-material entity essential for life, a type of necessary initial condition for life to exist.

2 Current concepts of information

The relevance of information in genetics and biology, in general, depends on an understanding of the notion of information itself. Information is a multifaceted concept that is relevant for understanding a diverse set of features, such as the degree of organisation of a system or the contents of a string of letters. It is not a simple task to provide a brief account of this crucial notion in modern science. Various ways to capture information have been suggested in the literature. Despite being a central concept across science, it is employed differently across disciplines. Matter and energy are modelled and well-studied in physics, but there is no universal model of information. It may even be inherently impossible to give a general definition of information, given the uncontained use of the term. Although information must be instantiated into physicality for storage or communication, it is generally accepted as a non-material entity with many attributes, both qualitative and quantitative. The philosophy of information (e.g. Floridi, 2002, 2010, 2016) is a general area of research investigating the conceptual nature and basic principles of information. In this section we review a number of notions of information, before relating them to Floridi's conceptual model of information in Sect. 3.

2.1 Ontological, epistemic and practical information

Three types of information can be distinguished, based on the way in which it relates to reality (Borgmann, 1999; Floridi, 2002; Osimani, 2014), and all of them are of relevance for biology. Ontological information is information as reality (such as the actual structure of a DNA sequence), epistemic information is information about reality (such as an agent, with a mind, having knowledge about the DNA sequence), whereas practical information is information for reality (such as the DNA sequence corresponding to instructions, that are either successfully transcribed and translated into a functional protein, or not).

Epistemic information corresponds to knowledge, and as such it can be further divided into three types (Pavese, 2021); acquaintance knowledge (to get to know persons), knowledge how (to learn certain skills), and knowledge that (to learn the truth value of propositions or facts). Whereas practical information concerns the end product of a series of instructions, knowledge how is a concept that rather focuses on an agent (with a mind) being able to perform such instructions. With this definition, if an enzyme (without a mind) carries out biological instructions successfully, this does

still not qualify as knowledge how. In Sect. 3 we will use the word know-how (rather than knowledge how) to characterize such effective algorithms.

Knowledge that is closely related to justified, true belief. Foundationalism and coherentism are normative theories for how beliefs are justified. Foundationalists argue that there are self-evident basic beliefs that need no justification, whereas all other beliefs must be grounded in basic beliefs in order to be justified. Coherentists argue that a belief is justified if it coheres with other beliefs. In any case, the grounding or coherence can be internal to the agent having the belief, or external. Regardless of how the justification part of knowledge that is defined, an agent is more motivated to acquire knowledge that about a proposition, if it carries some meaning (Sect. 2.2) to him.

2.2 Quantitative and qualitative/semantic information

Claude Shannon's well-known information theory relies on probability theory, as his model only considers the *statistical* properties of the symbols that form messages (Shannon, 1948). Details on Shannon's information theory and entropy can be found in the literature (Yockey, 2005). Here, we briefly emphasise the application of this concept in molecular biology. In a sequence of characters, the classical Shannon measure of information is purely a function of the probabilities of the character string: a quantitative theory without a semantic dimension (see Sect. 4.1).

However, Shannon recognised that his theory of information was not the last word in the mathematical modelling of information. Shannon's model of information theory was framed to address the problem of communication. Depending on the field of application, we must choose different approaches to information. In particular, the present paper is framed to address some of the challenges within genetics.

In addition to the quantitative ideas of information suggested by Shannon and Weaver (1949), there is a vital qualitative notion of information going back to Carnap (1947), who suggested the use of modal logic to understand these qualitative aspects of information. Warren Weaver acknowledged: 'In fact, two messages, one of which is heavily loaded with meaning and the other of which is pure nonsense, can be exactly equivalent, from the present viewpoint, as regards information' (Shannon & Weaver, 1949, p. 8). This is clearly the case if we concentrate on the numerical and engineering aspects of the communication. However, we may also choose to focus on the content and qualitative aspects. Although modern biology has frequently made use of the concept of information, it has to some extent avoided these qualitative concepts of meaning and semantics. But since the *semantic* concepts of information can be established in several ways (Floridi, 2015; Lopez-Ruiz, 2005; Weinberger, 2002; Atlan & Koppel, 1990), there is plenty of room for defining qualitative aspects of information within biology.

We will find in Sect. 4 that a semantic approach to information does not exclude a numerical approach. In fact, the modern notion of information should be conceived as a combination of these two aspects—statistics and semantic—that together capture some of the double nature of information. This dual nature of Shannon and semantic information is a basic dichotomy in the literature. In spite of this, there are common

features of quantitative and qualitative information, such as a tendency to protect itself. Indeed, quantitative, digital information is typically robust against perturbations and allows for accurate error correction, and the same is often true for qualitative information.

2.3 Information and elimination of possibilities

According to Fred Dretske (1981, p. 4), 'the amount of information associated with, or generated by, the occurrence of an event (or the realisation of a state of affairs)' should be identified with 'the reduction in uncertainty, the elimination of possibilities, represented by that event or state of affairs'. This idea of defining information in terms of the elimination of possibilities is the basis of the quantitative concepts of self-information and Shannon information (Sect. 4.1) and the use of a language, with a certain syntax, in order to define information. In this way, it is the syntax that allows some strings of letters and eliminates others. A further (and more qualitative) elimination of possibilities, among all syntactically valid text strings, occurs when only some of these syntactically valid text strings convey meaning (Gitt, 1989). Elimination of possibilities can also be used as a qualitative tool of information through the concept of possible worlds semantics (Floridi, 2015; Martinez & Sequoiah-Grayson, 2018), giving rise to a very attractive approach to the notion of semantic information. William A. Dembski stated that 'the ultimate act of information must then consist in separating out the actual world from among all possible worlds' (1998, chap. 4). Basically, in order to qualify as a definition of semantic information, we will require that the abovementioned rule for separating the true world from the other worlds is based on some type of code (Sect. 2.4), as well as some type of specification (Sect. 2.7) that provides meaning.

2.4 Information and its relation to codes, causation and syntax

An important example of the relevance of meaning in molecular biology and genetics, is its close connection to the idea of a code. This was predicted already by Erwin Schrödinger in his popular book "What is life?" (Schrödinger, 1945), and further highlighted by Francis Crick when he discovered the Central Dogma of Biology:

> Once information has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but the transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein (Crick, 1958, p. 153).

Stating that a code establishes correspondence between two objects (e.g., triplets of nucleic acids and amino acids) is equivalent to stating that one object is the meaning of the other, internally within the coded system. In this restricted sense of meaning, we cannot have codes without meaning or meaning without codes. When the code is between mental objects, meaning is a mental entity that is related to epistemic

information (Sect. 2.1). However, meaning is a genetic entity when the code is between genetic molecules. Similarly, semantic information has intentionality about something; it is directed to other things. Genes carry semantic information if and only if they are interpreted as such.

The concept of code, and its established correspondence between two objects, is closely related to causation. Predrag Sustar reduces the notion of genetic information to causation (Sustar, 2007), whereas Barbara Osimani defines genetic information as a special kind of cause, which causes something to be one way rather than another, by combining elementary units one way rather than another (Osimani, 2014).

The 'precise determination of sequence' quote from Crick's 1958 article refers to the way in which the nucleic acids or amino acids are arranged. This emphasizes the importance of language, and the syntactic nature of information (which we refer to as syntactic information). Much effort has been made to build up an integrated view of genetic information and its role in biology, incorporating such a syntactic nature of genetic information, contained in the coded DNA, and the semantic quality of processes by which genes specify biological forms and functions (i.e., biosemiotics) (Hoffmeyer, 2008; Jablonka, 2002; Maynard-Smith, 2000; Wills, 2016). Maynard-Smith maintained that bioinformation is both specific and intentional, and that genes are meaningful in the same way that words are in a language. Wills argues that the processes inside cells are under computational control of genetic programming, and Barbieri (2016) refers to the idea that 'life is chemistry plus information plus codes' as the *code paradigm*.

2.5 Natural information

The work of Dretske (1981) was among the first systematic and influential theory of semantic information. It builds directly on the technical-theoretical resources coming from Shannon's non-semantic information theory. The crucial idea was that the state of one system could carry information about the state of another, and that this information could provide a basis for the meaning of the signals mediating the sender–receiver relation. This kind of information is commonly known as 'natural information' and it is defined as a mind-independent, lawful relation between systems or events in the world, e.g., between fire (F) and smoke (s). According to Dretske (1981, p. 65), the information that one system carries about another can be expressed in probabilistic terms: A received signal r carries the information that s is F if the conditional probability of s being F, given r and background knowledge k, is 1; whereas the conditional probability of s being F, given only k, is less than 1.

This analysis contains two conditional probabilities. First the probability of s being F given both the signal r and background knowledge k; and second the probability of s being F given only the background knowledge k, in the absence of signal r. Hence, r carries the information that s is F, when the first conditional probability is one and the second is less than one.

Several authors have critically discussed Dretske's theory. His view has two notable shortcomings, one being that the requirement that signal r must raise probabilities

to unity as too strict (e.g., Millikan, 2000; Scarantino, 2015), and the other being that it fails to provide a way of determining the background knowledge k, that is, the reference classes of information-carrying states. The strictness problem can be relaxed by modifying Dretske's view to incorporate more nuanced notions of correlations, probabilities and initial conditions to maintain that imperfectly related events can carry natural information, and then aim to specify the conditions under which they do (Baker, 2021; Stegmann, 2015). An updated version of natural information can be stated as follows:

A received signal r being in state G carries information about the source s being in state F, relative to background knowledge k, if and only if

$$P(s \text{ is } F | r \text{ is } G \& k) > P(s \text{ is } F | k).$$

The transmission of positive information amounts to the presence of a positive correlation between the received signal and what the signal is about (relative to reference knowledge k). A more general definition of natural information is given in Appendix A.3.

Baker (2021) also states that the problem of identifying the relevant background knowledge is non-trivial, as it requires resources from outside of information theory. This means that we have to recognize a basic limit on what information-theoretic tools can do to tell us something about the content of intentional processes, since information theory by itself is useless in determining the semantic content in the understanding of biological features. Natural information is still an indispensable idea for much scientific work and it plays a crucial role for analyses that try to shed light on the way in which physical systems can exhibit intentional processes. Both philosophers and philosophically minded biologists have contributed to the ongoing foundational discussion on the ontological status of natural information, and its mode of description in biology. For recent overviews see Floridi (2016, chap. 22–23) and Godfrey-Smith and Sterelny (2016). We will return to this question in the discussion part of our paper.

Systems with codes and causation (Sect. 2.4) generate correlations between entities that are closely related to natural information, with a distinction that they need not refer to mindless, lawlike procedures. An invertible code generates a sender-transmitter relation that corresponds to Dretske's original definition of natural information, whereby the left-hand side of the above equation is 1. That is, if an invertible code translates s to r, a receiver that observes $r = G$ without error knows with certainty that F was translated (and not any other value of s). Note however that the genetic code is not invertible, since several codons code for the same amino acid. It still corresponds to natural information in the wider sense. Indeed, a received amino acid $r = G$ originates from a particular codon F that translates to G, with a probability that equals one divided by the number of codons that code for G. This probability, which corresponds to the left-hand side of the above equation, is larger than the probability that a randomly chosen codon equals F, that is, the right-hand side of the equation.

2.6 Instructive, blueprint, and hereditary information

Living organisms are unique because they are suitably structured and arranged in precise forms. All these structures, their composition and configuration, are at the centre of developmental and evolutionary biology: 'Developmental biology can be seen as the study of how information in the genome is translated into adult structure, and evolutionary biology of how the information came to be there in the first place' (Szathmáry & Maynard Smith, 1995). An obvious question is what the nature and characteristics of this extensive use of the concept of information are. According to Shea (2013), some aspects of an organism's development can be seen as reading information carried by its genome.

