PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: Marshall SM, Murray ARG, Cronin L. 2017 A probabilistic framework for identifying biosignatures using Pathway Complexity. *Phil. Trans. R. Soc. A* **375**: 20160342. http://dx.doi.org/10.1098/rsta.2016.0342

Accepted: 18 June 2017

One contribution of 18 to a theme issue 'Re-conceptualizing the origins of life'.

Subject Areas:
astrobiology, algorithmic information theory, complexity, synthetic chemistry

Keywords:
complexity, biosignature, Pathway Complexity, living–non-living threshold

Author for correspondence:
Leroy Cronin
e-mail: lee.cronin@glasgow.ac.uk

A probabilistic framework for identifying biosignatures using Pathway Complexity

Stuart M. Marshall, Alastair R. G. Murray and Leroy Cronin

School of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK

LC, 0000-0001-8035-5757

One thing that discriminates living things from inanimate matter is their ability to generate similarly complex or non-random structures in a large abundance. From DNA sequences to folded protein structures, living cells, microbial communities and multicellular structures, the material configurations in biology can easily be distinguished from non-living material assemblies. Many complex artefacts, from ordinary bioproducts to human tools, though they are not living things, are ultimately produced by biological processes—whether those processes occur at the scale of cells or societies, they are the consequences of living systems. While these objects are not living, they cannot randomly form, as they are the product of a biological organism and hence are either technological or cultural biosignatures. A generalized approach that aims to evaluate complex objects as possible biosignatures could be useful to explore the cosmos for new life forms. However, it is not obvious how it might be possible to create such a self-contained approach. This would require us to prove rigorously that a given artefact is too complex to have formed by chance. In this paper, we present a new type of complexity measure, which we call 'Pathway Complexity', that allows us not only to threshold the abiotic–biotic divide, but also to demonstrate a probabilistic approach based on object abundance and complexity which can be used to unambiguously assign complex objects as biosignatures. We hope that this approach will not only open up the search for biosignatures beyond the Earth, but

THE ROYAL SOCIETY PUBLISHING

also allow us to explore the Earth for new types of biology, and to determine when a complex chemical system discovered in the laboratory could be considered alive.

This article is part of the themed issue 'Re-conceptualizing the origins of life'.

1. Introduction

(a) Biosignatures

There have been many proposals for finding effective biosignatures, that is, unambiguous indicators of the influence of life in an environment. These include searching for atmospheric gases such as methane [1], looking for signs of a distinctive $^{56/54}$Fe isotope ratio [2], and searching for a biological impact on minerals and mineral assemblages [3], for fossils [4] or for distinctive patterns in the distribution of monomer abundance [5]. It has also been suggested that life on exoplanets could be detected by searching for a variant of the distinctive 'red-edge' spectral feature of the Earth [6], where there is a strong increase in reflectance in the 700–750 nm region of the spectrum due to vegetation. In this case, we cannot necessarily expect alien vegetation to share the spectral characteristics of terrestrial vegetation, but perhaps a spectroscopic signature could be observed at another wavelength. Additionally, extreme care should be taken to avoid misidentifying this effect with similar effects that can be caused by certain mineral formations. These two caveats highlight particular difficulties in trying to classify phenomena as biosignatures. The first difficulty is in ensuring that we cast the net wide enough to include biologies that may well differ fundamentally from our own. By remaining too tied to the details of terrestrial biology, we risk missing biosignatures presented to us due to our assumptions about what life must be like. The second difficulty is in avoiding false positives by ensuring that abiotic causes are ruled out. For example, shortly after the $^{56/54}$Fe ratio was suggested as a biosignature in 1999 [2], Bullen *et al.* [7] published in 2001 evidence that the same isotopic fragmentation could have an abiotic origin. In another example, a 2002 paper declared that magnetite crystals within Martian meteorite ALH84001 were 'a robust biosignature' [8]; however, potential abiotic processes to create such crystals have also since been proposed [9,10].

(b) Complexity measures

Discussion of the concept of complexity can be difficult, as there is currently no consensus on a single unambiguous definition of complexity [11]. In addition, descriptions of complexity and randomness are intrinsically related, and many definitions of complexity are specific to certain fields or applications and depend on an often-biased observer, which can lead to comparisons between intrinsically different things. This is a problem because it can result in misleading notions about which objects are more complex. We will describe below some existing measures of complexity, although, as there are a great many such measures, the list is far from exhaustive, and it is beyond the scope of this paper to make a detailed comparison of 'Pathway Complexity' with established complexity measures. Several complexity measures find utility in the realms of computation and information. In information theory, the 'Shannon entropy' [12] of a string of unknown characters, which can be used as a complexity measure, is a measure of how predictable the outcome of the string is, or equivalently how much information it contains, based on the probability of each possible character in the string. The Kolmogorov complexity of a known object [13] is the minimum necessary length of a programme that outputs the object, where, for example, strings containing many repetitions, or programmable patterns, would have lower complexity than those that are random. Logical depth [14] is a complexity measure closely related to Kolmogorov complexity, but is a measure of the number of computational steps required to generate the object from the shortest (or almost shortest) programme, rather than the size of the programme itself. Effective complexity [15] looks for a compressed description

