Biological information as choice and construction

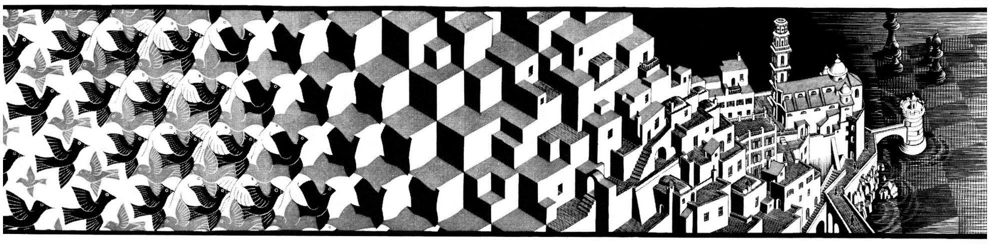
Arnaud Pocheville

*Department of Philosophy & Charles Perkins Centre,
The University of Sydney, NSW 2006, Australia*

https://arnaud.pocheville.science/

Abstract

A causal approach to biological information is outlined. There are two aspects to this approach: information as determining a choice between alternative objects, and information as determining the construction of a single object. The first aspect has been developed in earlier work to yield a quantitative measure of biological information that can be used to analyse biological networks. This paper explores the prospects for a measure based on the second aspect, and suggests some applications for such a measure. These two aspects are not suggested to exhaust all the facets of biological information.



Escher. Metamorphosis II. 1940

How to cite • Arnaud Pocheville (2018). “Biological information as choice and construction”. In: *Philosophy of Science.* Proceedings of the 2016 Biennial Meeting of the Philosophy of Science Association 85.5. Ed. by Wendy Parker.

1 Introduction

Biological development is classically assumed to reflect the expression of information accumulated in the genome during evolution (Mayr 1961; Jacob 1970). Major textbooks and popular science presentation of biology rely on this picture (e.g. Alberts et al. 2013). Leading biologists are also attracted to this view (Williams 1992, p. 10; Maynard Smith and Szathmary 1995; 2000; Jablonka 2002). On closer scrutiny however, the role of information in biology seems purely instrumental: it serves either as a metaphor or as a tool for big data analyses; biology does not yet have a theory of life as an information-processing phenomenon (Sarkar 1996; Godfrey-Smith 2000). The aim of this paper is to offer some scientific substance to such a theory.

Several theoretical and philosophical approaches have interpreted living systems as information-processing systems. One tradition identifies information with meaning, interpretation, and intentionality (Barbieri 2007; Shea 2007). A second tradition, which we espouse here, identifies information with patterns of association between objects (Dretske 1981).

We start from the sense of information introduced by Crick in his sequence hypothesis and central dogma of molecular biology, which was to become massively influential in biology: 'Information ... means the precise determination of sequence...' (Crick 1958, p. 153; see Kay 2000). Information here is causal (Šustar 2007). Crick introduced this conception in an attempt to understand how DNA and RNA carry biological specificity for the synthesis of proteins, an idea which parallels the modern contrast philosophers draw between specific causes and other necessary, background factors to obtain an effect (Woodward 2010). Griffiths and Stotz (2013, Ch. 4) have argued that Crick's sense of information vindicates the idea that factors other than DNA are also sources of information for biomolecules, a phenomenon they called 'distributed specificity'. This idea needs substantiation.

Acknowledgements • Maël Montévil is warmly thanked for insightful discussions and coining the language of the cell. A development of this paper is in preparation (Pocheville and Montévil n.d.). Two anonymous reviewers and the editor made numerous very helpful suggestions, and the audience at the 2016 PSA meeting provided kind feedback. Paul Griffiths and Karola Stotz mentored the project, and Maureen O'Malley provided kind encouragement. This research was supported by the Judith and David Coffey Life Lab (headed by Jean Yang), Charles Perkins Centre, The University of Sydney. This publication was made possible through the support of two grants from the Templeton World Charity Foundation, Inc. (TWCF0063 to Paul Griffiths, TWCF0242 to Arnaud Pocheville). The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Templeton World Charity Foundation, Inc.

Version history • First version: 4 June 2017; revised after review: 4 November 2017, 26 March 2018; editorial revisions: 13 June 2018, 13 July 2018, 31 July 2018; available online: 1 August 2018; editorial revisions: 5 October 2018; online: 8 October 2018.