Biologists also recognise the crucial importance and usefulness of information notions in the account of life and its origin. Nobel Prize winner and origin-of-life researcher Manfred Eigen equated the problem of life's origin with uncovering 'the origin of information' (Eigen, 1992). This relates to a common understanding of genetic information as a blueprint (an Aristotelian formal cause) for making the molecular elements that are responsible for the complexity and functionality at all levels of life, from DNA to the whole organism. Such a blueprint definition of genetic information is based on the fact that DNA is expressed in various ways. Griffiths (2017) has argued that it is common policy to consider the characteristics of chromosomes and genes as the expression and transmission of information, and he emphasises the prevailing challenge of capitalising on this in strict, scientific terms.

Much of modern genetics is a science of hereditary information, and the survival of each species is discussed in terms of transmission of genetic material from generation to generation in the unique informational narrative of living systems. For instance, Monod (1971) identified hereditary information with the structural morphology of organisms that are reproduced from one generation to the next.

These instructive, blueprint and hereditary notions of information are closely related in the sense that genomic information is transferred, either to a developmental structure, to molecular elements of a cell, or to genomic information of the next generation. When quantified, these three concepts of information are conceptually related to natural information (Sect. 2.5) where genomic information s is transferred to some structure r (either developmental, molecular element or next generation genomics). Such a gene-centred approach has been criticized though by Developmental Systems Theorists, who argue that environmental factors should also be considered as part of the causal determinants s (Griffiths, 2001). Adherents of teleosemantics have similarly argued that the gene-centred approach is too narrow, and that evolutionary causes of the genome should be included in the information concept as well (Bergström & Rosvall, 2011; Shea, 2013). Notice however that it is possible to use the analogy with natural information in order to combine environmental and/or evolutionary factors with a gene-centred view. Indeed, it is possible that these environmental/evolutionary factors lower the probability that s equals F, given the observed structure r and background knowledge k. As long as this probability remains larger than the conditional probability that s equals F, given only background knowledge k, it is still the case that r carries

information about s equals F. Alternatively, environmental/evolutionary factors can be put into the background knowledge k, in the definition of natural information.

2.7 Specified and functional information

A structure is *specified* if it involves some events or feature that cannot be defined only in terms of the units that make up the feature. Such a feature represents specified information (Dembski, 1998), for instance some function that the structure possesses or is able to perform. The feature is independent of the structure itself, in the sense that it corresponds to an external property or pattern of functional expression.

An important aspect of information within the cell is the fact that proteins exhibit such specificity, both 1D arrays and 3D geometries. The folding of proteins with their specific 3D shapes requires highly specific amino acid sequences. Within the set of possible sequences, only a very few will produce a set of functional and cooperative proteins in the cell (Axe, 2004; Bowie & Sauer, 1989; Tokuriki & Tawfik, 2009). Since the physicochemical properties of the amino acids allow a huge set of combinatorically possible arrangements, any particular sequence will necessarily be very improbable and rich in information load. These sequences are not only improbable but also functionally specific. The small set of functionally effective sequences reduces the larger set of possible combinations. Furthermore, this smaller set establishes an independent feature because it divides functional sequences from non-functional sequences. Hence, any actual amino acid sequence that meets such requirements is both highly improbable and specified about that independent new feature. Accordingly, the coding protein sequence possesses both *syntactic* and *specified* information (Meyer, 2003, p. 237).

Furthermore, the coding sequences are highly specific to the overall functional requirements of the cellular and intracellular networks. The cell transmits resources back and forth through its membranes, controls metabolism, and performs many other specific tasks. Each of these functional requirements, in turn, needs specific molecular elements, molecular machines (mostly made of proteins), and logistics systems to be realized.

The question arises whether specified information of a structure is sufficient for this structure having semantic information (Sect. 2.2) as well. If 'semantic information' is defined as 'subjectively meaningful information that is expressed syntactically as a string of characters and is understood by a conscious, epistemic agent' (Sect. 2.1), then clearly only real language carries conscious meaning. By this definition, it may seem at first that the information in proteins does not qualify as semantic information. But proteins function similarly as a software algorithm (Sect. 2.9), instructing effective processes within a complex material system via complex yet highly specified strings. In the same way as the precise sequencing of two bits (0 and 1) in a software procedure can perform a function within a technical environment, so too can the specific pattern of the 20 amino acids perform a *function* within the cell. Genetic information therefore uncovers its meaning not only through codes, causation and syntax (Sect. 2.4), but also through instruction or the actual production of formal biofunctions (which a conscious agent observes and recognizes), and this is a central issue penetrating much of biology (Newman, 2022). Similar to software and machine codes, the sequence specificity of

proteins occurs within the syntactic domain of an operative amino acid string. Thus, proteins possess both syntactic and specified information and, in this way, they carry meaning. In Fig. 1, we denote this process as an *algorithm* (cf. Sect. 2.9), since an algorithm has a syntax, and it is typically also functionally effective (the specification).

Twenty years ago, Jack Szostak published a paper in *Nature* that paved an important notion of specific information. Szostak argued that the meaning or functionality of a message is essential in molecular biology (Szostak, 2003). Since conventional information theory does not distinguish between functionality and non-functionality, Szostak pointed out the need for a new measure of information, which he called *functional information*. Together with his colleagues he introduced *functional information* in terms of a finite gene string as $-\log_2$ of the fraction of functional sequences that have fitness values (activity of a biopolymer) greater than a given value (Hazen et al., 2007). Szostak's definition was motivated by imagining a conic pile of protein molecules of all possible sequences sorted by a certain activity with the most active at the top. A horizontal plane across the conic pile signifies a given level of activity. As the plane gets higher, fewer sequences remain above it. The functional information needed to quantify that activity is $-\log_2$ of the fraction of sequences above the plane. This provides an immediate and quantitative measure of the difficulty of a task. More functional information is involved to specify molecules that perform more complicated tasks.

Functional information in the sense of Hazen et al. (2007) can be seen as a special case of specified information, where fitness (biopolymer activity) is used to specify a protein. This observation suggests that specified information of an observed object can also be defined in other contexts than functionality. Suppose we are dealing with some type of specification (such as function, degree of organization, or algorithmic complexity) that can be quantified. As shown in Appendix Appendix A, it is possible then to give a quantitative definition of specified information as $-\log_2$ of the fraction of objects that are at least as specified as the observed object's specification. An instance of such a specification is Werner Gitt's five levels of information (statistics, syntax, semantics, action, purpose) assigned to written text (Gitt, 1989). If these five levels are coded as 0, 1, 2, 3, 4, the numbers serve to quantify the degree of specification of a text. As shown in Appendix A.1, the higher the information level of a text is, the more specified information it has.

In Sects. 2.8 and 2.9 we will treat cybernetics and algorithms as two other special cases of specified information (cf. Appendix A.1).

2.8 Cybernetics

Cybernetics is the study of systems with circular causal processes, such as feedback loops (Wiener, 1948). In biology, feedback loops are important. Negative feedback loops (such as the regulation of body temperature) allow systems to remain in homeostasis. An example of a positive feedback occurs at the onset of contractions in childbirth. When contraction occurs, oxytocin is released into the body stimulating more contractions. On the molecular level, positive and negative feedback loops occur to decrease or increase the expression of genes. Since a circular causal process is

a specification of a system, cybernetic information can be seen as a special case of specified information (see Appendix A.1).

Suppose we widen the definition of cybernetics to the degree of organization of a system. Then Eigen's statement that explaining the origin of life boils down to explaining the origin of information, can be interpreted as explaining cybernetic information of a cell or an organism, in terms of a genetic blueprint information (cf. Sect. 2.6).

2.9 Algorithms and algorithmic information

The concept of algorithm has existed since antiquity (Chabert, 2012). An algorithm is an effective procedure, a way of performing something in a limited number of stages. Originally introduced in mathematics, it denotes a process leading to future utility that terminates after a finite number of steps. There are many detailed definitions for describing such algorithms. We restrict the definitions to those that most closely refer to algorithms used in information science. Therefore, an algorithm is understood as a set of steps or procedures that precisely define a finite sequence of operations (D'Onofrio et al., 2012). Starting from an initial state, the instructions of the algorithm describe a stepwise process that, when executed, proceeds through a finite number of successive states, eventually terminating at the final ending state. Such algorithms can be expressed with many types of notation, including natural languages, flowcharts, pseudocodes or programming languages. If the end state of the algorithm is seen as its goal or specification, we will show in Sect. 4.2 and Appendix A.1 that Kolmogorov algorithmic information is a special case of specified information, as defined in Sect. 2.7. This notion of algorithmic information can also be seen as a special case of practical information (Sect. 2.1), when the end product of a series of instructions is effective (achieves its goal).

This understanding of algorithms is defined from a computer science perspective, owing to the discrete nature of genetic systems and operations. Life is largely governed by algorithmic processes, much in the same way as linear digital programs, that is, a sequential string represented by command characters. Digital here refers to something discrete and definite. A simple computer program, for example, is directed by such a linear digital string of well-chosen binary commands represented by either '1' or '0'. The sequencing or syntax of these purposeful commands supports a growing functional hierarchy. An example of a highly complicated algorithm is the 3D folding of protein sequences. We already found Sect. 2.7 that fold itself is specified, since only a small fraction of amino acid sequences will generate functional proteins. The folding process is only based on the underlying physical principles of the amino acids in the sequence and the chemical context of the cell. Prediction of protein folds from the knowledge of the amino acid sequence remains an important challenge in the post-genomic era. This requires understanding of the folding pathway. The algorithm that folds proteins each and every time is the algorithm that nature employs. Equal amino acid sequences normally fold into the same 3D structure. Folding is rapid (within milliseconds to seconds). The pathway is also crucial, as some mutants that are stable in the native state will not fold because the folding pathway is blocked by the mutation. Protein

structure prediction has been studied using several algorithmic approaches (Hutson, 2019).

Moreover, genes and proteins are different from inorganic molecules not only because they have different functions and structures but also because they are initially generated in an entirely different way. Inorganic molecules are produced by self-assembly, and their shapes are determined by internal factors. Instead, genes and proteins are produced using molecular machines that physically combine their subunits with external templates. This means that genes and proteins are built based on external instructions, making them very different from conventional molecules. Thus, genes and proteins contain algorithmic information whereas inorganic molecules do not. Indeed, there seems to be no such algorithms with external instructions in the dead nature (Penrose, 1989), while genetics is filled with these kinds of algorithms. Life is a system with its own inherent set of biological instructions and algorithmic processing of information. Hubert Yockey made the following observation: 'There is nothing in the physico-chemical world that remotely resembles reactions being determined by a sequence and codes between sequences. The existence of a genome and the genetic code divides living organisms from non-living matter' (Yockey, 2000). In contrast to cells, chemical systems are not considered as being able to process information. Biological informational architectures distinguish them from other complex physical systems that do not display the same informational attributes. Life exhibits formalisms that cannot be generated or explained by physicodynamics alone, including its numerous biofunctional goals of staying alive (Abel, 2012). Matter, energy, cause-and-effect determinism, and the positive and negative feedback mechanisms of nature's order cannot foster formalisms such as a language with symbol systems, coding, decoding, logic, organisation (not to be confused with mere self-ordering), and succeeding functionality. These invariant markers of life are formal, not physical.

In 2016, the Royal Society of England published a special issue on regarding DNA as information (Wills, 2016). In these papers, it is emphasised that the biological DNA code not only transmits information but also translates it, and that the genetically encoded information even includes how to produce new transmitters and receivers. This special issue also mentions the programming aspects of DNA, which is not just a linear instruction but a program of algorithms, subroutines, feedback loops, and all the complexities that this entails. Genes are dynamic algorithms that, together with the environment in the cell, contribute to the growth, development, and control of the organism. Biological information can be used to specify molecular systems right down to the atomic level.

It is therefore reasonable to give the concept of instructional or algorithmic information a fundamental role in genetics as information can be seen as a basic concept of the general understanding of life, in line with fundamental concepts such as matter, time, and space. This is a well-motivated approach for a deeper understanding of genetics.