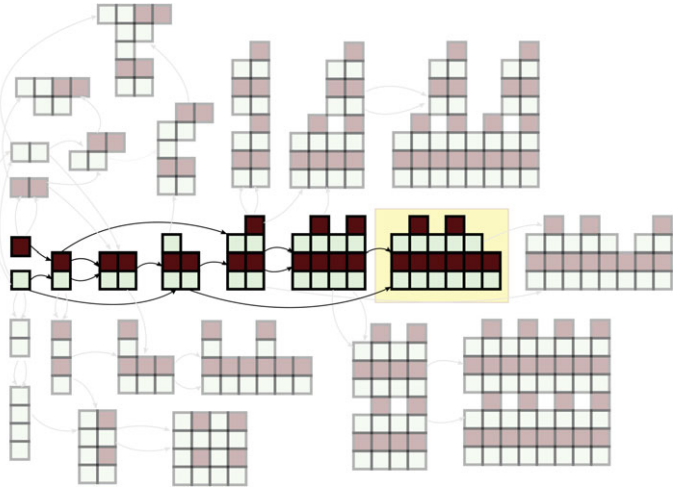

Figure 1. Illustration of a complexity pathway in blocks, with the target shown by the yellow box. A combinatorial explosion in structures is illustrated by the other faded structures shown, which are just a small set of the many alternative structures that could be constructed. (Online version in colour.)

of the regularities of an object. One can also examine the computational complexity [16] of an algorithm, which gives a measure of how the resources required increase with the size of the input. Stochastic complexity is another similar measure, but which looks at the shortest encoding of the object taking advantage of probabilistic models [17]. Measures such as Shannon entropy and Kolmogorov complexity are maximum for random structures, although one can argue that randomness is not necessarily complexity, and that maximum complexity lies somewhere between completely ordered and random structures [11].

There have been a number of suggestions for complexity measures on molecules [18] or crystal structures [19]. These range from those based on information-theoretic measures, to specific features of the chemical graph such as vertex degrees [20] and the number of subgraphs of the molecular graph [21]. In other applications, measures for complexity in graph theory [22], in tile self-assembly [23] and in biology in relation to genes and their environment have been proposed [24]. Here, we present the concept of 'Pathway Complexity', which identifies the shortest pathway to assemble a given object by allowing the object to be dissected into a set of basic building units and rebuilding the object using those units. Thus the Pathway Complexity can be seen as a way to rank the relative complexity of objects made up of the same building units on the basis of the pathway, exploiting the combinatorial nature of these combinations (figure 1).

2. Results and discussion

(a) Pathway Complexity as a biosignature

We propose a measure of complexity based on the construction of an object through joining operations, starting with a set of basic substructures, where structures already built in the process can be used in subsequent joining operations. The sequence of joining operations that constructs the objects can be defined as a complexity pathway, and the number of associated joining operations is defined as the complexity of that pathway. The complexity of the object with respect to the set of substructures is defined as the lowest complexity of any pathway that builds the object. We call this complexity measure 'Pathway Complexity' and it is illustrated in figure 2.

The motivation for the formulation of Pathway Complexity is to place a lower bound on the likelihood that a population of identical objects could have formed abiotically from an initial pool

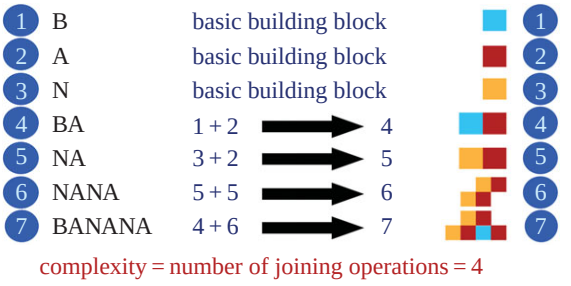

Figure 2. Complexity pathways for a text string and a simple block shape giving the Pathway Complexity to construct the word BANANA from its basic building blocks as 4. (Online version in colour.)

of starting materials, i.e. *without the influence of any biological system or biologically derived agent*. An object of sufficient complexity, if formed naturally, would have its formation competing against a combinatorial explosion of many other possible structures. If that given object was found in abundance, it would be a clear indication that life-like processes were required to navigate in the state space to that particular structure, rather than diffusing the starting materials through the state space and ending up with a diverse mixture of structures that may or may not contain the structure in question. The ability to reliably produce specific non-trivial trajectories through state space is, we propose, a characteristic unique to living systems. Therefore, if we can use Pathway Complexity to place a lower bound on the threshold above which a trajectory becomes non-trivial, we can then establish whether an object is undoubtedly of biological origin.

By following this reasoning, it can be proposed that living systems themselves are self-sustaining non-trivial trajectories in a state space. This means that the biosignatures produced by living systems are themselves non-trivial trajectories [25]. As such, Pathway Complexity bounds the likelihood of natural occurrence by modelling a naive synthesis of the object from populations of its basic parts, where at any time pairs of existing objects can join in a single step. In establishing the Pathway Complexity we are asking, in this idealized world, if the number of joins required would be low enough that we could have some population of the desired object rather than being overwhelmed by instances of all the other structures that could be created. Of course, some pathways may be more favoured than others (such as in chemical synthesis), but unless we have special pathways with 100% yield of each substructure on the pathway, then that fact merely pushes back the threshold. If we find anything significantly above the threshold, then this, we propose, is a general biosignature. By searching for complexity alone, whether of molecules, objects or signals, we do not have to make any assumptions about the details of the biology or its relation to our own biology. By using this new approach, we show below that a rigorous framework can be developed to search for agnostic biosignatures.