We explore here this idea and develop an approach to biological information as a measurable and distinctive aspect of biological systems. Our approach has two facets, inspired from respectively Shannon's and Kolmogorov's approaches in information theory. On the one hand, we have a measure of the relative, complementary influence of several causes of the same effect (Section 2). This concerns the choice between a set of alternative objects, and is blind to the information content of each object. This approach has been extensively discussed and applied elsewhere. On the other hand, we have measures of the complexity of a single object, independently of any particular set of alternatives (Section 3). These can measure the information inherent to a biomolecule and the quantity of information in a molecule that can be attributed to a particular source. The computability of these latter measures, however, is problematic; in practice, tentative measures ought to be used. The role of randomness in creating information is outlined (Section 4). We sketch potential developments for a Kolmogorov-inspired approach (Section 5), and argue that it is a potentially fruitful yet challenging biological research program (Section 6). The two approaches are not straightforwardly reducible to one another, and are not suggested to exhaust all aspects of biological information.

2 Causal specificity: information as choice

Recent work has defined an information-theoretic measure of the 'specificity' of a cause for an effect, the extent to which a cause precisely determines an effect, and applied this measure to biological problems (Griffiths et al. 2015; Calcott et al. 2017; Pocheville et al. 2017; Weber 2017). This work develops earlier, qualitative discussions of 'causal specificity' in philosophy (Woodward 2010) and converges with formal work on causation in complex systems theory (see Footnote 1).

Causal specificity is measured using Shannon information theory, which conceives information as a reduction in uncertainty (Shannon 1948; Cover and Thomas 2006). Uncertainty, measured in bits, can be understood as the average number of binary (yes/no) questions that are required to determine the value of an unknown variable. A variable is said to share mutual information with another variable when it reduces our uncertainty about that variable. Mutual information measures the association between two variables: the more two variables are associated, the more each of them answers questions about the value of the other. Causal specificity can be measured by the mutual information between values of a cause variable *set by an intervention* and the value of a putative effect variable (Griffiths et al. 2015, p. 538). Formally, the causal specificity of C for E when controlling for a putative background B is given by the following formula (Pocheville et al. 2017):[1]

$$I(\widehat{C}; E \mid \widehat{B}) = \sum_{b} p(\widehat{b}) \sum_{c} p(\widehat{c} \mid \widehat{b}) \sum_{e} p(e \mid \widehat{c}, \widehat{b}) \log_2 \frac{p(e \mid \widehat{c}, \widehat{b})}{p(e \mid \widehat{b})}.$$

The '$\widehat{\ }$' (hat) on a variable is an operator indicating that its value is set by an intervention, rather than observed (Pearl 2009).[2] This operator transforms the symmetrical mutual information, representing observed association, into an asymmetric measure of causal influence, representing how much experimentally intervening on C whilst controlling for B affects E. If C is not a cause of E, then $I(\widehat{C}; E \mid \widehat{B}) = 0$. Reciprocally, if C is a cause of E, then there exists at least one set of background variables B (which can be empty) such that $I(\widehat{C}; E \mid \widehat{B}) > 0$ (Pocheville n.d.[a]).

This measure of causal specificity seems to capture one aspect of Crick's, and the above cited biologists', conception of information as 'precise determination'. It can be used to compare the causal contributions of genetic and epigenetic causes to the production of biomolecules (Griffiths et al. 2015). It can be applied to objects other than biomolecules, and is a practical tool for the analysis of biological networks (Tononi et al. 1999; Calcott et al. 2017; Pocheville n.d.[b]).

There is, however, a blindspot in Shannon information theory: it is silent about the information content of the objects themselves. For example, it makes no difference to the amount of information that DNA carries about RNA whether the DNA strands are 3 or 1416 nucleotides long. What matters is only the number of values that the variable 'DNA' can take and the probability distribution over those values. Arguably, the longer the sequences, the greater the number of possible alternatives, thus the greater the potential causal specificity of these alternatives.[3]

[1] This measure has been previously proposed in cognitive sciences (Tononi et al. 1999) and in computational sciences (Korb et al. 2009). Closely related measures have been proposed by Ay and Polani (2008) and Janzing et al. (2013). Transfer entropy is another 'information as choice' measure – although correlational in character, not causal (Lizier and Prokopenko 2010) – which has been applied to the study of the origins of life (Walker et al. 2017).

[2] As an anonymous reviewer noticed, the term $p(\widehat{c} \mid \widehat{b})$ implies that the intervention on C be potentially dependent on a previous intervention on B, which seems to contradict the very idea of an intervention. In chosen applications however, one may decide that interventions are independent, and that $p(\widehat{c} \mid \widehat{b}) = p(\widehat{c})$. When the terms differ, the intervention on C can be thought of as a partial intervention, breaking all causal links pointing to C but some stemming from B.

