Research

Author for correspondence:
Antony M. Jose
e-mail: amjose@umd.edu

A framework for parsing heritable information

Antony M. Jose

Department of Cell Biology and Molecular Genetics, University of Maryland, Room 2136, Bioscience Research Building (Building #413), College Park, MD 20742, USA

AMJ, 0000-0003-1405-0618

Living systems transmit heritable information using the replicating gene sequences and the cycling regulators assembled around gene sequences. Here, I develop a framework for heredity and development that includes the cycling regulators parsed in terms of what an organism can sense about itself and its environment by defining entities, their sensors and the sensed properties. Entities include small molecules (ATP, ions, metabolites, etc.), macromolecules (individual proteins, RNAs, polysaccharides, etc.) and assemblies of molecules. While concentration may be the only relevant property measured by sensors for small molecules, multiple properties that include concentration, sequence, conformation and modification may all be measured for macromolecules and assemblies. Each configuration of these entities and sensors that is recreated in successive generations in a given environment thus specifies a potentially vast amount of information driving complex development in each generation. This entity–sensor–property framework explains how sensors limit the number of distinguishable states, how distinct molecular configurations can be functionally equivalent and how regulation of sensors prevents detection of some perturbations. Overall, this framework is a useful guide for understanding how life evolves and how the storage of information has itself evolved with complexity since before the origin of life.

1. Introduction

Analyses of living systems from molecular to population scales have revealed information storage and processing across multiple scales as key attributes of life [1]. The need to understand the behaviour of a basic unit of life—a single cell—in terms of an integrated framework for information handling has been previously articulated [2–5], but is yet to be developed. A single cell is often the bottleneck stage that separates successive generations, making it the minimal space for storing all heritable information (see electronic supplementary material for variations on the single-cell bottleneck). Such information in molecules is part of the 'nature' of organisms and does not include information transmitted when parents train progeny, which can be considered as 'nurture'. Cells and more complex living systems can change their information content by learning through interactions with their environment. However, their ability to transmit any such learned information from one generation to the next is limited by the available storage in the bottleneck stage and potentially other system constraints (e.g. inability of learned information to cross generational boundaries) [6]. To appreciate these limits, we need to consider the total amount of information that could be encoded using *all* molecules in the bottleneck stage. Such joint consideration of all heritable information that is transmissible using molecules will inform how complexity grows over evolutionary time, what constitutes nature versus nurture and how to synthesize new living systems.

To facilitate discussion of all heritable information, I begin by defining key terms introduced in an earlier article [6]: stores of information, stored information and cell code. *Stores* (n.) of information refer to molecules or arrangements of

molecules that hold information. This information can be transferred to other molecules or arrangements and the original store can be degraded or modified after such transfer. Therefore, molecules and arrangements of molecules can have *stored* (v.) information. *Cell code* (n.) refers to the heritable information that *encodes* the development of organisms from a bottleneck stage, which is minimally a single *cell*. Similar development in successive generations in a given environment presumably relies on similar cell codes assembled during bottleneck stages (see electronic supplementary material for more on assembly of cell codes).

The information in a cell code can be conceptually separated into two distinct forms [6]. One is the genome sequence, where information is stored in a linear sequence of bases, and the other is the recurring arrangement, where information is stored in the concentrations, configurations and interactions of molecules in bottleneck stages (see electronic supplementary material on cell code assembly). While the information content in this arrangement and the extent to which it is recreated is currently not easily quantified, it is clear that heredity relies on information that is held in multiple stores and transmitted across generations. This communication of heritable information through the development of an organism from one generation to the next has been likened to the transmission of messages through a communication channel from sender to receiver (e.g. [6,7]). Just as 'the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point' [8], the fundamental problem of heredity is that of reproducing at one bottleneck stage either exactly or approximately a cell code selected at the preceding bottleneck stage.

Each choice of molecules and their arrangement in bottleneck stages collectively stores heritable information and forms a message transmitted across generations. The average information content in a message chosen from among N possible messages is given by the Shannon entropy, $H = \sum_{i=1}^{N} -p_i \cdot \log_2 p_i$, where p_i is the probability of the ith message [8,9]. If the probability of selecting each message is equal, this expression simplifies to give the maximal information a message can carry, $H = \sum_{i=1}^{N} -1/N \cdot \log_2(1/N) = \log_2 N$ bits. Therefore, to determine the maximal heritable information in a living system, we need to enumerate all distinguishable states of its bottleneck stage (i.e. N). This exercise will provide a starting point for the joint analysis of all heritable information that needs to be transmitted across generational boundaries for the reproduction of living systems.