(b) Pathway Complexity: basic approach

The basic approach for determining the Pathway Complexity of an object is applicable when we are considering the construction of an object in its entirety from defined, basic subunits. The Pathway Complexity is calculated in the context of any possible objects that could be constructed from the same subunits. Later we will extend this approach to assessing the complexity of a class of objects that are not necessarily identical. We represent subunits of the object as vertices, and connections as edges, in a graph G. The vertices of G are grouped into equivalence classes, which, in the basic approach, would mean that subunits in the same equivalence class are identical. There may be multiple types of edges if there are different types of connection in the object. We then construct complexity pathways for G and establish their complexities using the following process. We start with a sequence containing only trivial 'fundamental' graphs representing each unique subunit in the object. A pair of graphs in the sequence is joined by adding one or more edges

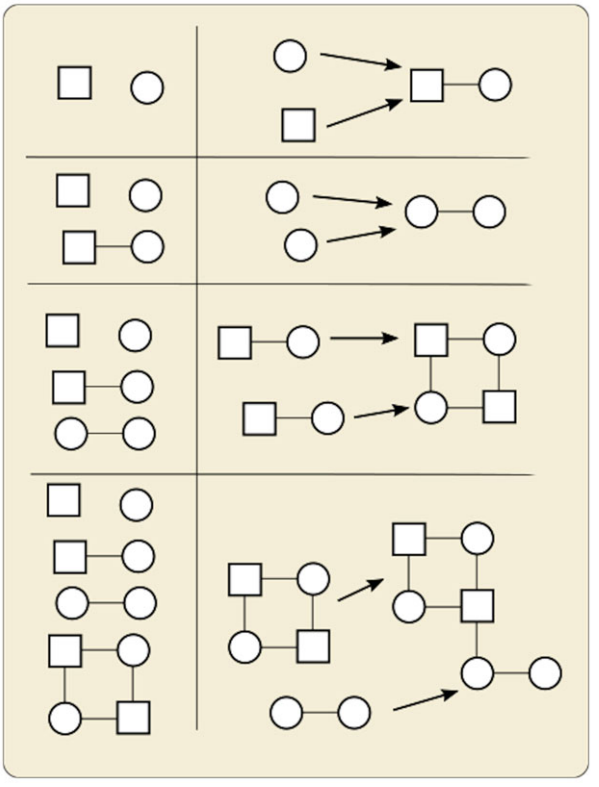

Figure 3. Illustration of a single complexity pathway, with the set of objects on the left and the joining operations on the right. At each step, the structure created by the operation is added to the set of objects and available for subsequent joining operations. (Online version in colour.)

between vertices of one graph and the other (figure 3). A pathway is complete when the sequence contains G, i.e. when the graph of the object has been constructed. The complexity of the pathway is the number of joining operations required to complete the pathway. The Pathway Complexity of G, and of the object with respect to the given substructures, is the smallest number of joining operations of any pathway. G may be a directed graph if the direction of a connection is important. For example, in a text string, different structures will result in connections left-to-right and right-to-left.

We can describe a search tree representing all different pathways, as at each point allowable combinations of different graphs in the set, with different edge types, joined at different combinations of vertices, would follow a pathway down a different branch of the tree, provided the graph resulting from the join is an induced subgraph of G. It should be noted that, while we are conceptually exploring the entire search tree, in practice it is not necessary to explore every pathway as described above, as algorithmic implementations including branch and bound, and other techniques, can reduce the computational burden.

(c) Choice of substructures

The choice of the basic substructures depends on the context of the desired complexity. For example, if we are establishing the complexity of the word 'banana', then we could select the set of unique structures {b, a, n}, where the complexity is relative to all other words that can be made from those three letters. In fact, we could reasonably extend this set to include all letters of the English language plus punctuation, so we could then compare the complexities of any arbitrary phrase in any language using that alphabet. For a chessboard, natural units to choose would be

{black square, white square}, and the complexity would then be relative to all patterns that can be made of black and white squares.

In selecting the set of basic subunits, we need to consider the class of objects that we are comparing. For example, if comparing a polymer to all polymers made of the same types of monomer, then the monomers could be our basic subunits, but if being compared to all molecules in general, then we would be likely to select atom types or bond types instead. Although the choice of basic subunits is context-dependent, a natural choice for the subunit is normally apparent from the nature of the class of objects being examined, which is the minimum set of subunits that would be sufficient to construct any object of that class.

(d) Mathematical formulation for the Pathway Complexity of graphs

The following is a mathematical formulation for establishing the complexity pathway of a graph, as described above.

Definition 2.1. A graph G can be *constructed in one step* from two graphs X and Y if and only if:

— X and Y are disjoint subgraphs of G,
— every vertex in G is in either X or Y, and
— every edge in G is either in X, in Y, or connects a vertex in X with a vertex in Y.