[3] The Shannon entropy of a source emitting sequences of length l asymptotically tends towards the expected Kolmogorov complexity (see next section) of the sequences as $l \to \infty$. Potential causal specificity and expected Kolmogorov complexity thus go hand in hand. We leave to future work to make this connection more explicit (see Grünwald and Vitányi 2003, p. 518; Li and Vitányi 2008, p. 187; Balduzzi 2011).

Still, in an actual case, the number of alternatives can be zero, and a very long DNA sequence can therefore have null causal specificity for its own transcript. Causal specificity represents a sense of information which enables us (or causes the system) to choose between a set of well-defined alternatives with a well-defined probability distribution. This is 'information as choice'. If what we are interested in is the information content of a single object, another branch of information theory, Kolmogorov complexity, is more appropriate. It is to this second aspect of information that we now turn.

3 Kolmogorov meets Crick: information as construction

Kolmogorov complexity can measure the complexity of a single object (Grünwald and Vitányi 2003; Li and Vitányi 2008). The intuitive idea is that the more complex the object, the longer its description needs to be. The Kolmogorov complexity is the length of the shortest description enabling one to reconstruct the object using a computer (or more precisely a universal Turing machine).[4] Kolmogorov complexity also provides a measure of the amount of information in an object *about* another object. This is measured by the algorithmic mutual information: it is the amount of program length that one saves when describing one object given a description of the other object for free. Algorithmic mutual information is symmetrical.

Obviously, the length of the shortest description will depend not only on the object at stake, but also on the language (the description method) used to write the program generating the object. However, the lengths of the shortest descriptions in two different languages will be the same up to a given translation constant, independent of the object itself. This is because the translation from one language to another can be described by a program, of which the length is fixed (which gives the constant of translation). In this sense, the Kolmogorov complexity is an objective property of the object.

A drawback of Kolmogorov complexity is that it is provably uncomputable – there is no computer program which, given any string as an input, returns its Kolmogorov complexity as an output. In itself, this is an interesting negative result: if what we are interested in is complexity in this sense, then what we want to know is simply uncomputable. In

[4] A Turing machine consists of a finite program capable of manipulating a linear list of cells (each containing a symbol, from a finite set of symbols), accessing one cell at a time. A universal Turing machine is one that can imitate any other Turing machine (Li and Vitányi 2008, p. 24) .

practice, one can bound the complexity of binary objects using diverse lossless compression methods (e.g. those used in the ZIP file format). Indeed, the compressed object is a (hopefully shorter) description enabling one, together with a decompression program, to reconstruct the initial object. The length of description is then the length of the compressed file plus a constant, the length of the decompression program. This measurement is tentative, not definitive, as other, potentially unknown compression methods might compress the object more. For the sake of the argument, we assume for the moment that we are given a reasonable compression method.

Kolmogorov complexity can be used to explore what Crick meant when he described the determination of proteins by nucleic acids as the "detailed residue-by-residue transfer of sequential information" (Crick 1970, p. 561), where nucleotides would form a quaternary alphabet and amino-acids a vigesimal one. Two kinds of questions can be addressed: about how much information there is in a given biological object, and, closer to Crick's thinking, about how much information in an object comes from another.[5]

The complexity of a strand of DNA, for instance, can be approached by measuring the length of the compressed sequence. Telomeres provide an interesting limit-case. They are nucleotide sequences at the end of chromosomes, consisting of a repetitive pattern (e.g. TTAGGG in humans, and many other species). Telomeres are elongated by an enzyme, called telomerase, which embeds an RNA-sequence as a template (Hiyama et al. 2009). It is not difficult to come up with a program describing a given telomeric sequence in a compact way. Whatever the length of a telomere, it can be described by a template for the repeated pattern and the number of repeats (Figure 1, Algorithm 1). A naive observer would surely think that telomeres do not contain much information, and in particular not much *sequential* information. The intuition here coincides with the low Kolmogorov complexity of these sequences.