2. Static and dynamic storage of information

The logical requirements for self-replication have been explored in two-dimensional universes called cellular automata using abstract 'machines' [10].[1] Of particular relevance are self-replicating machines that use the same store of information in two distinct ways: (i) as instructions whose interpretation leads to the construction of an identical copy of the machine and (ii) as data to be copied without interpretation and placed in the copied machine.

This scheme for making self-replicating machines avoids the infinite regress of instructions stored within instructions and is often presented as analogous to the process of self-replication in living systems with the 'instructions' being held in DNA. However, the instructions for replicating a

machine can be held either in static tapes (e.g. the von Neumann universal constructor, figure 1a) or in dynamic tapes (e.g. the Langton loop, figure 1b).

These two different types of instruction storage can be viewed as occurring simultaneously in living systems as the 'static' genome sequence and the 'dynamic' recurring arrangement of molecules [6].

The transmission of the genome sequence from one generation to the next occurs along a lineage of cells that each go through cell division cycles. As a result, numerous additional cycling stores can carry information across generations. For example, the information for copying a genome is stored in an arrangement of molecules that changes during replication such that the genome is usually copied with a period of one cell division cycle. Additional cycling stores of information are clearly recognizable in oscillations that occur at different temporal and spatial scales relative to generation time. These include oscillations in the chemical modifications of molecules (e.g. approx. 24 h period in cyanobacteria [13] despite an approx. 12 h cell cycle [14]), in the localizations of molecules (e.g. approx. 40 s period in cell lines that have an approx. 24 h cell cycle [15]), in the collective morphology of embryonic cells (e.g. approx. 0.5 h period in sea anemone that have an approx. 0.5 h cell cycle [16]) and in the activity levels of circuits (e.g. approx. 24 h circadian rhythms in non-cycling neurons [17]).

From these considerations, the following realizations emerge about living systems:

(1) The transmission of form and function across generations relies on many stores of information that cycle with different periods that could each in principle range from less than the duration of one cell division cycle to more than that of one generation.
(2) The relative phases of the many cycles within the continuous lineage of cells between generations create distinct states over time such that the cell code for the development of an organism is approximated at the start of each generation.

Thus, the integrated process of self-replication cannot be artificially parsed into the static genome that holds all the instructions to be interpreted by the dynamic molecular machines in the cell.

3. Information in self-replicating machines

Consideration of the total information stored in a self-replicating machine can clarify the different stores of information required for replication and sharpen the corresponding unknowns in living systems. For example, consider the self-replicating universal constructor (figure 1a), which has a 'machine' that has 6329 parts with 32 states per part and uses an instruction tape that has 145 315 parts with two states per part. The maximal information stored in this machine could be enumerated by separately considering three different stores that each have analogues in living systems: (i) the configuration or shape of the machine, (ii) the instruction tape, and (iii) the parts of the machine.