Definition 2.2. A *Complexity Pathway* of a graph G relative to a set of m single-vertex graphs is defined as a sequence of graphs $G_{-m+1}, G_{-m+2}, \ldots, G_0, G_1, G_2, \ldots, G_n$ such that:

— $G_n = G$,
— for $i < 1$, G_i is a single-vertex graph, and
— for $i \geq 1$, G_i can be constructed in one step from two graphs G_j and G_k, with $j, k < i$.

Definition 2.3. The *Pathway Complexity* C of G is the length of the shortest complexity pathway of G, minus the number of single-vertex graphs in that pathway (i.e. n in Definition 2.2). In other words, C is the smallest number of construction steps, as defined in Definition 2.1, that will result in a set containing G.

(e) Characteristics of Pathway Complexity

The Pathway Complexity of an object generally increases with size, but decreases with symmetry, so large objects with repeating substructures may have lower complexity than smaller objects with greater heterogeneity. In addition, the history dependence and recursive nature of the measure mean that internal symmetries are also accounted for if they lie on the shortest pathway. For example, an object may be asymmetric but have a symmetric feature in it that can be constructed through duplication prior to the asymmetric parts being added on. Those duplicated structures may themselves contain substructures with similar duplications, which are accounted for recursively. In this way, we can describe the construction of structures through repeated duplication and addition of subunits.

Pathway Complexity has an upper bound of $N_v - 1$, where N_v is the number of vertices on the graph. This represents joining two fundamental graphs in the first step, and then adding one more at a time until the object is constructed. One lower bound of Pathway Complexity is $\log_2 N_v$, which represents the fact that the simplest way to increase the size of an object in a pathway is to take the largest object so far and join it to itself, e.g. we can make an object of size 2 with one join, 4 with 2 joins, 8 with 3 joins, etc. An illustration of the upper and lower bounds of Pathway Complexity can be seen in figure 4, with the orange regions being forbidden due to the above boundary conditions. The green portion of the figure is illustrative of the location in the complexity space where life might reasonably be found. Regions below can be thought of as being potentially naturally occurring, and regions above being so complex that even living systems might have

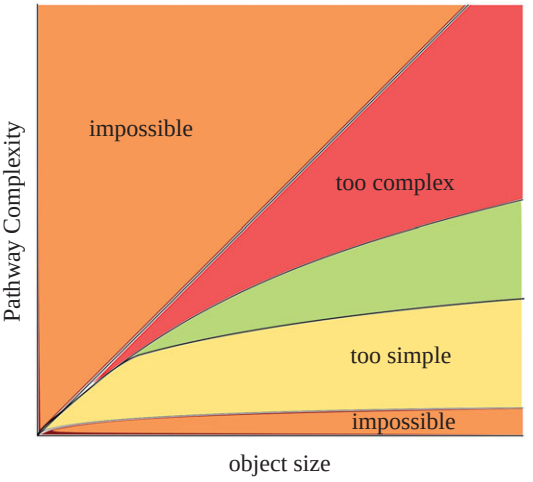

Figure 4. An illustrative graph of complexity against size of the state space. Orange regions are impossible as they are above or below the bounds of the measure. The green region is where living systems may be most probable, where structures are neither too simple to be definitively biological, nor too complex to exist at all.

been unlikely to create them. This is because they represent structures with limited internal structure and symmetries, which would require vast amounts of effort to faithfully reproduce. In exploring this region, we can attempt to find these boundaries, and examine the rate at which living systems can increase their complexity, and the limitations on that increase.

3. Example: text

Pathway Complexity can be used to examine text strings, finding the shortest complexity pathway by leveraging internal regularities. In the following example, we used an algorithm to analyse four strings of text to establish their Pathway Complexity. The following text strings were used, each of them 60 characters long. For consistency, we have converted the strings to lower case without space or punctuation.

(1) A random sequence of letters: 'anpncsaveuoaklkgobqfdfoqtyilrzausbcbsxfclanbipcwizl majbualbs'
(2) Some text from *Green Eggs and Ham* [26] by Dr Seuss: 'iamsamiamsamsamiamthats amiamthatsamiamidonotlikethatsamiamdo'
(3) Some text from *Dracula* [27] by Bram Stoker: 'myfriendwelcometothecarpathiansiam anxiouslyexpectingyousleep'
(4) A highly repetitive sequence: 'redrumredrumredrumredrumredrumredrumredrumred rumredrumredrum'

Intuitively, one would expect the ascending order of complexity to be 4, 2, 3, 1 (with 2 simpler than 3 as *Green Eggs and Ham* is known for its simplicity and repetition). This ordering was confirmed by the algorithm, which found the Pathway Complexity of (4) to be 9, of (2) to be 25, of (3) to be 53 and of (1) to be 57.

The maximum possible Pathway Complexity for any 60-character sequence is 59, so we would expect a random string to have a value close to this. In our random string (1), there are two repetitions in the sequence that the pathway has leveraged to reduce the complexity to 57, which are repetitions of 'an' and 'bs'.

In (3), the passage from *Dracula*, the pathway found has used repetition of 'ec', 'ia', 'an', 'th' and 'ousl'.