The situation looks quite different for coding sequences. There does not seem to be, at first sight, as easy a way to compress these sequences as we did with telomeres, and their Kolmogorov complexity is probably substantially higher: a program to reconstruct a coding sequence may have to spell it out explicitly – or at least to spell out significant aspects of the sequence (Fig. 1, Alg. 2). This lower compressibility coincides with the intuition that coding sequences carry sequential information – and even that it is their function to carry sequential information. However,

[5] On the algorithmic approach in biology, see e.g. Yockey (2005, p. 170) and especially Chaitin (1979; 2012), and the discussion by Artmann (2008, pp. 32-37). We lack space to review the independent convergences and divergences of the here proposed account.

coding sequences do not carry a maximal amount of sequential information: as an anonymous reviewer noticed, coding sequences are structured, and are thus expected to be compressible to some extent – as are non-coding, so-called 'junk' DNA-sequences containing a significant number of repetitive elements and duplications.[6] Note that the intrinsic amount of information in a sequence is independent of whether the sequence is inserted in a region which will actually undergo transcription. Arguably, even a coding sequence carries no information *about* any transcript if it is not transcribed, but it nevertheless carries sequential information *tout court*.

We now turn to the second question, asking how much information there is in an object *about* another object. For the sake of the argument, we suppose that the world is as Crick supposed in 1958: the accuracy of information transfers is high, which we idealize by assuming that transcription (of DNA into RNA) and translation (of RNA into polypeptides) are error-free, deterministic processes. We ignore splicing and other post-transcriptional processes, which will be treated elsewhere. As described above, one can estimate the amount of information in DNA about RNA by their algorithmic mutual information, that is: $I(DNA : RNA) = K(RNA) - K(RNA \mid DNA^*)$.[7] The shared information between DNA and RNA is substantial: the transcription process is all about replacing the nucleotides by their complementary ones, with the proviso that As in the coding DNA-sequence are complementary to Us (not Ts) in the RNA-sequence. To see this sharing of information, compare the lengths of an algorithm spelling out the RNA explicitly (similar to Alg. 2) and one treating transcription generically (Fig. 1, Alg. 3). The difference in length would increase with sequence length. This corresponds to the fact that sequential information is transferred from DNA to RNA through transcription. If transcription is errorless, the sequential information in RNA which does not come from DNA, measured by the remainder complexity $K(RNA \mid DNA^*)$, is a constant, independent of the sequence. Algorithmic mutual information between biological sequences has been used in the past decade with various aims, such as the building of phylogenetic trees according to the amount of information needed to transform one DNA sequence into another (e.g. Chen et al. 2000; Li et al. 2001; Chen et al. 2002; Vinga 2014, for a review).

Since the 'true' Kolmogorov complexity is uncomputable, an algorithmic approach relies on a bet – that the language of description and compression methods capture interesting and relevant aspects of the object at stake. This is not to say that the approach is necessarily entirely

[6] Whether the compressibility of sequences is an inevitable aspect of their biological function is precisely a question we wish to address in the long term.

[7] Where DNA^* is the shortest program generating DNA.

arbitrary: once these methods are agreed on, researchers can agree on the measures obtained for finite sequences. If a particular language gives particularly interesting results (e.g. saving biological appearances, leading to new questions, predictions and generalizations), then this language becomes a theoretical entity worth discussing in its own rights. In the remainder of the paper, we outline what we deem desirable features for such a language, and substantially develop the algorithmic approach to take into account the fact that biological systems are not, strictly speaking, deterministic, universal Turing machines. What we aim at is not an application of conventional algorithmic information theory to biology, but a specifically biological approach to information inspired by the Kolmogorov branch of information theory.

Algorithm 1: Function synthesizing a telomeric sequence of a defined length.

```
Function SynthesizeTelomere(n, RNA-template)
    input  : $n$: number of iterations,
             RNA-template: ‘CCCAAUCCC’
    output : a given telomere
    for $i \leftarrow 1$ to $n$ do
        apply telomerase using RNA-template
    end
end
```

Algorithm 2: Function synthesizing a given (bit of) coding DNA from scratch.

```
Function SynthesizeCodingSequence
    input  : none
    output : a given coding sequence
    synthesize ‘... GCAGTA GAATTCCGAGCAACT GAACGAGCAGTAGAA ... ’
end
```

Algorithm 3: Function synthesizing a RNA given a DNA as an input.

```
Function SynthesizeRNA(DNA)
    input  : a given DNA
    output : a given RNA
    while in DNA do
        transcribe nucleotide
        move forward one nucleotide
    end
end
```

Algorithm 4: Function spelling out transcription in (some) details.

```
Function Transcribe(nucleotide)
    input  : nucleotide
    output : corresponding nucleotide
    if $A$ then return $U$
    else if $C$ then return $G$
    else if $G$ then return $C$
    else if $T$ then return $A$
end
```

Figure 1: Four algorithms illustrating an algorithmic approach to biological functioning.