2.10 Active information

A passive perspective on genetic information is incomplete, as genetic information is *active*, not merely passive facts. If a structure has features brought in by an external agent as exogenous information, we say that the structure possesses *active information* (Dembski & Marks, 2009a, 2009b).

The concept of active information was originally introduced in the context of computer search, but we argue it has bearing on semantic information in biology as well, in order to discuss the cause of information. Recall from Sect. 2.7 that semantics requires more than the internal/syntactic definition meaning of Sect. 2.4, between objects that are connected through a code. It also requires structures that are meaningful, for instance an effective algorithm. This meaning is typically some external specification (Sect. 2.7). If an external agent actively *causes* this external specification, the semantic information is also active information. In Appendix A.3 we illustrate active information in the context of Werner Gitt's five hierarchical levels of information (Gitt, 1989), defined in Sect. 2.7. Semantic information is the third level, whereas the two highest levels originate from an active and conscious sender. The latter two levels typically correspond to a positive active information, since the sender desires some action of the receiver or has some purpose with the sent message. However, it is important to note that the concept of active information is applicable whether the agent is conscious or some lawlike behaviour brought in from outside. It is described in Appendix A.3 how active information may be associated with natural information (Sect. 2.5).

2.11 Summary of information concepts

To summarize Sect. 2.1–2.10, we have found that Shannon's concept of information relies on statistical properties alone, syntactic information additionally requires data that obeys some syntax, and semantic information (on top of syntax) additionally requires some type of meaning.

For many notions of information, one structure carries information about another (the two structures are correlated). For instance, a received signal carries natural information about the state of a system, if it makes this system more likely, given the background knowledge that is at hand (Sect. 2.5). Mathematically this is closely related to the information that one object brings about another, given that a code between them exists (Sect. 2.4). The notions of instructive, blueprint or hereditary information (Sect. 2.6) are related to natural information as well. In order for these notions of information to convey more than internal meaning within the system (Sect. 2.4), there must be some external specification of the two correlated structures. That is, semantic information (Sects. 2.2 and 2.7) not only requires syntactic information, and/or correlation between structures, but also specified information (Sect. 2.7), such as objects of a coded system that produce some effect. If the system can be recognized to carry meaning by an external agent (the receiver, cf. Section 2.1), such as intentionality or functionality, the semantic information is epistemic as well. If the specified feature of a structure is actively brought in from outside by an agent (the sender), the structure also has active information (Sect. 2.10).

Although meaning is mainly a qualitative aspect, it can sometimes be quantified as degree of specification. If so, it is possible to compute specified information (Appendix A) for a structure conveying meaning. Functional information, for instance, is a special case of this notion of specified information, with degree of functionality (reaction rate) used as specification.

3 Floridi's conceptual model

Semantic information was discussed in Sects. 2.2 and 2.7. When we focus on semantic information, we may consider the different types and distinctions that take part in the complicated interaction and flow of information. Luciano Floridi, at Yale University and the University of Bologna, is one of the most well-known information theorists. He has proposed a model for this interaction, as illustrated in Fig. 1 and Table 1.

Floridi clearly distinguished between *data* and *well-formed data* in his model. The idea is that data can be utterly meaningless signals and they will not be well-formed unless they meet certain syntactic requirements (Sect. 2.4). In addition, data becomes *semantic information* when there is also a code used, and an external specification, to make them *meaningful*, either as *instructional information* (i.e., instructions on how to do something, cf. Sect. 2.9) or *factual information* (i.e., ontologically how something is in reality, cf. Sect. 2.1). Furthermore, it appears from the model that the factual information is either true or false. Factual semantic content is the most common way to understand information and it is also one of the most important, because information as true semantic content is a necessary condition for learning as well as for knowledge

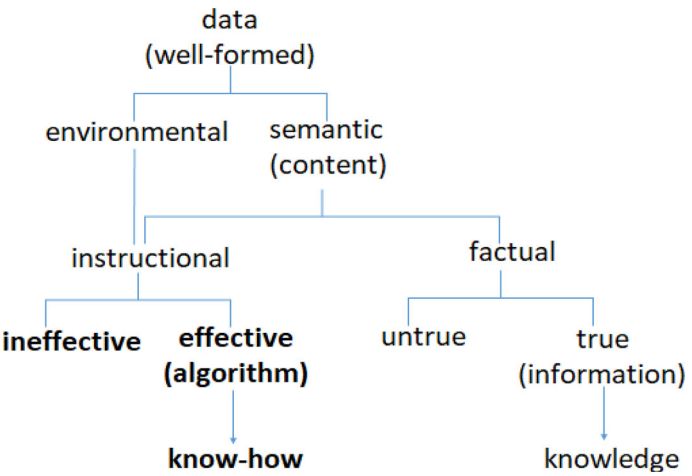

Fig. 1 Extended version of Luciano Floridi's general map of the information spectrum (Floridi, 2010). The extensions presented in the present paper are in **bold**. 'Well-formed' means that the data are composed in line with the rules (syntax) governing the system in question. Knowledge corresponds to acquaintance knowledge or knowledge how (Sect. 2.1). If the word instructional is replaced by skills of a conscious agent, and these skills are either ineffective or effective, then effective skills correspond to knowledge-how rather than know-how

Table 1 Concepts from Floridis' conceptual model of information (Fig. 1), and their relevance to some of the notions of information outlined in Sect. 2. Note that semantic information, in the sense of Sect. 2.2, is a slightly wider concept than in Floridis' model, since it also incorporates environmental information that corresponds to effective instructions (algorithms)

Concept from Floridis' model	Relation to notions of information in Sect. 2
Well-formed data	Data with a syntax (syntactic information, Sect. 2.4). The syntax eliminates some possible data sets, those that are not well formed (Sect. 2.3).
Environmental information	Data with a syntax, which brings information about something (natural information, Sect. 2.5). If this information is actively infused by an external agent, it is also an instance of active information (Sect. 2.10).
Instructional information	The data, with a syntax, are instructions for reality (practical information, Sect. 2.1). A closely related concept is cybernetic information (Sect. 2.8), since causal feedback loops (just as instructions) can either be effective or ineffective.
Ineffective instructional information	The instructions do not achieve any goal, with no specified information (Sect. 2.7).
Effective instructional information	The instructions achieve a goal (algorithmic information, Sect. 2.9). This goal is a specification of the algorithm (specified information, Sect. 2.7) and it provides a meaning (semantic information, Sect. 2.2).
Semantic information	Data with a syntax that involves a code. This code serves as a specification (specified information, Sect. 2.7) and it provides meaning (semantic information, Sect. 2.2). The code eliminates some possible well-formed data (Sect. 2.3). If the code is intentionally brought in by an external agent, it corresponds to active information (Sect. 2.10).
Factual information	Data with a syntax and a code, with details of how something is in reality (ontological information, Sect. 2.1).
Untrue factual information	Misinformation about reality (Sect. 2.1).
True factual information	An epistemic agent who learns about a true statement acquires epistemic information (Sect. 2.1). If the agent can justify what he learns, he has also acquired knowledge-that (Sect. 2.1).

acquisition of an epistemic agent (Sect. 2.1). This idea has been elaborated by Hössjer et al. (2022) in order to define a mathematical model for learning and knowledge acquisition.

Floridi also introduces another important type of information, or another way in which we regularly use the term 'information'. 'Environmental information' of the type often found in nature is included in his model. For example, smoke that is properly decoded provides information about fire. This type of natural information deals with how one thing can bring information about another thing in a system when there is a clear correlation between them, as outlined in Sect. 2.5. Strictly speaking, environmental information does not need to be naturally caused though, such as when a car engineer builds a red-light indicator signalling a low battery (Floridi, 2010, p. 32).

Concerning Floridi´s model, it should be noted that information science traditionally relates to semantic information, whereas environmental information is not part of information science. Floridi has much to say about semantic information, and less to say about environmental information. Floridi adopts a qualitative view of information as well-formed, true, and meaningful content. He does not adopt a quantitative measure.

A significant example is the genetic information in a cell's DNA and protein systems. Stegmann (2005) introduced the term *instructional information* to describe the type of information found in the genes in our DNA system, and Floridi applied the same term (Floridi, 2010, chap. 6). As mentioned in Sect. 2.1, genetic information is practical information, i.e., instructional information *for* something (in this case, the construction and maintenance of life), not just *about* something. This means that 'the medium (the genes) is the message'. The genes carry important chemical information about their own proper interpretation (metalanguage). In a way, they are self-interpretive, although the right context (cell environment) is also essential in the interpretation process. It seems natural to expand Floridi's model because information in genes can be informational sources in different ways. For instance, genetic information can be instructive, blueprint, hereditary (Sect. 2.6) or algorithmic (Sect. 2.9). It includes both software (formulation level) and hardware (execution level). Broken or damaged genes are examples of ineffective instructional information.

But since the abovementioned hardware is effective, there is also semantic information in gene instructions programs (Sect. 2.2). They have an external specification in terms of function for the proteins that the genes code for (Sect. 2.7). When processed, the effective information produces a nontrivial formal function. Merely describing a gene does not prescribe or produce results. Hence, mere description needs to be dichotomised from instruction, as instructional (or practical/algorithmic) information does far more than describe. The processing of genetic information is just as formal as the information it processes. Effective instructional information provides 'how-to' information, by prescribing, steering, and controlling physical interactions. It 'breathes fire into the equations and makes a universe for them to govern' (Hawking, 1988, p. 174). Know-how differs from knowledge-how in the sense that instructions (not the skills of conscious agent) are effective, and it differs from knowledge that in the sense that it does not relate to factual statements.

We should also mention that Floridi has expanded his understanding of genetic information over the years, although he has always classified it as instructional information. In a 2010 publication (Floridi, 2010, chap. 6), he classifies genetic information as environmental information because he considers it less demanding than a semantic interpretation, and in danger of losing its useful and concrete procedural sense. He also stressed that environmental information might be meaningful independently of an intelligent producer/informer. In the same publication, he drew a line between environmental and instructional information, as shown in Fig. 1. However, this line was omitted in another publication five years later (Floridi, 2015). Although this line was absent in his 2015 publication, he confirmed that he wanted to prioritise the diagram with the line (personal communication, 7 December 2020), as shown in Fig. 1. As we see it, there is no reason to deny that structures in nature, as long as they have an external specification, can be both meaningful and instructional, corresponding to

semantic information. This, of course, raises the question of how a structure obtains meaning and even instructional function. This presumably relates to basic theories and discussions on one of the hardest problems in semantics: 'The data grounding problem' – How can data acquire their meaning? (Crnkovic & Hofkirchner, 2011). Meaningful data may be an ontological concept embedded in information carriers. Meaning is not only an epistemic concept, in the mind of the user (Sect. 2.1).

4 Defining and quantifying genetic information

Natural science is largely concerned with quantification and measurement, and considerable effort has been invested in finding a general measure of information, although some researchers have expressed doubts about the possibility of measuring information in a fully satisfactory manner. What is said (that is, information) is not necessarily the same as being able to quantify how much it is said. There may be no direct quantifiable framework for mathematical biology in the same manner as such a framework is well-established in mathematical physics (Chaitin, 1979). Genetic information cannot at present be measured in a general manner, and the same is true for genetic meaning.

However, in genetics, as in any science, introducing measurable quantities is central to how we study subjects and frame our theories. And as argued in Sect. 2.2, under certain circumstances it possible to use such quantities to quantify genetic information with a syntactic or semantic element. These measures of information involve presumptions or specifications regarding signs, observers, and reference states that require careful consideration of the basic aspects of the system.

In this section we will give some examples of how to measure genetic information, that form a bridge between quantitative and qualitative information (Sect. 2.2). The most common representation of information is a linear sequence of symbols. A protein sequence of length L is described as a discrete random variable $\boldsymbol{X} = (X_1, \ldots, X_L)$, where X_j, $j = 1, \ldots, L$ is the amino acid of site j. Such protein sequences can be directly inferred from DNA sequences through the genetic code.