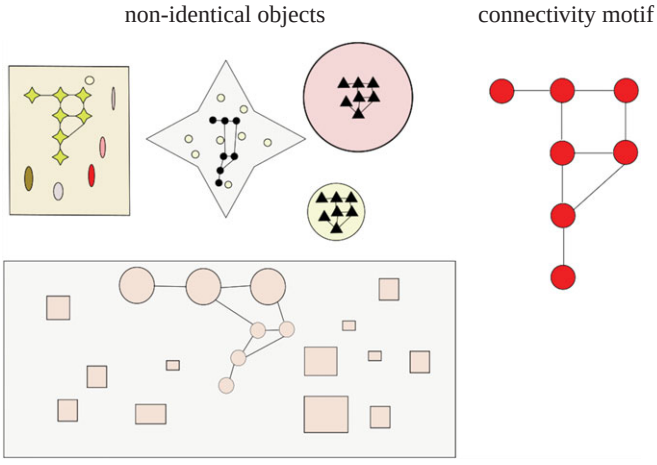

Figure 5. Illustration of the general approach. The same connection motif can be found in all of these shapes, as shown in the graph on the right. Even though the structures and their components are quite different, we can extract the same graph from them and establish its Pathway Complexity. (Online version in colour.)

In (2), the pathway constructs 'am' and then uses that in 'sam' and 'iam'. It then constructs 'samiam' from that pair, and adds letters 't', 'h', 'a', 't', to make 'thatsamiam'. These are then used in the final pathway where 'samiam' and 'thatsamiam' are repeated two and three times, respectively. The pair 'do' is also constructed and used twice in the final pathway.

The pathway in (4) constructs the phrase 'redrum' from individual letters and then duplicates that to make 'redrumredrum', further duplicating that to make 'redrumredrumredrumredrum' and then 'redrumredrumredrumredrumredrumredrumredrumredrum'. Finally, 'redrumredrum' is added to give the result.

4. Pathway Complexity: general approach

We can extend the basic complexity measure above to cope with assessing the complexity of a group of objects that contain identical connection motifs (figure 5). In this case, we examine a population of objects and abstract out a common graph based on connected subunits that share features. For example, if examining a set of cups or mugs, then we can create a common graph of 'handle connected to body', regardless of potential variations in size/colour etc. If examining a set of human beings, then we could create a common graph of bone connectivity, ignoring variations in size/shape of individual bones, or any material in the body other than bones.

5. Choice of subunits and connections

In the general case, we define an archetypal set of connected subunits $S = \{s_1, s_2, s_3, \dots\}$, along with a set of equivalence classes for the subunits $P = \{p_1, p_2, p_3, \dots\}$, and a function $f : S \rightarrow P$. Here, f maps members of S into equivalence classes in P based on defining characteristics for each of the p_i. For example, looking at bones in the human body, there could just be a single class for 'bone' with the characteristic 'made of bone', or different classes for different types of bone distinguished by some characteristic of that type of bone (e.g. tibia, sternum). There may be characteristics of members of S not considered by f and these can be thought of as 'noise'. For example, the same type of bone will vary in shape and size across individuals, but we are only interested in the characteristics that define the mapping onto P. Connections are defined by mapping into another set of equivalence classes $E = \{0, e_1, e_2, \dots\}$ by some function $g : S \times S \rightarrow E$. Here, E contains the 0 element to represent 'not connected', and the e_i represent different types of connection. Here, a

connection could be an actual physical connection, or it could be some more abstract relationship. We then define the archetypal graph G, in which vertices are members of S, with categories $f(s_i)$, and an edge exists between s_i and s_j if $g(s_i, s_j) \neq 0$, with edges categorized by $g(s_i, s_j)$. In the general case, we are looking at a class of objects to which the above rules can be applied to extract a graph isomorphic to G. In this case, members of S are not necessarily substructures that can rebuild an object in its entirety, but rather are shared connection motifs common to a number of objects that we consider to be similar/related.

In the construction of G, it is important that, if a substructure is included in S, all equivalent substructures, as defined by f, are also included in S. This is to prevent overestimation of complexity by selecting a more complex subgraph of G through exclusion of some member of S. For example, one could remove some internal symmetries of a skeleton by selectively erasing some of the bones.

6. Pathway Complexity in the general case

The procedure for constructing complexity pathways on G, and defining the Pathway Complexity of the object, then follows that of the basic approach, only now we are establishing the Pathway Complexity of the selected archetypal graph G that is contained within the whole class of objects. In this way, we can bound the complexity of sets of objects that are non-identical but that clearly share features in such a way that they have some relationship to each other, and establish if the relationship of those motifs exceeds the complexity threshold for a biological source. Greater specificity can provide a higher bound (e.g. specifying the type of each bone in a skeleton, rather than labelling each vertex of the graph as a bone, will result in a higher complexity value).