4 Randomness as the source of information

We made several idealising assumptions in the previous sections. Let us now relax the assumption that cellular processes are deterministic. Our argument here will remain theoretical: we lack room to take sides on whether, and how, randomness is actually realized in biology.[8]

Random events, by definition, cannot be determined in advance by an algorithm. This means that randomness in the generation of a sequence creates information *de novo*. In biological terms, this means that any random point mutation, any error of transcription, etc., if they are genuinely random, can create information in the Kolmogorov sense. As we have seen in the previous section, this also means that randomly generated sequences contain more information than highly structured sequences. From an algorithmic point of view, randomness is, ultimately, the only way to create information.

This information need not always be functional, that is, of any use to the cell. That it may *sometimes* be so, however, is a reasonable assumption. There are several biological examples suggesting that randomness plays a key role in biological functioning (Kupiec 1983; Heams 2014). Gene shuffling in the immune system of jawed vertebrates provides one such example regarding biological sequences. It enables a great variety of antibodies to be produced, orders of magnitude more numerous than the genes producing them, increasing the chance of matching potentially threatening antigens (Cooper and Alder 2006).

This tension between information and function is why it is crucial to distinguish them. One might be interested in how information flows in biological systems without committing oneself to a particular account of biological function. More importantly, if one is interested in whether and how information leads to function, a concept of biological information as necessarily biologically functional will beg the question.

5 A language for the cell

Kolmogorov complexity allowed us to flesh out the idea of information as construction. Now we need to kick that ladder away and ask what information as construction actually looks like in living systems. We suggest that it ought to be measured using a particular programming

[8] Doing so properly would require developing an account of measurement in biology, as those developed for deterministic chaos and quantum indeterminacy. In deterministic chaos, any finite measurement of the initial condition will leave aside information (the amount of which is infinite) which will manifest itself after a certain time in the system (Montévil 2018, § 2.1, 2.3). Quantum indeterminism represents another entry to a physicalist view of the appearance of information (Stamos 2001, and the responses).

language: the language of the cell itself, in which available programming functions mimic actual operations by which molecules are produced. It goes without saying that what we evoke here is not the 'true' language, but a model of a language of the cell.

The idea of a language of the cell takes us away from treating cells as universal Turing machines and from the genuine Kolmogorov complexity K, to consider a more biological algorithmic complexity, the Kolmogorov complexity in the chosen biological language (hereafter denoted K_B). For instance, algorithmic mutual information is symmetric – there is as much information in DNA about RNA as there is in RNA about DNA, i.e. $I(RNA : DNA) = I(DNA : RNA)$. But not all operations are possible in a cell. A central feature of molecular biology is that flows of information are asymmetrical. Crick's 'central dogma' (still widely held today) states which flows of information between biomolecules are possible and which are not. If no reverse-transcriptase is present, for instance, no information can flow from RNA to DNA. In 'biologically' algorithmic terms, this means that a biological program aiming at reconstructing a DNA-sequence being given an RNA-sequence as an input would fare no better than a program being given no input, and we would obtain $K_B(DNA \mid RNA^*) = K_B(DNA)$. This means that we would get, as for a biological analog of algorithmic mutual information, $I_B(RNA \to DNA) = K_B(DNA) - K_B(DNA \mid RNA^*) = 0$. (The subscript '$_B$' again denotes that the measure is defined using the chosen biological language, and the arrow now reflects that it can be asymmetric.[9]) The reciprocal, as we have seen above, is very different: when DNA is transcribed into RNA, $K_B(RNA \mid DNA^*) = C$, where C is a constant not depending on the sequences. Assuming, for the sake of presentation, that $K_B(RNA) = K_B(DNA)$, we would obtain $I_B(DNA \to RNA) = K_B(RNA) - K_B(RNA \mid DNA^*) = K_B(DNA) - C$. Thus, contrary to its genuine counterpart, biological algorithmic mutual information would not be expected to be always symmetrical, reflecting the directionality of possible information flows.

In the same vein, not all sequences can be produced by a given cell. In algorithmic information theory, a universal Turing machine can emulate any other Turing machine, which means that there is no sequence that a particular machine can produce that a universal machine cannot produce. By contrast, if the cell lacks a programming function, for instance, if it lacks a nucleic acid template, or if some nucleotides do not belong to its alphabet, then some sequences may be impossible to produce. In this case, the information needed to produce an impossible sequence is ill-defined, in other words, the amount of information needed to pro-

[9] We follow the notation used for one asymmetrical, causal version of Shannon mutual information (Ay and Polani 2008).

duce the sequence is indefinite. Even on an evolutionary time-scale, the amount of information needed to acquire the programming function (if it is acquired) and produce the previously-impossible sequence could be orders of magnitude greater than the length of the sequence.