The information stored in the shape of the machine is incalculably large because we have to consider the universe of shapes from which the particular assembly of parts that
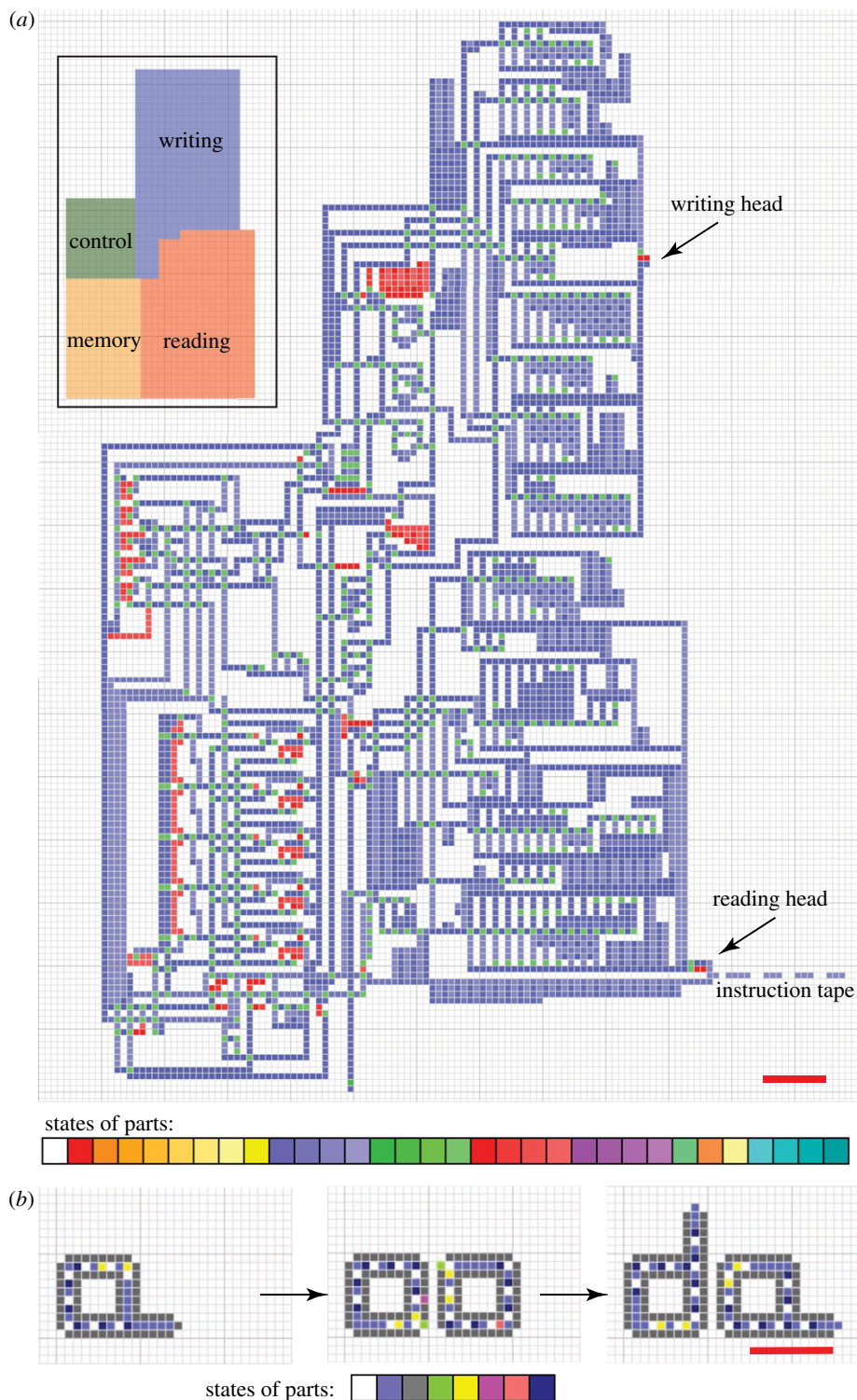
Figure 1. Self-replicating 'machines' with instructions held in a static tape or in a dynamic tape have been implemented in cellular automata. (*a*) Implementation of John von Neumann's design of universal constructor [11]. (Top) The universal constructor in the starting configuration. (Inset) Schematic of broad regions within the universal constructor. (Bottom) The 32 states of the parts that make up the machine (see [11] for the meaning of each colour). (*b*) Implementation of the Langton loop [12]. (Left) The loop in the starting configuration. (Middle) A replication intermediate showing the use of all states. (Right) Loops near the end of one round of replication. (Bottom) The eight states used for replicating the loop (see [12] for the meaning of each colour). Red bar indicates scale for comparing (*a*,*b*). See Methods in the electronic supplementary material for additional details.

make the machine was selected (see electronic supplementary material, text and figure S1 for a proof). This information is akin to the information required for getting together the particular collection of molecules that constitutes each current living system and has accrued since before the origin of life along the lineage of every living system. Because the unknown information in all historical environments (i.e. past available complements of molecules) needs to be taken into account to determine what life accrued bit by bit [18], the magnitude of this information is incalculable.

The maximal information that can be stored in the instruction tape that has $N = 145\,315$ parts with two states each is given by $H = \log_2 2^N = N = 145\,315$ bits. This store is analogous to the linear genome where the information is stored in the sequence of the four bases in DNA (A, T, G, C). For such a genome of length L, $H = \log_2 4^L = 2L$ bits.