With this approach, we can examine complex patterns within non-identical structures comprising non-identical parts. As an extreme example, if we were to find sets of entirely different objects (pebbles, bits of wood, etc.) joined by lengths of string on a beach, we could then construct G using 'any object' as a vertex and 'joined by string' as an edge. If the objects were all joined in pairs, then G would be simple and indeed one could imagine plausible physical effects for such a phenomenon. However, if G were particularly complex and abundant, i.e. the same complex pattern were found in multiple locations, one would have to consider that some biological agency was involved. Note here that characteristics such as the lengths of the string or the shapes of the object are not considered—the connected structures could be entirely different sizes and made of completely different things, but the identical complex connectivity motif common to all of them would be enough to make a judgement on the probability of a naturally occurring origin from that perspective alone.

7. Finding threshold between non-biological and biological systems

To assess a reasonable threshold for a given set of objects, we can examine the likelihood of objects of varying complexity being constructed randomly [28]. For example, if we examine a large random text string, and look at the abundance of repeating fragments up to a certain size, we can get some idea of how the abundance of repeated fragments of increasing size drops off as the size increases. To illustrate this point, we have generated a random string of 100 000 characters and plotted the number of repeats of string fragments of different sizes up to size 8 (figure 6). We can see here in plot (*a*) that the number of repeated units drops off dramatically (note that the *y*-axis is logarithmic in plots (*a*) and (*b*)), with very few repeats above length 4 or 5. By splicing in the word 'complex' 1000 times at random positions (plot (*b*)), we dramatically increase the number of repeating units at larger sizes. The difference in the number of repeats can be seen in plot (*c*), with a large difference starting at size 4. From this we can tell that we would expect to find a rather large number of repeats of size 2 and 3, but finding any abundance of repeated strings of size 7 or 8 suggests some internal structure. We can then set a reasonable threshold at 3–5 sigma higher than the random data, suggesting that if we find an abundance of repeats of greater sizes, we have a biosignature.

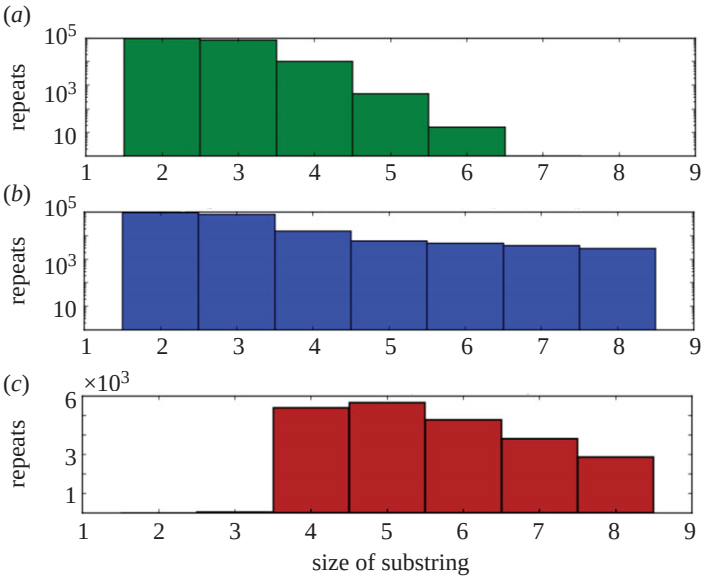

Figure 6. (*a*) Number of repeated substrings of each size in a 100 000 character random text string. (*b*) As in (*a*) but where the same random string has the word 'complex' inserted at random locations 1000 times. (*c*) The difference in the number of repeated substrings between (*b*) and (*a*). (Online version in colour.)

Although a useful indicator, this thresholding exercise does not capture the full details of the complexity measure. A stochastic model is currently in development to give a more accurate assessment of populations of larger structures being found among all the structures that could be created in the same time period. This will then be developed in tandem with an experimental system to investigate if a given dataset contains a biosignature or not.

8. Variant: Recursive Tree Complexity

A variant on the concept of Pathway Complexity as described above is what we have called 'Recursive Tree Complexity'. In this variant, we establish a complexity pathway by partitioning the object graph into a number of different subgraphs. Then the complexity of that pathway is established as the complexity of each unique subgraph, plus the number of times it is duplicated. If the subgraph is a single vertex, then it contributes 1 to the complexity. The procedure is repeated recursively on the unique subgraphs, while adding 1 to the complexity for each time they are duplicated, and the entire structure will eventually be broken down to single-vertex graphs. The partitioning which gives the lowest total complexity is defined as the Recursive Tree Complexity. The Recursive Tree variant provides a slightly different model of the natural construction of objects. In this variant, the parts that have come together to make a particular substructure cannot be leveraged to create multiple completely different structures. In the Recursive Tree variant, one can think of different structures developing separately and then being brought together, rather than all structures being available at all times in one pool. Any pathway in the Recursive Tree variant can also be made by the Pathway Complexity process, but it may not be the shortest pathway and may include redundant steps. Thus Recursive Tree Complexity is an upper bound for Pathway Complexity. Note also that because in Pathway Complexity the first step is a joining step, but in the Recursive Tree variant we effectively lay down a single fundamental structure first, the equivalent pathways in the Recursive Tree variant will be 1 greater than in the Pathway Complexity measure, and this must be accounted for when comparing the two.