Granted that some operations are impossible, how are we to describe the set of primitive programming functions that, by contrast, are possible?

As we have seen above, the complexity of an object depends on the language used to describe it. An example will flesh out this idea. Assume, say, that 'Transcribe' is given for free by the language, and that the description of the function is short: say, just a few letters. Contrast this with a DNA sequence of several kilobases. This DNA sequence appears much more informational than the function 'Transcribe'.[10] Now, imagine that 'Transcribe' is not given for free by the language, but that one has to write a program for this function, using other, more primitive, available functions. We exemplify such a program in Alg. 4 (Fig. 1) – it could be made much longer by describing explicitly the dynamics of chemical bonds in a binary manner (assuming for the sake of the argument that this would be feasible). Conversely, the description of a long DNA-sequence can be very short. For instance, nominal genes are usually described, not by their full sequence, but by a nickname like 'p53'. This nickname is enough, on most occasions, for biologists to communicate about the processes at stake. A language can lack the function 'Transcribe' but have a built-in function 'P53' dedicated to returning the full sequence of the gene. In such a language, descriptions of transcription would be complex (informational) and that of DNA simple. Thus, one needs to be cautious about the language of description before assigning any particular object a privileged informational role – much in the same way as one needs to be cautious about specifying the probability distributions when using Shannon information theory.

We propose that the primitive functions should be those which enable us to understand the processes of interest. Assume, for instance, that our interest lies in understanding the flows of sequential information between biological polymers. Then assuming that 'Transcribe' and 'Translate' are given as primitive functions is fine: if they are errorless, they are not difference-makers as regards the final sequences of the products (an assumption that we made in Alg. 3). Generally speaking, it makes sense to consider as primitives those operations which are not difference-makers as regards the outputs of interest, and as inputs those very difference-makers: the genericity for functions, and the specificity

[10] Many biologists and some philosophers routinely ascribe to DNA a privileged informational role. One way to reconstruct this idea is to consider that they implicitly assume such a language.

for inputs. Incidentally, it is good algorithmic practice to write functions for generic operations and give them specific variables as inputs. This is not unlike causal specificity: once the generic functional relationships in the causal model are set, information flows from difference-makers.

6 Payoff of the approach

The algorithmic approach sketched above may promote research on biological systems, although not without significant challenges.

Even the best current specifically designed compression algorithm may overestimate biological complexity. This is because the algorithm may not have compressed the object sufficiently. On the positive side, compression algorithms may actually tend to parallel the biological processes that have produced the sequences at stake. For instance, if DNA translocation is frequent, then an algorithm that pays due attention to translocation should be more likely to compress a DNA-sequence. Conversely, considering that most strings are random in the algorithmic sense, it is highly unlikely that a series of refined algorithms will converge, if they do convergence, towards something other than the processes involved in producing the sequences. It is highly unlikely that the cell will, by chance, produce a string which is compressible by other means than some of its own means of production (or the corresponding models of these means). In other words, improving these algorithms may yield a better grasp of functions which are in fact available in the language of the cell.

Just as an algorithm may overestimate complexity, however, it can also *underestimate* biological complexity. Because cells are not universal Turing machines, a biological sequence may be more complex than its algorithmic counterpart. For instance, a cell may need a complex process to resist random perturbations when duplicating a sequence, while a universal Turing machine, being deterministic, would not. Similarly, a short sequence may require a complex machinery or a complex evolutionary history to produce it. Just as biological complexity can shortcut algorithmic complexity (when a cell generates randomness), it can also exceed it.

7 Conclusion

This paper aimed to give substance to the idea of biological information – an idea that has grounded significant aspects of informal biological thought for the past fifty years. Crick's seminal use, in molecular biology, of the term 'information', meaning the precise determination of sequence, is grounded on causation, not meaning or representation. We inflected this idea in two ways, corresponding to two aspects of information theory: the precise determination of a single output from a set of alternatives ('information as choice'), and the precise determination of the sequence of a single output ('information as construction'). These two aspects can be traced back to Crick, whose idea of information as construction – to rephrase in our terms – was an attempt to provide an explanation of information as choice, in the sense of biological specificity (Crick 1958; Crick 1970). This suggests that Griffiths' and Stotz's (2013) idea of distributed specificity is theoretically richer than initially envisioned.

Information as choice is captured by causal specificity, proposed elsewhere to be measured by the Shannon mutual information between values of a cause set by an intervention and observations of the effect. This measure can be applied to causal graphs, such as those representing gene regulatory or animal signalling networks, and has numerous potential applications in biology.