The biochemical *function* of a protein or a protein family is determined by direct empirical experimentation, and links information content to functionality (Adami & Nitash, 2022). Function is an objective feature because it is the same for all observers. In categorical terms, genetic function is possibly the best nominal depiction of the cellular 'meaning' of a sequence (cf. Section 2.7). It may be interpretated as an instance of instructional information in Floridi's general map (Fig. 1). This is also semantic information, since on top of instructions or syntax, the function is a type of specification that conveys meaning to an external observer.

As shown in Table 2, Barbieri (2016) has altogether portrayed five properties of genetic sequences. Originally Barbieri defined these five properties of a single genetic sequence $\boldsymbol{X}$, but it may be generalised to a family (f) alignment of sequences gathered into a matrix $\underline{\boldsymbol{X}}_f$. In the following sections we will discuss how these five properties of Table 2 possibly can be operationalised and measured within genetics.

Table 2 An overview of five distinct characteristics of protein sequences $\boldsymbol{X}$ (Barbieri, 2016; Thorvaldsen & Hössjer, 2023) and their scale of statistical measurement

Property of genes	Scientific framework	Statistical measure level	Section
Probability	Self-information: $I(\boldsymbol{X})$	Metric scale (bits)	4.1
Complexity	Algorithmic information: length	Metric scale (bits)	4.2
Distance	Relative distance: $D(\boldsymbol{X}_1, \boldsymbol{X}_2)$	Metric scale (bits)	4.3
Organic information	Function F in the context of a cell: joint variable $[\boldsymbol{X}, F]$	Joint [scale (bits), nominal]	4.3
Organic meaning	Cellular and intracellular networks, logistics in a living system	Joint nominals (categories)	4.4

Organic information and meaning are usually considered non-numerical entities but are objective observables in genetics and hence fundamental nominal data type. However, function F can be viewed as a specification of $\boldsymbol{X}$ that makes it possible to define its functional information as a type of specified information (Sects. 2.7 and 4.3 and Appendix A). Together with the instructions carried by $\boldsymbol{X}$ this makes it possible to quantify a notion of genetic information of $\boldsymbol{X}$ which has a semantic element (Sects. 2.1, 2.9 and 4.3). On the other hand, it is more challenging to quantify genetic information for cellular networks (Sect. 4.4)

4.1 Self-information of amino acid sequences

Let $\boldsymbol{X} = (X_1, \ldots, X_L)$, where X_j, $j = 1, \ldots, L$ is the amino acid of site j, be a protein sequence of length L. What is commonly referred to as *self-information* (corresponding to the first property of Table 2) in information theory may be applied to measure the information content, or 'surprisal', to each of the 20 amino acids x:

$$I(x) \stackrel{\text{def}}{=} \log_2 \frac{1}{p_x} = -\log_2(p_x),\ x = 1,2, \ldots 20,$$

where p_x is the probability of each amino acid, x. The use of logs to measure information goes back to Hartley (1928). Consider a large pool of N amino acids, with frequencies $p_1, \ldots, p_{20}$, and assume that one of these N amino acids is drawn randomly. Then $I(x)$ quantifies how much the number of possibilities (about which amino acid from the pool that was sampled) decreases after observing x, due to the fact that other possibilities are eliminated (Sect. 2.3). As stated by Dretske (1981, p.529), self-information reflects the fundamental intuition behind information. The *expected* (mean) self-information per amino acid

$$E(I) = \sum\nolimits_{x=1}^{20} p_x I(x)$$

is equivalent to the classical Shannon formula for entropy (or Shannon uncertainty). The measure of self-information for an entire sequence $\boldsymbol{X}$ of amino acids is positive

and additive if the components of $\boldsymbol{X}$ are independent. The unit of the self-information is 'bit', since base two is used in the formula for the logarithm.

An orthologue protein family f is commonly represented by the *alignment* of its sequences. 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 the amino acid of protein m at site j, or a gap. The (vertical) column vector $\boldsymbol{X}_j$ contains the amino acids at a single *site* j along a multiple sequence alignment of a protein family. We will define measures of information for protein families $\underline{X}_f$ in Sect. 4.3.

4.2 Algorithmic information of amino acid sequences

Algorithmic information uses the notion of algorithm to measure the amount of information. As discussed in Sect. 2.9, since algorithms contains instructions with an effect (an external specification), this type of information can be viewed as an instance of specified information (Sect. 2.7). And since a program is also effective, this external specification adds a semantic element to algorithmic information.

Formally, *Kolmogorov algorithmic information*, or complexity, of a finite string $\boldsymbol{X}$ of bits is the length of the shortest computer program that generates string $\boldsymbol{X}$ and stops (Kolmogorov, 1965). According to this measure (which corresponds to the second property of Table 2), the amount of information contained in data is equal to the shortest program that can reproduce it. This establishes an inverse connection between informativeness and predictability. Kolmogorov complexity is related to the compression of data and is sometimes called *descriptional complexity*. A nontrivial string may be incompressible and requires an algorithm or instruction set of complexity, such as the system it describes. Whereas algorithmic information theory uses the notion of a universal computer, and the resources required to reproduce data on that computer, it does not represent contingencies in terms of probabilities. Shannon's theory on information, on the other hand, is founded on the concept of probability, and its relation to contingency. This is true for the definition of specified information in Sect. 2.7 as well. It is shown in Appendix A.1 that Kolmogorov complexity is a special case of specified information, in spite of the fact that the latter uses probabilities for its definition.

By stating that Kolmogorov complexity measures the amount of information in a given string, one does not mean that it is actually possible to perform such a measurement. Numerous programs will generate $\boldsymbol{X}$, but the exact Kolmogorov measure of complexity has a disadvantage of being algorithmically unknowable, as there is no general method to compute it (Cover & Thomas, 2006). The notion of Kolmogorov complexity is primarily theoretical. However, it is possible to obtain an upper bound for it, and accordingly, it is bounded without being computed exactly. Several protein compression methods have been proposed in the literature. For a review we refer to Hosseini et al. (2016). Interestingly, the Kolmogorov complexity of $\boldsymbol{X}$ may be estimated from its output frequency distribution (Soler-Toscano et. al 2014).

4.3 Extending the Shannon measure of information to protein families

Yockey (1977) pioneered the application of information theory to protein sequences by estimating amino acid variability at each position in the primary sequence of the protein family cytochrome c. He found the information per amino acid to be 2.953 bits.

In a recent paper Thorvaldsen and Hössjer (2023) have demonstrated how variants of the Shannon information measure can be applied to molecular sequence data sets of protein families. This constitutes a series of very specific analyses, constrained by the assumptions about the underlying probability distributions of sequences from which data have been obtained. The reference distribution $\boldsymbol{p} = (p_1, \ldots, p_{20})$ on the set of amino acids is derived directly from the genetic code, where each amino acid x is assigned a prior probability $p_x = n_x/61$ proportionally to its constituting number n_x of codons (between 1 and 6), with corresponding a self-information $-\log_2(n_x/61)$ between 5.93 and 3.35 bits. This distribution assigns the same probability to each of the 61 non stop codons of the genetic code. It corresponds to a non-informative prior on the set of codons and hence is a natural starting point, from 'first principles' thinking, to model maximal ignorance about the codon distribution before any data has been analysed. It relies on using the Principle of Insufficient Reason (Bernoulli, 1713), or the principle of maximum entropy (Jaynes, 2003) for the prior codon distribution.

The concept of self-information is of relevance for quantifying natural information (Sect. 2.5). Suppose an amino acid $r = x$ has been observed, and we want to find out which codon s that was translated into r. Consider a particular codon F that translates to r. Observing an amino acid $r = x$ conveys information about the statement $s = F$, since the probability of this event increases from $1/61$ to $1/n_x = 1/\big(61\, p_x\big)$. Indeed, after having observed $r = x$, the statement $s = F$ make us less surprised, corresponding to a decrease $\log_2(61) - \log_2(61\, p_x) = I(x)$ of self-information. Consequently, $I(x)$ quantifies the amount of natural information carried by x.

Based on self-information it is also possible to estimate the information content of various protein domains and families $\underline{\boldsymbol{X}}_f$ in different ways. Table 3 gives an overview of three such information content quantities, and it summarises some of their properties (see Thorvaldsen & Hössjer, 2023 for more details).

As described in Table 3, a conditional version of the commonly used *Mutual information* from information theory may be applied to sites of aligned sequences. Mutual information captures all dependencies between two random variables. It measures how much the Shannon uncertainty for one random variable (which in our context is an amino acid sequence with a prior distribution) is expected to decrease when knowledge of another random variable (which in our context is an amino acid sequence with marginal distributions at all sites in agreement with the empirical distributions of the observed amino acid sequence) is taken into account. High mutual information indicates a large reduction in uncertainty. The conditional mutual information quantifies the corresponding observed (not expected) reduction in Shannon entropy.

Durston et al. (2007) applied conditional mutual information to compute the information content based on a uniform prior distribution $\boldsymbol{p}$ with $p_x = 1/20$ for each amino acid x. The sum of the contributions at each position of the alignment leads

Table 3 Three different quantities for measuring bits of information $I(\underline{X}_f)$ for a protein family $\underline{\boldsymbol{X}}_f$ of M amino acid sequences of length L, with $\boldsymbol{p} = (p_1, \ldots, p_{20})$ the prior distribution of amino acids, $\boldsymbol{q}_j = (q_{1\,j}, \ldots, q_{20,\,j})$ the observed empirical distribution of amino acids at site j of the protein family, whereas $\boldsymbol{r}_j = (r_{1\,j}, \ldots, r_{20,\,j})$ contains the probabilities r_{xj} of not rejecting an amino acid x, when such an amino acid is sampled from a large reservoir of amino acids (distributed according to the prior $\boldsymbol{p}$) in order to build up site j of the protein family

Quantity	Estimate of $I(\underline{X}_f)$	Per site range [min, max]	Value at site with conserved amino acid x
Conditional mutual information	$-L\boldsymbol{p}\cdot\log_2\boldsymbol{p} + \sum_{j=1}^{L}\boldsymbol{q}_j\cdot\log_2\boldsymbol{q}_j$	[− 0.18, 4.14]	4.14
Expected active information	$-\sum_{j=1}^{L}\boldsymbol{q}_j\cdot\log_2\boldsymbol{p} + \sum_{j=1}^{L}\boldsymbol{q}_j\cdot\log_2\boldsymbol{q}_j$	[0, 5.93]	$-\log_2 p_x$
Functional information	$-\sum_{j=1}^{L}\log_2(\boldsymbol{p}\cdot\boldsymbol{r}_j) = -\log_2\prod_{j=1}^{L}\boldsymbol{p}\cdot\boldsymbol{r}_j$, where $r_{xj} = \frac{q_{xj}/p_x}{\max(q_{1\,j}/p_1, \ldots, q_{20,\,j}/p_{20})}$	[0, 5.93]	$-\log_2 p_x$

Note that $I(\underline{X}_f) = \sum_{j=1}^{L} I(X_j)$ is additive over sites, for all three quantities. The two rightmost columns depict, for each quantity, the range of values $I(\boldsymbol{X}_j)$ can take per site j, and the value of $I(\boldsymbol{X}_j)$ at sites j for which x is conserved ($q_{xj} = 1$), respectively. The models are elaborated further in the text

to the conditional mutual information of the entire protein family. They examined the lower bounds for the conditional mutual information (in units of bits) for 35 protein families, with lengths L ranging between 33 and 949 amino acids, and computed an information content between 46 and 2416 bits. An improved approach is to apply the reference distribution defined in the beginning of Sect. 4.3 as prior (Thorvaldsen & Hössjer, 2023).

As mentioned in Sect. 2.10, *Active information,* I^+ was introduced by Dembski and Marks to handle infusion of knowledge in random search algorithms (Dembski & Marks, 2009a, 2009b). It was later applied to population genetics by Díaz-Pachón and Marks (2020). A general statistical framework for estimating active information is provided by Díaz-Pachón and Hössjer (2022).