Figure 7 illustrates an example of the difference between Pathway Complexity and the Recursive Tree variant. In the former (*a*), the substring 'CO' can be constructed and then used

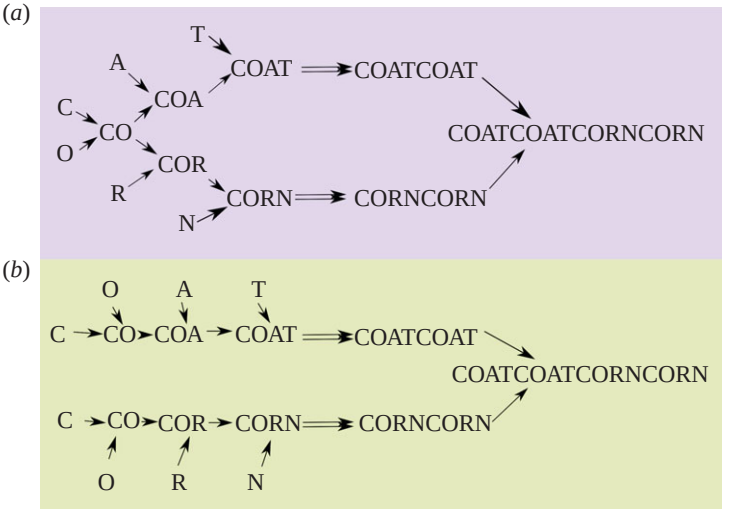

Figure 7. The Pathway Complexity measure (*a*) and Recursive Tree variant (*b*) are used to construct the phrase 'COATCOATCORNCORN'. (Online version in colour.)

to make both 'COA' and 'COR'. In the Recursive Tree variant (*b*), such sharing is not allowed, and the 'CO' that goes to make 'COA' is constructed separately from that which goes to make 'COR'.

A general mathematical formulation of the Recursive Tree variant is given below:

A **single complexity measure** c_P of an object partitioned into a given multiset P is

$$c_P = \sum_{i=1}^{|K|} C_{K_i} + D(P),$$

where K is the set of unique objects in P, C_{K_i} is the Recursive Tree complexity of the ith member of K and $D(P)$ is a function of the multiplicity of the objects in P (initially, $D(P)$ is the total number of duplicated objects, or formally $D(P) = \sum_{i=1}^{|P|} (|P_i| - 1)$).

The **recursive tree complexity** C of an object is equal to 1 if the object cannot be partitioned further (it is atomic/its graph is a single vertex); otherwise it is equal to the minimum single complexity measure

$$C = \begin{cases} 1, & |P| = 1 \; \forall \, P, \\ \min c_P, & \text{otherwise.} \end{cases}$$

9. Conclusion

It is clear that biological and biologically derived systems have an ability to create complex structures, whether proteins or iPhones, that is not found elsewhere in nature. Assessing the complexity of such artefacts will be instrumental in searching for undiscovered biospheres, either on Earth [29] or elsewhere in the Solar System, and would make no assumptions about the details of the biology found. We propose Pathway Complexity as the natural measure of complexity for the production of artefacts. In this context, we argue that there is a critical value of Pathway Complexity above which all artefacts must be biologically derived. This approach provides a probabilistic context to extending the physical basis for life detection proposed by Lovelock [30]. In further work, we will show how this applies to a range of other systems, and propose a series of experimental approaches to the detection of objects and data that could be investigated as a possible biosignature. In the laboratory, we are interested in using this approach to develop a system that can explore the threshold between a non-living and living system. Pathway Complexity may also allow us to develop a new theory for biology. This might inform a

new way to search for life in the laboratory in terms of the complex products a system produces and if they could have arisen in any abundance by chance, rather than trying to measure the intrinsic complexity of the living system itself.

Data accessibility. This article has no additional data.

Authors' contributions. L.C. conceived the initial concept and developed the idea with S.M.M., who then together built the algorithm. S.M.M. and A.R.G.M. expanded the concept and did the simulations. All the authors wrote the paper together.

Competing interests. We declare that we have no competing interests.

Funding. We gratefully acknowledge financial support from the EPSRC for funding (grants EP/P00153X/1; EP/L0236521/1; EP/J015156/1), The John Templeton Foundation Grant ID 60625, EVOBLISS EC 611640, ERC (project 670467 SMART-POM) and the University of Glasgow.

Acknowledgement. We thank Dr Alon Henson for useful discussions.