Information as construction is captured by the Kolmogorov complexity of a sequence and the algorithmic mutual information between two sequences. These measures capture the intuition that there is something in common between a program generating a sequence and the biological processes of transcription and translation. We insisted, however, that there is more to biology than discrete, deterministic computing: randomness plays a central role in biological functioning. A similar point could be made regarding the non-discrete nature of biological phenomena. From the point of view of Kolmogorov complexity, randomness creates information. Such information is not necessarily functional, and distinguishing between information and function is a necessary step towards better understanding how information can lead to function.

We proposed that biological algorithmic complexity ought to be measured using a biologically relevant programming language – the language in which the cell performs its own operations. In such a language, some operations, such as reverse-translation, will be impossible. This means that the biological complexity of a sequence can far exceed its own length – making it very different from non-biological algorithmic complexity. In planned future work, we will take up the challenge of fleshing out the 'language of the cell', and articulating the choice and construction aspects of biological information.

References

Alberts, Bruce et al. (2013). *Essential Cell Biology, Fourth Edition.* Garland Science. 863 pp.

Artmann, Stefan (2008). “Biological information”. In: *A Companion to the Philosophy of Biology*. Ed. by Sahotra Sarkar and Anya Plutynski. John Wiley & Sons.

Ay, Nihat and Daniel Polani (2008). “Information flows in causal networks”. In: *Advances in complex systems* 11.1, pp. 17–41.

Balduzzi, David (2011). “Information, learning and falsification”. In: *arXiv preprint arXiv:1110.3592*.

Barbieri, Marcello (2007). *Introduction to biosemiotics: The new biological synthesis*. Springer Science & Business Media.

Calcott, Brett, Arnaud Pocheville, and Paul E. Griffiths (2017). “Signals that make a difference”. In: *The British Journal for the Philosophy of Science*.

Chaitin, Gregory J. (1979). “Toward a mathematical definition of “life””. In: *The Maximum Entropy Formalism*. Ed. by R. D. Levine and M. Tribus. MIT Press, pp. 477–498.

— (2012). *Proving Darwin: making biology mathematical*. Vintage.

Chen, Xin, Sam Kwong, and Ming Li (2000). “A compression algorithm for DNA sequences and its applications in genome comparison”. In: *Proceedings of the fourth annual international conference on Computational molecular biology*. ACM, p. 107.

Chen, Xin, Ming Li, Bin Ma, and John Tromp (2002). “DNACompress: fast and effective DNA sequence compression”. In: *Bioinformatics* 18.12, pp. 1696–1698.

Cooper, Max D. and Matthew N. Alder (2006). “The Evolution of Adaptive Immune Systems”. In: *Cell* 124, pp. 815–822.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of information theory*. John Wiley & Sons.

Crick, Francis (1958). *On Protein Synthesis*. Symposium of the Society for Experimental Biology XII. New York: Academic Press.

— (1970). “Central dogma of molecular biology”. In: *Nature* 227.5258, pp. 561–563.

Dretske, Fred I. (1981). *Knowledge and the Flow of Information*. Basil Blackwell. 273 pp.

Godfrey-Smith, P. (2000). “On the Theoretical Role of "Genetic Coding"”. In: *Philosophy of Science*, pp. 26–44.

Griffiths, Paul E., Arnaud Pocheville, Brett Calcott, Karola Stotz, Hyunju Kim, and Rob Knight (2015). “Measuring Causal Specificity”. In: *Philosophy of Science* 82.4, pp. 529–555.

Griffiths, Paul E and Karola Stotz (2013). *Genetics and Philosophy: An introduction*. New York: Cambridge University Press.

Grünwald, Peter D. and Paul M. B. Vitányi (2003). “Kolmogorov Complexity and Information Theory. With an Interpretation in Terms of

Questions and Answers". In: *Journal of Logic, Language and Information* 12.4, pp. 497–529.

Heams, Thomas (2014). "Randomness in biology". In: *Mathematical Structures in Computer Science* 24.3.

Hiyama, Keiko, Eiso Hiyama, and Jerry W. Shay (2009). "Telomeres and Telomerase in Humans". In: *Telomeres and Telomerase in Cancer*. Ed. by Keiko Hiyama. Springer Science & Business Media.

Jablonka, Eva (2002). "Information: its interpretation, its inheritance, and its sharing". In: *Philosophy of Science* 69.4, pp. 578–605.

Jacob, François (1970). *La logique du vivant: une histoire de l'hérédité*. Gallimard. 362 pp.

Janzing, Dominik, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf (2013). "Quantifying causal influences". In: *The Annals of Statistics* 41.5, pp. 2324–2358.