In our context, $I_j^+(x) = \log_2\left(q_{xj}/p_x\right)$ is the active information associated with a change of frequency of amino acid x at site j from the prior probability p_x to the observed relative frequency q_{xj} in the protein family. Analogously, a change in the

frequency of an amino acid distribution from $\boldsymbol{p}$ to $\boldsymbol{q}_j$ at site j corresponds to the expected active information

$$E\left(I_j^+\left(X_j\right)\right) = E_{q_j}^{(p)} - E_{q_j}^{(q_j)} = -\boldsymbol{q}_j \cdot \log_2 \boldsymbol{p} + \boldsymbol{q}_j \cdot \log_2 \boldsymbol{q}_j$$
$$= \sum_{x=1}^{20} q_{xj} \log_2 \frac{q_{xj}}{p_x} = D_{KL}\left(\boldsymbol{q}_j || \boldsymbol{p}\right),$$

where expectation is over $\boldsymbol{q}_j = \left(q_{1j}, \ldots, q_{20,j}\right)$, the observed empirical distribution of amino acids at site j (cf. Table 3), whereas $\cdot$ refers to the dot product between two vectors of equal length. Hence, the expected active information at site j is equivalent to the information-based Kullback–Leibler divergence D_{KL} between $\boldsymbol{p}$ to $\boldsymbol{q}_j$ (Kullback & Leibler, 1951). Motivated by continuity, we define $0 \cdot \log 0 = 0$. It follows that the expression for the expected active information in Table 3 equals the total Kullback–Leibler divergence $\sum_{j=1}^{L} D_{KL}\left(\boldsymbol{q}_j || \boldsymbol{p}\right)$ between the prior and posterior distributions $\boldsymbol{p}$ and $\boldsymbol{q}_j$, summed over all sites. Each term of this expression is always non-negative and quantifies the directed 'distance', or relative information, between two probability distributions over the same sample space, with $D_{KL} = 0$ being the most similar (the probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}_j$ are identical). However, despite its many useful properties, the Kullback–Leibler divergence is still an asymmetric measure and thus does not qualify as a common metric of spread; it also does not satisfy the triangle inequality (Cover & Thomas, 2006). Consequently, when active information is employed, the Kullback–Leibler divergence gives rise to a relative distance (the third property of Table 2) $D\left(\boldsymbol{X}, \boldsymbol{X}_j\right) = D_{KL}\left(\boldsymbol{q}_j || \boldsymbol{p}\right)$ between two amino acid sequences $\boldsymbol{X}$ and $\boldsymbol{X}_j$, with amino acid distributions $\boldsymbol{p}$ and $\boldsymbol{q}_j$ respectively, that does not satisfy the triangle inequality.

The third model in Table 3 is *Functional information*. This notion of information was introduced in Sect. 2.7. The version of functional information in Table 3 is approximate, since it does not make direct use of the empirically observed function F from the fourth property of Barbieri's Table 2 (see Appendix A.2 for more details). It rather applies an indirect method inspired from rejection sampling (Wells et al., 2004), with a proposal distribution $\boldsymbol{p}$ and a target distribution $\boldsymbol{q}_j$ at site j. This sampling procedure is repeated independently for all sites $j = 1, \ldots, L$. It is hypothetically assumed 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 $\underline{X}_R$ of amino acids with distribution $\boldsymbol{p}$, and an amino acid x at site j is retained (not censored) with probability r_{xj}. The censoring mechanism gives rise to the measure of information expressed in Table 3, with the non-censoring probability viewed as an approximation of the fraction of functional amino acid sequences (Thorvaldsen & Hössjer, 2023).

The same expression also corresponds to the functional information as introduced by Jack Szostak in his important paper in *Nature* (Szostak, 2003) and subsequently studied in Hazen et al. (2007). As mentioned in Sect. 2.7, Szostak and colleagues specified *functional information* in terms of a gene string as $-\log_2$ of the tiny fraction of functional sequences that have fitness values (activity of a biopolymer) greater than

a specified value (Hazen et al., 2007).[1] For the approximate version of functional information of Table 3, the non-censored sequences are defined as functional (F), so that this fraction is the probability that a random sequence will not be censored.

The approximate measure of functional information of Table 3 also satisfies the triangle inequality (Thorvaldsen & Hössjer, 2023), but it does not qualify as a standard distance measure (it is not symmetric). In the mathematical literature such spaces are often named a *quasi-metric space*, or a "mountainous" space, since the effort of going upward to the top of a mountain is not the same as descending downhill to the starting point (Khamsi, 2015). This corresponds to using a non-symmetric relative distance (the third property of Table 2) $D(\boldsymbol{X}, \boldsymbol{X}_j) = -\log_2(\boldsymbol{p} \cdot \boldsymbol{r}_j)$ between two amino acid sequences $\boldsymbol{X}$ and $\boldsymbol{X}_j$ with amino acid distributions $\boldsymbol{p}$ and $\boldsymbol{q}_j$ respectively, where $\boldsymbol{q}_j$ is obtained from $\boldsymbol{p}$ and $\boldsymbol{r}_j$ as described in Table 3.

The models with expected active information and functional information have the advantage of being non-negative, and the latter quantity additionally approximates the functional information specified by Szostak. Expected active information does not meet the usual criteria of a distance measure, because it is both asymmetric and path-dependent. In contrast to the functional information, it does not satisfy the triangle inequality. On the other hand, the conditional mutual information quantity needs no correction for random sequences to have approximately zero information, whereas both the expected active information and functional information quantities will need such a correction term, as derived by Thorvaldsen and Hössjer (2023). The functional information and the conditional mutual information have successfully been applied on large multiple sequence alignment data derived from the Cath (Sillitoe et al., 2021) and Pfam (Mistry et al., 2021) databases (Thorvaldsen & Hössjer, 2023; Thorvaldsen et al., 2010).

4.4 Cellular and intracellular networks and logistics

Genes are not merely epistemic descriptions. These epistemic features only constitute a subset of the overall properties of genes, and although they are very useful, they are limited and inadequate for addressing many forms of instruction and control. Genes are also prescriptions for metabolic success. Abel (2009, 2012) issued the term *prescriptive information* to describe the sources and nature of coding controls, regulations, and algorithmic processing (cf. Sects. 2.4, 2.8 and 2.9). Such prescriptions are universally instantiated in all the known living cells. Prescriptive information instructs genetic functions in such a way as to realize a prescribed set of logic gate programming choices (Abel, 2009; D'Onofrio, 2012). Without such steering of physicochemical interactions, metabolic pathways and cycles would be unattainable to merge into a cooperative and holistic metabolism of a cell, where a non-trivial formal organisation is achieved.

[1] The definition of $I^+(A) = \log_2[Q(A)/P(A)]$ of active information, for a set A of amino acid sequences of length L and prior distribution P, can also be viewed upon as a generalization of specified information, (which corresponds to the special case $Q(A) = 1$, with A the set of specified states), if an external agent actively caused A to happen with certainty ($Q(A) = 1$). Cf. Díaz-Pachón and Hössjer (2022) and Appendix A for further details.

The functionality of the expressed genes is strongly context dependent. Only a very small subset of these molecules is at work (i.e., meaningful) at a given moment in the big intercellular network, whose functioning is dependent on preconditions of temperature, salinity, and pH. That milieu also involves other expressed proteins. Biological functionality is subject to informational control and feedback, so that the rules may change dynamically with time in a manner that is a function of the current state of the organism. Information of biological systems is not only a way to describe states but also an integrated property of the whole system. It has also causal efficacy (Walker & Davies, 2013, see also Sect. 2.4). A statistical framework to capture and measure the interactions of a cellular network will be exceedingly complex and remains one of the challenges of system biology. For instance, the functional information of such an interaction network would correspond to $-\log_2$ of the fraction of states of the system that correspond to a functioning network (cf. Sects. 2.7 and 4.3). However, the main challenge is firstly to define exactly what a functioning network means, and secondly to define a probability distribution on the set of possible states of networks in order to approximate the fraction of functional states (in a certain environment, at a specific time).

Although it is difficult to define the information of a whole network, it is sometimes possible to define the specified information of some of its subcomponents. Biological information of such a component can occur in both analogue and digital forms. An example of digital information is binary functionality (a component that either functions or not, cf. Appendix A.2). An example of analogue information is morphogens, where patterns of electric membrane potentials serve as morphological templates in 3D space (Levin, 2017). In either case, as long as the degree of specification of the component can be quantified, it is also possible to quantify its specified information (Sect. 2.7 and Appendix A).

Accordingly, a whole living system has a large informational narrative. It includes genes and gene products that are joined in an immensely organised network of information flow through the cell. Researching the protein interaction networks of all proteins in an organism is one of the crucial challenges in biology and a crucial part of systems biology. We may also describe this as process information or logistics of cellular and intracellular networks, corresponding to the fifth property of Table 2.

5 Discussion and concluding remarks

In this article we have reviewed the use of qualitative and quantitative notions of information within biology, as a first step towards developing a taxonomy of information of biology. In particular, we analysed information measures as applied to distinct dimensions of the genomic machinery by distinguishing their statistical and epistemological merits. We have argued that semantic information can be accessed for protein families, since they involve codes (the syntactic part) and exhibit function (the specification part). As we observed in Sect. 4.3, an asymmetric measure of functional information exists for protein families as a proxy for semantic information, while no similar measure exists at present for cellular and intracellular networks (Sect. 4.4). Under certain conditions the asymmetry of the functional information measure may be given a nice

interpretation, as the probability of acquiring information is typically less than the probability of losing information. For instance, deleterious mutations tend to erode information of protein families, and they are much more frequent than benign mutations, that correspond to a gain of information. We also found that protein families are defined both by a qualitative epistemology together with their quantitative measures in bits, whereas cellular networks are at present only described by their epistemic merits. This observation is interesting and may generally be reflecting scientific work in progress.

The quantification of functional information for a given (observed) protein family employs all recognised sequences of this family. Our approximation of functional information in Sect. 4.3 treats function as a binary feature, where protein families are sampled and the non-censored protein families are regarded as functional. The original functional information definition in Szostak (2003) rather treats functionality as a continuous feature, corresponding to *degree* of biofunction (e.g., the reaction rate). This more general definition of functional information is essentially equivalent to the definition of specified information in Sect. 2.7, regarding degree of functionality as the specification (see Appendix A).

Genes may have similar biochemical functions, without any noticeable sequence similarity. These *isoenzymes* vary in sequence, but catalyse the same reaction (Guzzi et al., 2012). Semantic similarity measures have been developed and applied as biomedical ontologies, and are used to connect genes and proteins based on the similarity of their functions rather than on their sequence resemblance. However, since the methods used in Sect. 4.3 are based on sequence similarity, they only work for orthologue sequences and must handle isoenzymes as separate groups. These methods can be used though to estimate (and compare) the information content of each ortholog group.

5.1 Is the reference to information in genetics just a metaphor?

Various objections have been raised to implementing informational concepts in physico-chemical areas of biology, like in genetics. There are two main positions on the ontological status of information in genetics (Kim et al., 2015), which relate to the basic discussion we mentioned in the introduction: Is the reference to information in genetics just a metaphor (the first position) or is it not (the second position)? This debate is possibly one of the deeply dividing issues in modern science, and here we only briefly review and discuss a few aspects of the arguments.

Although most molecular biologists would see no serious controversy in characterising DNA and proteins as 'information-bearing' molecules, and thereby using notions from information science, some philosophers of biology have challenged this strategy. The first position, that information in biology and genetics is a metaphor, is known as 'the physicalist thesis'. It has been held by a number of scientists and philosophers (Chargaff, 1963; Sarkar, 1996, 2000; Mahner & Bunge, 1997; Griffith, 2001; Boniolo, 2003; Levy, 2011). Adherents of the first position argue that while information is undoubtedly a useful *metaphor* to describe genetic systems, in the end, all biological complexity is, at least in principle, reducible to basic physics and chemistry. Scholars supporting this view consider the description of genes as content-bearing a

thin one, and they are skeptical about strong notions and ascription of intentional properties.