The maximal information that can be stored in the machine that has $N = 6329$ parts with 32 states each is given by $H = \log_2 32^N = 5N = 31\,645$ bits. This store could be analogous to everything other than the genome sequence within the bottleneck stage. However, unlike in cellular automata, discrete states of living systems are not easily defined. Calculating the information in this potentially vast store requires definition of the biologically relevant states of the bottleneck stage. For any given genome, knowing the rest of the cell code for different organisms is a prerequisite for constructing living systems of varying complexity.

Here, I develop a framework for all the heritable information in a living system in terms of what that system can sense about itself and its environment. This framework is useful for guiding the experimental analysis of living systems and potentially for the design and analysis of other persistent adapting systems.

4. Heritable information in living systems

The spatial arrangement of the genome and everything else within the bottleneck stage could change over the course of development such that similar arrangements are reached with a period of one life cycle. As a result, molecules that are part of the recurring cell code could play different roles throughout development and defy permanent classification based on their roles. For example, an abundant maternal RNA that is simply used as a source of nucleotides in the developing embryo could at a later stage become a message that is translated into a protein. Nevertheless, a temporary classification during the bottleneck stage is necessary to enumerate the bits of information stored in cell codes. To facilitate this enumeration in units that are relevant for each living system and its environment, I propose considering entities, their sensors and the sensed properties.

4.1. Entities

An entity is a molecule[2] or association of molecules within a living system or in the environment that interacts with the living system. A cell code can include entities that are measured through interaction with other entities sometime during the life cycle and also entities that are never measured, which can be considered as by-products made by the processes of life. Such effectively inert and unmeasured entities could nevertheless non-specifically contribute to molecular crowding at the bottleneck stage and thereby affect interactions among other entities. While the number of all molecules in a cell is large but countable, the combinatorial associations of molecules could make the total number of effective entities (N) larger still. Cellular components that are entities or parts of entities include small molecules such as ATP, water, ions, metabolites, etc., for which perhaps only concentrations are discerned by sensors, and macromolecules such as individual proteins, RNAs, polysaccharides, etc., for which concentrations, sequences and conformations may all be discerned by sensors.

4.2. Sensors

A sensor is an entity or an association of entities within the living system that responds to changes in other entities with changes in its properties such that these changes can result in subsequent changes in the rest of the living system or its environment. A sensor could sense entities within the system (N total) or in the outside environment (O total) that interacts with the system (e.g. salts, nutrients, etc.). An entity that binds or collides with another entity without any specific downstream consequences is not considered a sensor (e.g. one water molecule bumping onto a membrane). An entity could be a part of multiple sensors. For example, a protein complex formed by the association of A, B and C proteins could be detecting and responding to the concentration of ATP, while another protein complex made of A and C could be detecting GTP. Conversely, multiple sensors could be measuring the same entity. For example, the many kinases in the cytosol are all potentially sensitive to the levels of a common pool of ATP. By these definitions, ATP itself can be a sensor because its levels change in response to production by a synthase and this change is communicated to the kinases that respond to changes in ATP levels. All sensors are entities, but not all entities are sensors.

4.3. Properties

A property is an attribute of an entity that is relevant for a living system because a sensor exists that can respond to changes in the values of that attribute. The number of different values for a property of an entity depends on the sensor and on the regulatory constraints of the system. Consider two sensors that can detect changes in the number of molecules of a particular RNA: a protein Lo responds when the numbers increase by 10 and a protein Hi responds when the numbers increase by 100. These two proteins would thus each 'see' different numbers of measurable units for the same property (number of molecules) of the same entity. However, not all detectable values for a relevant property of an entity could be attainable because of the regulatory constraints of the system. For example, if the RNA accumulated in steps of 50 molecules at a time, then many of the values measurable by Lo are never available in the living system because the system changes in steps that are larger than the measuring step of the Lo sensor.

4.4. Environment

Organisms develop as open systems interacting with the environment. Therefore, some entities in the environment are measured and reacted to by the living system throughout development. Even for a constant environment, some entities may be measured by different sensors that are active at different times during development. As a result, living systems are really system–environment combinations. Aspects of 'sensing' considered for interactions within the system are also relevant for interactions with the environment. Specifically, the sensed attributes of entities in the environment depend on the nature of the evolved sensors in the system and molecular crowders in the environment can modify the interaction between sensors in the system and entities in the environment.