Kay, Lily E. (2000). *Who Wrote the Book of Life?: A History of the Genetic Code*. Stanford University Press. 476 pp.

Korb, Kevin B, Lucas R Hope, and Erik P Nyberg (2009). "Information-Theoretic Causal Power". In: *Information Theory and Statistical Learning*. Ed. by Frank Emmert-Streib and Matthias Dehmer. Boston, MA: Springer US, pp. 231–265.

Kupiec, J. J. (1983). "A probabilist theory for cell differentiation, embryonic mortality and DNA C-value paradox". In: *Speculations in Science and Technology* 6.5, pp. 471–478.

Li, Ming, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul Kearney, and Haoyong Zhang (2001). "An information-based sequence distance and its application to whole mitochondrial genome phylogeny". In: *Bioinformatics* 17.2, pp. 149–154.

Li, Ming and Paul M. B. Vitányi (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd edition. Springer Science & Business Media. 809 pp.

Lizier, J. T. and M. Prokopenko (2010). "Differentiating information transfer and causal effect". In: *The European Physical Journal B* 73.4, pp. 605–615.

Maynard Smith, John and Eors Szathmary (1995). *The Major Transitions in Evolution*. Oxford University Press. 362 pp.

— (2000). *The Origins of Life: From the Birth of Life to the Origin of Language*. OUP Oxford. 179 pp.

Mayr, Ernst (1961). "Cause and effect in biology". In: *Science* 134, pp. 1501–1506.

Montévil, Maël (2018). "Possibility spaces and the notion of novelty: from music to biology". In: *Synthese*, pp. 1–27.

Pearl, Judea (2009). *Causality: Models, Reasoning, and Inference*. 2nd Edition. New York: Cambridge University Press. 486 pp.

Pocheville, Arnaud. “New tools for thinking about causation”. In: *Manuscript.*

— “Signals that represent the world”. In: *Manuscript.*

Pocheville, Arnaud, Paul E. Griffiths, and Karola Stotz (2017). “Comparing causes – an information-theoretic approach to specificity, proportionality and stability”. In: *Proceedings of the 15th Congress of Logic, Methodology and Philosophy of Science.* 15th Congress of Logic, Methodology, and Philosophy of Science. Ed. by Hannes Leitgeb, Ilkka Niiniluoto, Elliott Sober, and Seppälä, Päivi. London: College Publications.

Pocheville, Arnaud and Maël Montévil. “Giving substance to biological information”. In: *Manuscript.*

Sarkar, Sahotra (1996). “Biological information: A skeptical look at some central dogmas of molecular biology”. In: *The Philosophy and History of Molecular Biology: New Perspectives.* Ed. by Sahotra Sarkar. Vol. 183. Boston Studies in the Philosophy of Science. Dordrecht: Kluwer Academic Publishers, pp. 187–232.

Shannon, Claude (1948). “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* 27, pp. 379–423.

Shea, Nicholas (2007). “Representation in the genome and in other inheritance systems”. In: *Biology & Philosophy* 22.3, pp. 313–331.

Stamos, David N. (2001). “Quantum indeterminism and evolutionary biology”. In: *Philosophy of Science* 68.2, pp. 164–184.

Šustar, Predrag (2007). “Crick’s Notion of Genetic Information and the ’Central Dogma’ of Molecular Biology”. In: *The British Journal for the Philosophy of Science* 58.1, pp. 13–24.

Tononi, Giulio, Olaf Sporns, and Gerald M. Edelman (1999). “Measures of degeneracy and redundancy in biological networks”. In: *Proceedings of the National Academy of Sciences* 96.6, pp. 3257–3262.

Vinga, Susana (2014). “Information theory applications for biological sequence analysis”. In: *Briefings in Bioinformatics* 15.3, pp. 376–389.

Walker, Sara Imari, Paul C. W. Davies, and George F. R. Ellis (2017). *From Matter to Life: Information and Causality.* Cambridge University Press. 517 pp.

Weber, Marcel (2017). “Discussion Note: Which Kind of Causal Specificity Matters Biologically”. In: *Philosophy of Science* 84.3, pp. 574–585.

Williams, George C. (1992). *Natural Selection: Domains, Levels, and Challenges.* Oxford University Press. 222 pp.

Woodward, James (2010). “Causation in biology: stability, specificity, and the choice of levels of explanation”. In: *Biology & Philosophy* 25.3, pp. 287–318.

Yockey, Hubert P. (2005). *Information theory, evolution, and the origin of life.* Cambridge University Press.