For instance, Stuart (1985, p. 441 ff) suggested that the description of biological processes in terms of informational transfer is treated as 'a metaphor with deeply anthropomorphic content', and that we, for this reason, should look for another (non-anthropomorphic) approach. Kay (2000) described the application of information theory to biology as mistaken, mainly because Shannon's theory lacks the idea of meaning. She contended that the term "information", as applied in biology, is just a metaphor since the term designates nothing real. Kay explained the origin of the use of the term information within biology because of various social forces that were operating within the techno-culture of the time (1994, 2000).

Similarly, Sarkar (1996, 2003) reasoned that the concept of information has limited theoretical relevance in biology, because it lacks explanatory or predictive power. Like Kay, he seems to regard the concept of information as a redundant metaphor that lacks ontological substance and empirical references. Sarkar calls his account of genetic information 'deflationary' (2003). However, Stegmann (2009) argues that the deflationary theory does not capture four essential features of the ordinary concept of genetic information: intentionality, exclusiveness, asymmetry (DNA to proteins and other developmental outcomes, but not vice versa), and causal relevance. The deflationary definition of genetic information is therefore disconnected from what is customarily meant by genetic information.

Levy (2011) argues that the most reasonable interpretation of informational notions in biology is fictional—metaphors rather than descriptions that are based on genuine semantic properties of macromolecules and cells. However, he also argues that appeals to information bear theoretical weight by allowing us to reason via a fiction about real causal properties. According to this view, invocations of information in biology are non-literal descriptions playing a genuine role in biological understanding. Informational language is what Levy call a liminal metaphor—one that operates near the threshold of the noticeable.

In an interesting chapter on 'Evolution, Theology and Biosemiotics', Robinson (2010, pp. 179–219) has argued that the philosophical critique of 'naive uses of information terminology in biology' is well founded. On the other hand, Robinson also points out (in line with the second position, whereby information is not just a metaphor) that it might be a mistake to eliminate all semiotic concepts from theoretical biology. He asks: 'In particular, has the seductiveness of the mathematical theory of information diverted attention from the possible relevance of semantic information—'meaning' informatio—to the origin of life?' (2010, p. 196). This is exactly the point suggested in Peircean biosemiotics that semiotic concepts and ideas are needed to obtain a satisfactory understanding of life. This mainly calls for the use of the qualitative aspects of information, with the implication that information in biology is more than a metaphor.

Even though Shannon's information theory has a limited application in describing biological systems, it has been successful in quantitatively assessing the complexity of biomolecules. We may define information in the general sense as 'all that which is communicated', and hence, the information within a living cell is much greater than its protein sequences. All parts of the cell, including the DNA, RNA, protein molecules, lipids and carbohydrates are in steady communication with each other.

Thousands of different types of interactions take place within the cell. Although we briefly treated cellular networks in Sect. 4.4, in the present paper we mainly study biological information in the limited context of the array of information encoded within a cell's proteins (Sect. 4.3). Information carried by a gene in a living cell is instructional in the sense that it guides the production of a specific protein. The gene opens up some possibilities and excludes others (Sect. 2.3). In this way, the gene indicates which of the logically possible worlds could or could not be actual. This corresponds well with the qualitative aspects of semantic, syntactic, specific, and algorithmic information, discussed in Sects. 2.2, 2.4, 2.7 and 2.9.

For instance, experimental work has established the functional specificity of the sequences of nucleotides in DNA and amino acids in proteins. Thus, the term information used in genetics refers not only to syntactic but also to functional properties of living systems. As Crick explained shortly after their discovery of the molecular structure of DNA, 'By information I mean the specification of the amino acid sequence in protein' (Crick, 1958, p. 144), and 'Information means here the *precise* determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein' (Crick, 1958, p. 153). This 'specification' and 'precise determination' can be associated not only with codes and syntax (Sect. 2.4), but also with functional properties of proteins, as quantified by functional information (Sects. 2.7 and 4.3).

Adherents of the abovementioned second position on the ontological status of information in genetics, consider information as *intrinsic* to the process of living systems. If this is correct, life has to be classified as distinct from other types of physical systems, as we know of no other category of physical systems where information is mandatory to specify its state and process (Walker & Davies, 2013). Since Mendel's time, when scientists began to search severely for what would be involved in explaining the mechanism of heredity, biologists have anticipated the need for some substance or feature in living organisms possessing such properties (Alberts et al., 1983, p. 21). Davies has argued that the 'specific randomness' of DNA base sequences constitutes the basic secret on the nature and origin of life (Davies, 1998, p. 120). Once an informational framework (such as syntactic and specified information, cf. Sections 2.4 and 2.7) is in place, it allows us to capture properties of the objects and processes in question. In this way the informational language serves as a way of pointing to the real (literally true) causal roles of those elements in terms of codes and specifications that an external epistemic agent recognizes (Sect. 2.1). Metaphoric language may just be reflecting the preliminary state of art of scientific work in progress, until more is learnt about the real cause and quantification of these specifications. By studying the informational architecture of cellular networks, Kim et al. (2015) concluded (along the lines of the second point of view) that information is definitely intrinsic to life and they argued that there is increasingly strong support for this viewpoint. This suggests that informational architecture is necessary to account for life, and even for the emergence of life.

Today it is commonly accepted by many eminent biologists that the sense of information rendered by Shannon's mathematical information theory is legitimate, relevant, and useful in several parts of biology. The analysis of genetics in terms of information causes a deeper understanding beyond mathematics, which may reveal several other important characteristics. Some have drawn on the teleosemantic theory in philosophy

to make sense of this kind of approach, extending Shannon's theory of information to the semantic notions of information discussed in Sect. 2.2. A minority tradition has argued that the enthusiasm for information in biology has been a theoretical wrong turn, and that it fosters naive distortions of our understanding of the roles of interacting causes within biology, with an implicitly dualist ontology. However, the support for this skeptical response is fading, as a plausible relationship between the qualitative and quantitative notions of information based on scientific measurements is being established, as described in Sect. 4 and Appendix A.

5.2 Concluding remarks

Genes accommodate instructions, which are a type of procedural information (or practical information, as explained in Sect. 2.1). In this sense, genes represent proper informational entities. This interpretation of genetic information is compatible with, but still goes beyond, Shannon's probabilistic theory of information, yet being less demanding than a full semantic interpretation, which also requires external specification of the effects of the instructions, in terms of an effective algorithm (Sect. 2.9). Informational concepts, beyond Shannon's probabilistic definition, therefore, have robust application at the level of genes. The explicit introduction of *functional information* is one way of quantifying *specified information*. We have argued that functionality not only is an external specification, but that it also provides meaning (Sect. 2.7). Functional information therefore serves to bridge the quantitative and semantic information concepts (Sect. 2.2) and it also brings about a statistical framework and testable hypotheses on the role of information in genes and genetic systems. But more generally, the basic division between syntactic and semantic information still requires better coupling and coherent understanding to develop a synthetic theory of information for genetic systems (Sect. 4.4). One way forward is to interpret semantics (for instance function) as a higher degree of specification than syntax, along the lines of Gitt's five levels of information (cf. Sects. 2.7 and 2.10).

The idea that genetic sequences are a type of molecularly coded information is already well accepted in current research. In the present study, the *algorithmic* nature of genes was applied in Sect. 3 to classify genetic information, making use of the *functional* categories of Fig. 1, as is commonly performed in biology. Each category of Fig. 1 may be analysed further using a quantitative measure of *self-information* (Sect. 4.1). Information measured in this manner, merely by the reduction of the relevant uncertainty within a given category of Fig. 1, compared to another larger frame of Fig. 1 that includes this category, is an important, albeit restricted, notion of information (Sect. 2.3). It does not cover the full range of information concepts. As discussed in Sects. 2.2–2.3, 2.7 and 2.9, self-information does not cover, for instance, the common-sense conception of information in human cognition and communication or algorithmic information theory. But as mentioned in Sects. 2.2, 2.4, 2.5 and 2.7, probability-based information concepts can still be used to discern other types of information, in terms of uncertainty reduction, when the discernment mechanism is related to codes or function.

Functional information is a joint concept, whose definition involves a probability-based self-information as well as a specification based on function (cf. the fourth property of Table 2). It is therefore a valuable approach for quantifying information in the context of genetics. This approach was elaborated in Sects. 2.7, 4.3 and Appendix A, and Fig. 2 shows how this relation connects to, and may provide a deeper understanding of, genetic information. The crucial parts of Fig. 2 are the two dashed lines, which demonstrate how various ways of quantifying information to the left, are also used to quantify some information with a semantic context (such as function) to the right.

As noted in Sect. 4.4, grand unified theory of biological information may be long ahead, perhaps even fundamentally unreachable, given the uncontained use of the term. There is much more information present in a biological system than can be counted by simple, direct observation; therefore, its quantification by observing frequencies of DNA or amino acid variants only amounts to gross bias by discarding. The biological organisms exist within a set of hierarchical levels from DNA to ecology (Farnsworth et al., 2012; Griffiths, 2017). The flow and representation of information in ecological

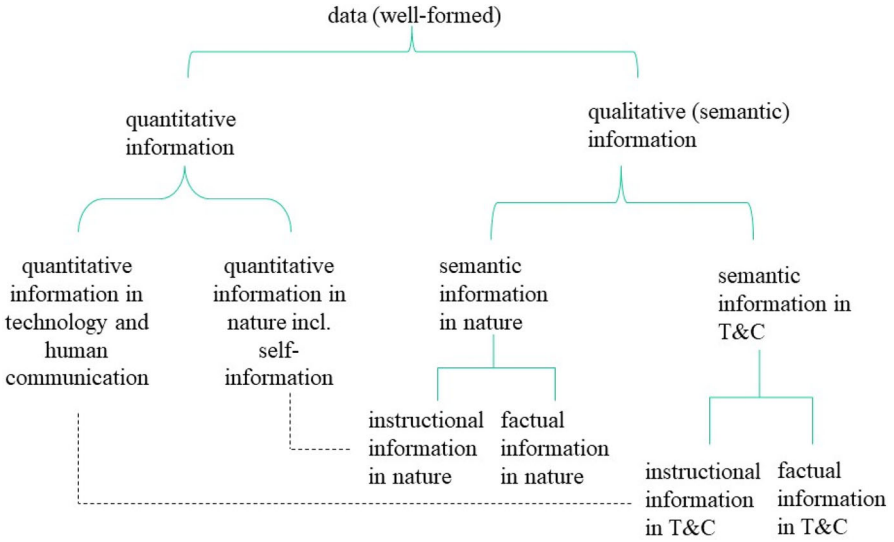

Fig. 2 An illustration of the relation between quantitative and qualitative information. As in Fig. 1 'well-formed' means that the data (or patterns) are composed in line with the rules (syntax) governing the system in question. Left part: For data with syntactic as well as quantitative information, the data have some statistical properties that convey information beyond the syntax itself. Note that this is not the same thing as Shannon information, since the latter only involves probabilities, not codes and syntax. Right part: For data with semantic information some type of specification is required beyond syntax, such as an entity that functions, an algorithm that is effective and achieves something when executed, or some message that conveys meaning beyond the syntax itself. The dashed lines between 'quantitative (syntactic) information' and 'semantical information' signify that quantitative information *is typically* used in semantic information (both instructional and factual), and that semantic information may itself be quantified (such as estimating the probability of obtaining by chance a given sentence with meaning). This holds for information in nature as well as for information in technology and human communication (T&C)

systems is information processing, which integrates information in multiple forms (O'Connor et al. 2019). Developing methods for observing, tracing, and quantifying information remains an object of research across disciplines. We hope that the present paper may serve to expand the common understanding of biological information, and be part of the process of creating a taxonomy for biological information.