References

1. Oze C, Jones LC, Goldsmith JI, Rosenbauer RJ. 2012 Differentiating biotic from abiotic methane genesis in hydrothermally active planetary surfaces. *Proc. Natl Acad. Sci. USA* **109**, 9750–9754. (doi:10.1073/pnas.1205223109)
2. Beard BL, Johnson CM, Cox L, Sun H, Nealson KH, Aguilar C. 1999 Iron isotope biosignatures. *Science* **285**, 1889–1892. (doi:10.1126/science.285.5435.1889)
3. Banfield JF, Moreau JW, Chan CS, Welch SA, Little B. 2001 Mineralogical biosignatures and the search for life on Mars. *Astrobiology* **1**, 447–465. (doi:10.1089/153110701753593856)
4. Cady SL, Farmer JD, Grotzinger JP, Schopf JW, Steele A. 2003 Morphological biosignatures and the search for life on Mars. *Astrobiology* **3**, 351–368. (doi:10.1089/153110703769016442)
5. Dorn ED, Nealson KH, Adami C. 2011 Monomer abundance distribution patterns as a universal biosignature: examples from terrestrial and digital life. *J. Mol. Evol.* **72**, 283–295. (doi:10.1007/s00239-011-9429-4)
6. Seager S, Turner EL, Schafer J, Ford EB. 2005 Vegetation's red edge: a possible spectroscopic biosignature of extraterrestrial plants. *Astrobiology* **5**, 372–390. (doi:10.1089/ast.2005.5.372)
7. Bullen TD, White AF, Childs CW, Vivit DV, Schultz MS. 2001 Demonstration of a significant iron isotope fractionation in nature. *Geology* **29**, 699–702. (doi:10.1130/0091-7613(2001)029<0699:DOSAII>2.0.CO;2)
8. Thomas-Keprta KL, Clemett SJ, Bazylinski DA, Kirschvink JL, McKay DS, Wentworth SJ, Vali H, Gibson EK, Romanek CS. 2002 Magnetofossils from ancient Mars: a robust biosignature in the Martian meteorite ALH84001. *Appl. Environ. Microbiol.* **68**, 3663–3672. (doi:10.1128/AEM.68.8.3663)
9. Golden DC *et al.* 2004 Evidence for exclusively inorganic formation of magnetite in Martian meteorite ALH84001. *Am. Mineral.* **89**, 5–6. (doi:10.2138/am-2004-5-602)
10. Cooper GJT *et al.* 2011 Osmotically driven crystal morphogenesis: a general approach to the fabrication of micrometer-scale tubular architectures based on polyoxometalates. *J. Am. Chem. Soc.* **133**, 5947–5954. (doi:10.1021/ja111011j)
11. Ladyman J, Lambert J, Wiesner K. 2013 What is a complex system? *Eur. J. Phil. Sci.* **3**, 33–67. (doi:10.1007/s13194-012-0056-8)
12. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423. (doi:10.1145/584091.584093)
13. Kolmogorov AN. 1963 On tables of random numbers. *Sankhya Indian J. Stat. Ser. A* **25**, 369–376. (doi:10.1016/S0304-3975(98)00075-9)
14. Bennett CH. 1995 Logical depth and physical complexity. In *The universal Turing machine: a half century survey* (ed. R Herken), pp. 227–257. Vienna, Austria: Springer.
15. Gell-Mann M, Lloyd S. 1996 Information measures, effective complexity, and total information. *Complexity* **2**, 44–52. (doi:10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X)
16. Stock-Meyer L. 1987 Classifying the computational complexity of problems. *J. Symb. Logic* **52**, 1–43. (doi:10.2307/2273858)

17. Rissanen J. 2017 Stochastic complexity and modeling. *Ann. Stat.* **14**, 1080–1100. (doi:10.1214/aos/1176350051)
18. Nikolic S, Trinajstic N, Tolic I. 2000 Complexity of molecules. *J. Chem. Inf. Comput. Sci.* **40**, 920–926. (doi:10.1021/ci9901183)
19. Krivovichev S. 2014 Which inorganic structures are the most complex? *Angew. Chem. Int. Edn.* **53**, 654–661. (doi:10.1002/anie.201304374)
20. Randić M, Plavšić D. 2002 Characterization of molecular complexity. *Int. J. Quantum Chem.* **91**, 20–31. (doi:10.1002/qua.10343)
21. Dehmer M, Barbarini N, Varmuza K, Graber A. 2009 A large scale analysis of information-theoretic network complexity measures using chemical structures. *PLoS ONE* **4**, e8057. (doi:10.1371/journal.pone.0008057)
22. Kim J, Wilhelm T. 2008 What is a complex graph? *Physica A: Stat. Mech. Appl.* **387**, 2637–2652. (doi:10.1016/j.physa.2008.01.015)
23. Soloveichik D, Winfree E. 2007 Complexity of self-assembled shapes. *SIAM J. Comput.* **36**, 344–354. (doi:10.1137/S0097539704446712)
24. Adami C, Ofria C, Collier TC. 2000 Evolution of biological complexity. *Proc. Natl Acad. Sci. USA* **97**, 4463–4468. (doi:10.1073/pnas.97.9.4463)
25. Cronin L, Walker SI. 2016 Beyond prebiotic chemistry. *Science* **352**, 1174–1175. (doi:10.1126/science.aaf6310)
26. Seuss T. 1960 *Green eggs and ham*. New York, NY: Random House.
27. Stoker B. 1897 *Dracula*. Edinburgh, UK: Archibald Constable.
28. Adami C, Labar T. 2017 From entropy to information: biased typewriters and the origin of life. In *From matter to life: information and causality* (eds SI Walker, PCW Davies, GFR Ellis), pp. 130–154. Cambridge, UK: Cambridge University Press.
29. Davies PCW. 2011 Searching for a shadow biosphere on Earth as a test of the 'cosmic imperative'. *Phil. Trans. R. Soc. A* **369**, 624–632. (doi:10.1098/rsta.2010.0235)
30. Lovelock JE. 1965 A physical basis for life detection experiments. *Nature* **207**, 568–570. (doi:10.1038/207568a0)