A taxonomy of biological information is important, not the least since information is a conceptual key to a proper understanding of reality. Despite the large body of evidence that information plays a vital role in genetic systems, information handling and esteem does not yet occupy much space in mainstream theories and textbooks. This inconsistence is problematic and separates genetics from other exact sciences where quantitative techniques are widely implemented. But new scientific and philosophical thinking and work may eventually lead to a full recognition of a ternary informational "domain" that exists alongside the domains of spacetime and energy/matter.

As emphasized in Sects. 2–4, biological information has both a probabilistic, linguistic/syntactic, semantic and algorithmic dimension over an observable dataset. Through the representation and pragmatic assessment of these features of data, there is substantial justification for considering and discussing basic aspects of biology in terms, fundamental notions and ideas well-known from information theory (particularly in the area of technology and human communication, T&C) that go well beyond Shannon information. In this paper, we have presented some recent advances in the measurement of genetic information as a joint variable of function and sequence data (the fourth property of Table 2), and connected this approach to Floridi's general map of information. Functionality is a natural informational concept to use in T&C, and we have argued that it is important in biology as well. As summarized in Fig. 2, notions of information known from the study of T&C have so far turned out to be useful in biology. If there is a need for introducing a completely new and different information notion to deal with problems in genetics, it still has to be shown. The burden of proof falls on anyone who claims that there is such a need. Without such a demonstration it must be obvious to apply the principle of Ockham's razor assuming that the notion of information being developed within the study of T&C is also highly appropriate to deal with the informational questions and problems in genetics. Not as a metaphor, but as a truly inherent property of life.

Appendix A: Mathematical treatment of specified information and related quantities

The purpose of this appendix is to give a more mathematical treatment of specified information in order to show that several other notions of information, treated in this paper, are either special cases or at least closely related to specified information. To this end, let $\mathcal{X}$ be a sample space, P a probability measure for generating random outcomes $\boldsymbol{X}$ on this sample space and $f : \mathcal{X} \to \mathbb{R}$ a real-valued specificity function (Montañez, 2018), with $f(\boldsymbol{x})$ telling how specified outcome $\boldsymbol{x}$ is. For each $\boldsymbol{x} \in \mathcal{X}$ we let

$$A_{\boldsymbol{x}} = \{\boldsymbol{y} \in \mathcal{X}; f(\boldsymbol{y}) \geq f(\boldsymbol{x})\}$$

be the set of outcomes at least as specified as each $\boldsymbol{x}$. The specified information

$$I_{\boldsymbol{x}} = -\log_2 P(A_{\boldsymbol{x}}) = -\log_2 P(f(\boldsymbol{X}) \geq f(\boldsymbol{x})) \tag{1}$$

of $\boldsymbol{x}$ is defined as the self-information of $A_{\boldsymbol{x}}$. Equation (1) is consistent with the original definition of specified information (Dembski, 1998), that a feature $\boldsymbol{x}$ has specified information if i) it is unlikely to occur by chance and ii) it has an independent specification. Indeed, whenever $I_{\boldsymbol{x}}$ is large, property i) is satisfied since $P(A_{\boldsymbol{x}})$ is small, whereas ii) is satisfied since $A_{\boldsymbol{x}}$ is constructed from the specificity function f. Formula (1) is essentially used in Thorvaldsen and Hössjer (2020) and Díaz-Pachón and Hössjer (2022) as a definition of specified information.

A.1: Special cases of specified information

A number of special cases of (1) can be inferred. For instance, if $\mathcal{X}$ is the set of polymers of a certain length L and $f(\boldsymbol{x})$ is the degree of biofunction of $\boldsymbol{x}$, then (1) reduces to the definition of functional information given in Szostak (2003).

If $\mathcal{X}$ is the set of text strings of given length L, and $f(\boldsymbol{x}) \in \{0,1, 2,3, 4\}$ refers to the level of information of $\boldsymbol{x}$ according to the hierarchical taxonomy of Gitt (1989), described in Sect. 2.7, then $I_{\boldsymbol{x}}$ quantifies the amount of information associated with hierarchy $f(\boldsymbol{x})$. Indeed, the higher level of information $\boldsymbol{x}$ conveys (the larger $f(\boldsymbol{x})$ is), the larger is $I_{\boldsymbol{x}}$.

If $\mathcal{X}$ is the set of systems of a certain type, and $f(\boldsymbol{x})$ either refers to the amount of circular causality of $\boldsymbol{x}$, or the degree of organization of $\boldsymbol{x}$, then $I_{\boldsymbol{x}}$ quantifies the amount of cybernetic information (Sect. 2.8) of $\boldsymbol{x}$.

Suppose $\mathcal{X}$ is a countably infinite space of all sequences, from a given alphabet, of finite length. For each $\boldsymbol{x} \in \mathcal{X}$ let $f(\boldsymbol{x}) \in \mathbb{Z}_+$ refer to the shortest binary computer program able to produce $\boldsymbol{x}$ as an output (or end state/goal). Assuming that each finite binary string corresponds to a computer program that produces an output in $\mathcal{X}$, it follows that $P(\{\boldsymbol{x}\}) = 2^{-2f(\boldsymbol{x})}$ defines a probability measure on $\mathcal{X}$. It can then be seen that $I_{\boldsymbol{x}}$ equals the Kolmogorov complexity $f(\boldsymbol{x})$ of $\boldsymbol{x}$ minus 1. This follows by taking $-\log_2$ of the identity

$$\begin{aligned} 2^{-I_{\boldsymbol{x}}} &= P(\boldsymbol{X} \geq \boldsymbol{x}) = \sum_{\boldsymbol{y}; f(\boldsymbol{y}) \geq f(\boldsymbol{x})} 2^{-2f(\boldsymbol{y})} \\ &= \sum_{n=f(\boldsymbol{x})}^{\infty} \sum_{\boldsymbol{y}; f(\boldsymbol{y})=n} 2^{-2f(\boldsymbol{y})} = \sum_{n=f(\boldsymbol{x})}^{\infty} 2^n \cdot 2^{-2n} = 2^{1-f(\boldsymbol{x})}. \end{aligned}$$

Hence algorithmic complexity (Sect. 2.9) is formally a special case of specified information. However, as mentioned in Sect. 2.9, the specificity function f is not computable, and moreover the random measure P involves f.

A.2: Functional information with a deterministic or random, binary-valued specification

Let $\mathcal{X}$ refer to the set of amino acid sequences of a given length L, whereas $f(\boldsymbol{x}) \in \{0,1\}$ tells whether $\boldsymbol{x} = (x_1, \ldots, x_L)$ corresponds to a functioning protein (1) or not (0). Define the probability measure P on through

$$P(\{\boldsymbol{x}\}) = \prod_{j=1}^{L} p_{x_j},$$

where $\boldsymbol{p} = (p_1, \ldots, p_{20})$ is the vector of prior amino acid probabilities defined in Sect. 4.3. Then

$$I = -\log_2 P\left(\sum_{\boldsymbol{x} \in A} P(\{\boldsymbol{x}\})\right) = -\log_2 P(A) \tag{2}$$

refers to the amount of functional information associated with the set of amino acid sequences

$$A = \{\boldsymbol{x} \in \mathcal{X}; f(\boldsymbol{x}) = 1\}$$

that correspond to a functioning protein. Note that (2) is a special case of the functional information definition (1) of Szostak (2003), for binary-valued specificity functions. Let us now consider the case of a random and binary-valued specificity function, with

$$r_{\boldsymbol{x}} = P(\boldsymbol{x} \text{ is functioning}) = P(f(\boldsymbol{X}) = 1 | \boldsymbol{X} = \boldsymbol{x}) = \prod_{j=1}^{L} r_{x_j j},$$

where $\boldsymbol{r}_j = (r_{1j}, \ldots, r_{20j})$ contains the non-censoring probabilities of Table 3 from site j. Let $\boldsymbol{X} = (X_1, \ldots, X_L) \sim P$ refer to a randomly chosen amino acid sequence. Then

$$I = -\log_2 P(\boldsymbol{X} \text{ is functioning}) = -\log_2 \prod_{j=1}^{L} \boldsymbol{p} \cdot \boldsymbol{r}_j \tag{3}$$

is equivalent to the functional information of Table 3.

Note that (2) is a special case of (3) when the specificity function is deterministic, that is, when $r_{\boldsymbol{x}} = 1(\boldsymbol{x} \in A)$. Equation (2) is more appropriate to use when it is possible to determine empirically whether a structure $\boldsymbol{x}$ functions or not, whereas (3) is applicable when functionality cannot be determined empirically, but only the probability $r_{\boldsymbol{x}}$ of $\boldsymbol{x}$ functioning.

A.3: Active and natural information

Suppose we have a second probability measure Q on $\mathcal{X}$, and let $A \subset \mathcal{X}$ be a subset of the sample space. Define

$$I^+(A) = \log_2 \frac{Q(A)}{P(A)}, \tag{4}$$

when Q refers to the probability distribution of an external agent who brings about an outcome in $\mathcal{X}$, then $I^+(A)$ is the active information of A (Sects. 2.10 and 4.3). In the context of Werner Gitt's five levels of information, $\mathcal{X}$ may be taken as the set of text strings from a certain alphabet of length L, whereas $A = \{x \in \mathcal{X}; f(\boldsymbol{x}) \geq 3\}$ are those text strings for which the sender requires an action, possibly also having a purpose with the message. If P corresponds to a randomly produced text, whereas Q designates a text generated by such an external and conscious agent, then $I^+(A)$ quantifies the amount of external information this agent brings about in order to produce a text that mediates action or purpose.

However, (4) can also be associated with natural information (Sect. 2.5), when $\mathcal{X}$ is the space of possible source signals s, whereas $P(A) = \text{Prob}(s \text{ is in } A|k)$ and $Q(A) = \text{Prob}(s \text{ is in } A|r \text{ is } G\&k)$ refer to the conditional probabilities of the source s belonging to A, given only background knowledge k, or given background knowledge k as well as an observation r in state G. According the weaker definition of natural information of Sect. 2.5, given background knowledge k, r is G conveys natural information about the event s is in A if and only if $I^+(A) > 0$. As mentioned in Sects. 2.5 and 2.6, this definition of natural information includes codes, as well as instructive, blueprint, and hereditary information. The main difference between the natural information and active information interpretations of (4) is that typically, for natural information, no external agent is associated with producing the received signal r, whereas for active information, such an agent produces an outcome with distribution Q.

Acknowledgements The authors wish to thank two anonymous reviewers, whose comments significantly improved the quality of the manuscript.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Steinar Thorvaldsen, Ola Hössjer and Peter Ørhrstrøm. The first draft of the manuscript was written by Steinar Thorvaldsen, and Appendix A was written by Ola Hössjer. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Funding Open access funding provided by UiT The Arctic University of Norway (incl University Hospital of North Norway).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abel, D. L. (2009). The GS (genetic selection) principle. *Frontiers in Bioscience, 14*, 2959–2969. https://doi.org/10.2741/3426

Abel, D. L. (2012). Is life unique? *Life, 2*, 106–134. https://doi.org/10.3390/life2010106

Adami, C., & Nitash, C. G. (2022). Emergence of functional information from multivariate correlations. *Philosophical Transactions of the Royal Society a., 4*, 89. https://doi.org/10.1098/rsta.2021.0250

Alberts, A., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. D. (1983). *Molecular biology of the cell* (p. 21). Garland.

Atlan, H., & Koppel, M. (1990). The cellular computer DNA: Program or data. *Bulletin of Mathematical Biology, 52*, 335–348.

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

Baker, B. (2021). Natural information, factivity and nomicity. *Biology and Philosophy, 36*, 26. https://doi.org/10.1007/s10539-021-09784-4

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

Bergström, C. T., & Rosvall, M. (2011). The transmission sense of information. *Biology and Philosophy, 26*(2), 159–176. https://doi.org/10.1007/s10539-009-9180-z

Bernoulli, J. (1713). Ars Conjectandi. Thurneysen Brothers.

Boniolo, G. (2003). Biology without Information. *History and Philosophy of the Life Sciences., 25*, 255–273.

Borgmann, A. (1999). *Holding on to reality*. The University of Chicago Press.

Bowie, J., & Sauer, R. (1989). Identifying determinants of folding and activity for a protein of unknown sequences: Tolerance to amino acid substitution. *Proceedings of the National Academy of Sciences of the USA, 86*, 2152–2156. https://doi.org/10.1073/pnas.86.7.2152

Carnap, R. (1947). *Meaning and necessity*. The University of Chicago Press.

Chabert, J.-L. (2012). *A History of algorithms: From the pebble to the microchip*. Springer.

Chaitin, G. J. (1979). Toward a mathematical definition of 'life.' In R. D. Levine & M. Tribus (Eds.), *The maximum entropy formalism.* MIT Press.

Chargaff, E. (1963). *Essays on nucleic acids*. Elsevier.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley.

Crick, F. (1958). On protein synthesis. *Symposium for the Society of Experimental Biology, 12*, 138–163.

Crnkovic, G. D., & Hofkirchner, W. (2011). Floridi's "open problems in philosophy of information", ten years later. *Information, 2*, 327–359. https://doi.org/10.3390/info2020327

Davies, P. (1998). *The fifth miracle* (p. 120). Simon and Schuster.

Davies, P. C. W., & Walker, S. I. (2016). The hidden simplicity of biology. *Reports on Progress in Physics, 79*, 102601. https://doi.org/10.1088/0034-4885/79/10/102601

Dembski, W.A. and Marks II, R.J. (2009a). Bernoulli's principle of insufficient reason and conservation of information in computer search. In Proceeding of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, 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 a, Systems & Humans, 5*(5), 1051–1061. https://doi.org/10.1109/TSMCA.2009.2025027

Dembski, W. A. (1998). *The design inference eliminating chance through small probabilities cambridge studies in probability, induction, and decision theory*. Cambridge University Press.

Dembski, W. A. (2014). *Being as communion*. Ashgate.

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

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

D'Onofrio, D. J., Abel, D. L., & Johnson, D. E. (2012). Dichotomy in the definition of prescriptive information suggests both prescribed data and prescribed algorithms: Biosemiotics applications in genomic systems. *Theoretical Biology and Medical Modelling, 9*(1), 8. https://doi.org/10.1109/10.1186/1742-4682-9-8

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.

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*, 47. https://doi.org/10.1186/1742-4682-4-47

Eigen, M. (1992). *Steps towards life: A perspective on evolution* (trans. by Paul Woolley). Oxford University Press, p. 12.

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

Floridi, L. (2002). What is the philosophy of information? *Metaphilosophy, 33*, 123–145. https://doi.org/10.1111/1467-9973.00221

Floridi, L. (2010). *Information*. Oxford University Press.

Floridi, L. (2015). Semantic conceptions of information. In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/information-semantic/. (Accessed 21 August 2021).

Floridi, L. (Ed.). (2016). *The Routledge handbook of philosophy of information*. Routledge.

Gitt, W. (1989). Information: The Third Fundamental Quantity. *Siemens Review, 56*(6), 36–41.

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

Griffiths, P. E. (2001). Genetic information: A metaphor in search of a theory. *Philosophy of Science, 68*(3), 394–412.

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

Hawking, S. (1988). *A brief history of time*. Bantam Books Toronto.

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

Hoffmeyer, J. (2008). *Biosemiotics: An examination into the science of life and the life of science*. University of Scranton Press.

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

Hutson, M. (2019). AI protein-folding algorithms solve structures faster than ever. *Nature News*. https://doi.org/10.1038/d41586-019-01357-6

Hössjer, O., Díaz-Pachón, D. A., & Rao, S. (2022). A formal framework for knowledge acquisition: Going beyond machine learning. *Entropy*, *24*, 14–69.

Jablonka, E. (2002). Information: Its Interpretation, its inheritance and its sharing. *Philosophy of Science, 69*, 578–605. https://doi.org/10.1086/344621

Jaynes, T. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Khamsi, M. A. (2015). Generalized metric spaces: A survey. *Journal of Fixed Point Theory and Applications, 17*, 455–475. https://doi.org/10.1007/s11784-015-0232-5

Kay, L. E. (1994). Who wrote the book of life? Information and the transformation of molecular biology. *Science in Context, 8*, 601–634. https://doi.org/10.1017/S0269889700002210

Kay, L. E. (2000). *Who wrote the book of life?* Stanford University Press.

Kim, H., Davies, P., & Walker, S. I. (2015). New scaling relation for information transfer in biological networks. *Journal of the Royal Society Interface, 12*, 20150944. https://doi.org/10.1098/rsif.2015.0944

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Informations Transmission, 1*(1), 1–7.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*, 79–86. https://doi.org/10.1214/aoms/1177729694

Levin, M. (2017). Molecular bioelectricity: How endogenous voltage potentials control cell behavior and instruct pattern regulation in vivo. *Molecular Biology of the Cell, 25*(24), 3835–3850. https://doi.org/10.1091/mbc.e13-12-0708

Levy, A. (2011). Information in biology: A fictionalist account. *Noûs, 45*(4), 640–657.

López-Ruiz, R. (2005). Shannon information, LMC complexity and Rényi entropies: a straightforward approach. *Biophysical Chemistry, 115*(2–3), 215–218.

Mahner, M., & Bunge, M. (1997). *Foundations of biophilosophy*. Springer.

Martinez, M. and Sequoiah-Grayson, S. (2018). Logic and information. In *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/logic-information/. (Accessed 21 August 2021).

Maynard-Smith, J. (2000). The concept of information in biology. *Philosophy of Science, 67*, 177–194.

Meyer, S. C. (2003). DNA and the origin of life: information, specification, and explanation. In J. A. Campbell & S. C. Meyer (Eds.), *Darwinism, design, and public education.* Michigan State University Press.

Millikan, R. (2000). What has natural information to do with intentional representation? (appendix B). In R. Millikan (Ed.), *On clear and confused ideas* (pp. 1–18). Cambridge University Press.

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

Monod, J. (1971). *Chance and necessity: An essay on the natural philosophy of modern biology*. Alfred A. Knopf.

Montañez, G. D. (2018). A unified model of complex specified information. *Biocomplexity, 2018*(4), 1–26. https://doi.org/10.5048/BIO-C.2018.4

Newman, S. A. (2022). Inherency and agency in the origin and evolution of biological functions. *Biological Journal of the Linnean Society*. https://doi.org/10.1093/biolinnean/blac109

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

Osimani, B. (2014). Causing something to be one way rather than the other. *Kybernetes, 43*(6), 865–881. https://doi.org/10.1108/K-07-2013-0149

Pavese, C. (2021). Knowledge how. In E. N. Zalta (Ed.), *(edn) The stanford encyclopedia of philosophy.* Stanford University.

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.

Robinson, A. (2010). *God and the World of Signs. Trinity, Evolution, and the Metaphysical Semiotics of C. S. Peirce*. Philosophical Studies in Science and Religion, Vol. 2. Brill, Hotei Publishing.

Sarkar, S. (1996). Biological information. A skeptical look at some central dogmas of molecular biology. In S. Sarkar (Ed.), *The philosophy and history of biology* (pp. 187–231). Kluwer Academic Publishers.

Sarkar, S. (2000). Information in genetics and developmental biology. *Philosophy of Science, 67*, 208–213. https://doi.org/10.1086/392771

Sarkar, S. (2003). Genes encode information for phenotypic traits. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 259–272). Blackwell.

Scarantino, A. (2015). Information as a probabilistic difference maker. *Australasian Journal of Philosophy, 93*(3), 419–443. https://doi.org/10.1080/00048402.2014.993665

Schrödinger, E. (1945). *What is life? The Physical aspect of the living cell*. Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

Shea, N. (2013). Inherited representations are read in development. *British Journal for the Philosophy of Science, 64*, 1–31.

Sillitoe, I., et al. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research, 49*(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

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

Stegmann, U. E. (2005). Genetic information as instructional content. *Philosophy of Science, 72*(3), 425–443. https://doi.org/10.1086/498472

Stegmann, U. E. (2009). DNA, inference, and information. *The British Journal for the Philosophy of Science, 60*(1), 1–17. https://doi.org/10.1093/bjps/axn041
Stegmann, U. E. (2015). Prospects for probabilistic theories of natural information. *Erkenntnis, 80*, 869–893. https://doi.org/10.1007/s10670-014-9679-9
Stewart, I. (1999). *Life's other secret: The new mathematics of the living world*. Penguin.
Stuart, C. I. J. M. (1985). Physical models of biological information and adaption. *Journal of Theoretical Biology, 113*, 441–454.
Sustar, P. (2007). Crick's notion of genetic information and the 'central dogma' of molecular Biology. *British Journal for the Philosophy of Science, 58*(1), 13–24. https://doi.org/10.1093/bjps/axl018
Szathmáry, E., & Maynard Smith, J. (1995). The major evolutionary transitions. *Nature, 374*, 227–232. https://doi.org/10.1038/374227a0
Szostak, J. (2003). Functional information: Molecular messages. *Nature, 423*, 689. https://doi.org/10.1038/423689a
Tokuriki, N., & Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology, 19*, 596–604. https://doi.org/10.1016/j.sbi.2009.08.003
Thorvaldsen, S., Flå, T., & Willassen, N. P. (2010). DeltaProt: A software toolbox for comparative genomics. *BMC Bioinformatics, 11*, 573. https://doi.org/10.1186/1471-2105-11-573
Thorvaldsen, S., & Hössjer, O. (2020). Using statistical methods to model the fine-tuning of molecular machines and systems. *Journal of Theoretical Biology*. https://doi.org/10.1016/j.jtbi.2020.110352
Thorvaldsen, S., & Hössjer, O. (2023). Estimating the information content of genetic sequence data. *Journal of the Royal Statistical Society Series C, 72*(5), 1310–1338. https://doi.org/10.1093/jrsssc/qlad062
von Neumann, J. (1961). *Collected works*. Pergamon Press.
von Neumann, J., Aspray, W., & Burks, A. W. (1987). *Papers of John Von Neumann on computing and computer theory*. MIT Press.
Walker, S. I., & Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of the Royal Society Interface, 10*, 79. https://doi.org/10.1098/rsif.2012.0869
Weinberger, E. D. (2002). A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Bio Systems, 66*, 105–119. https://doi.org/10.1016/S0303-2647(02)00038-2
Wells, M. T., Casella, G., & Robert, C. P. (2004). Generalized Accept-Reject sampling schemes. Institute of Mathematical Statistics Lecture Notes. *A Festschrift for Herman Rubin, 45*, 342–347.
Wheeler, J. A. (1990). Information, physics, quantum: The search for links. In W. H. Zurek (Ed.), *Complexity, entropy, and the physics of information.* Addison-Wesley.
Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine (2nd revised ed., 1961)*. MIT Press.
Wills, P. R. (2016). DNA as information. *Philosophical Transactions of the Royal Society A, 374*, 2063. https://doi.org/10.1098/rsta.2015.0417
Yockey, H. P. (1974). An application of information theory to the central dogma and the sequence hypothesis. *Journal of Theoretical Biology, 46*, 369–406. https://doi.org/10.1016/0022-5193(74)90005-8
Yockey, H. P. (1977). On the information content of cytochrome. *Journal of Theoretical Biology, 67*, 345–376. https://doi.org/10.1016/0022-5193(77)90043-1
Yockey, H. P. (2000). Origin of life on earth and Shannon's theory of communication. *Computers & Chemistry, 24*(1), 105–123. https://doi.org/10.1016/S0097-8485(99)00050-9
Yockey, H. P. (2005). *Information theory, evolution, and the origin of life*. Cambridge University Press.