4.5. Forces

Living systems can generate and be exposed to many kinds of forces, which are not being explicitly considered in the framework for heritable information developed here. Rather, they are implicitly accounted for in the properties of entities. For example, an entity experiences/exerts gravity because of its

mass, electrostatic attraction because of its charge, tension because of its elasticity, etc. Thus, heritable information captured using entities, sensors and properties can account for relevant forces in the living system or in the environment.

4.6. Configurations

The number of ways in which molecules can be arranged in the bottleneck stage such that they can be distinguished by the system provides an upper bound for the information that can be stored in a cell code, which is the subset of configurations that are nearly reproduced during the bottleneck stage of successive generations. The maximal number of such distinguishable configurations of a living system for a given number of interacting entities in the environment is given by the product of the number of possible genome sequences and the number of possible systems and their coupled environments that can support each genome sequence. Assuming that each system–environment combination generates one characteristic set of unmeasured entities that contribute to crowding effects, the number of distinguishable configurations for a living system and its environment during the bottleneck stage (C_{tot}) is given by

$$C_{tot} = X^L \left(\sum_{i=1}^{B} E_i \left(\sum_{j=1}^{S_i} S_j \left(\sum_{k=1}^{P_j} P_k \right) \right) \right) \qquad (4.1)$$

C_{tot} is the genomes × system–environments, X the number of types of bases in the genome, L the length of the genome in base pairs, E the measured entity (total B in the bottleneck stage: N_b in system, O_b in environment), S the measuring sensor (total S_i for ith entity), $= f(Y)$, where each $Y \subseteq \{E_1, E_2, …, E_N\}$ i.e. a configuration of entities, N per life cycle, P the attainable and measurable values of property (total P_j for the jth sensor of the ith entity).

This entity–sensor–property framework enumerates all distinguishable configurations as a product of four terms that encapsulate the maximal numbers of distinct states in two stores of information: X^L enumerates all possible genome sequences, which are replicating stores of heritable information, and $\sum_i E_i \sum_j S_j \sum_k P_k$ enumerates all potentially recurring arrangements of interacting molecules, which are cycling stores of heritable information. Such enumeration without considering rearrangements of chemical bonds within any molecule can be thought of as biological entropy and is less than the chemical entropy of an organism, which was initially estimated allowing for rearrangements of chemical bonds to be approximately 4.2×10^{10} bits for *Escherichia coli* [19].

It is clear that the replicating store cannot uniquely predict the cycling store as evidenced by most distinguishable cell types of the human body all having the same genome sequence. However, interdependence of the two stores and compatibility with the perpetuation of life reduce this maximal number of distinguishable states of the bottleneck stage. In other words, fewer configurations can act as heritable cell codes (Cell code$_{tot}$ < C_{tot}) because of mutual constraints between the arrangement of molecules and the genome sequence in living systems. First, some genome sequences may not be sufficiently complex to support any living system (e.g. a genome of all As, all Gs, all Cs or all Ts). Second, each genome sequence *constrains* but does not dictate the number and kinds of entities that could be part of any cell that contains the genome (e.g. DNA sequence constrains RNA

sequence, which constrains protein sequence). Third, the genome sequence may also constrain the total number of possible arrangements of molecules within any cell—i.e. the number of cell states and cell types—in a given environment. Fourth, the lineage of cells that connects two generations may be incapable of supporting some cell types because of the need to return to the cell code at the start of each generation within the context of a living system (i.e. some differentiated cell types may be irreversible within the context of the living system, although many can be transformed into pluripotent stem cells *in vitro* [20]). The number of all possible cell codes, however, is likely greater than that seen in evolved organisms because the historical process of evolution is not expected to allow exploration of every cell code (i.e. C_{tot} > Cell code$_{tot}$ > Cell code$_{evol}$).

Cell codes of varying complexity have evolved over time to specify the development of each organism that has ever existed [6]. Cell codes could in principle differ in the relative amounts of information stored in the genome sequence versus in the arrangement of molecules. The interdependence of these two stores of information invites exploration of the relationship between their storage capacities over evolutionary time. Consider the consequences of adding into a pre-existing cell code a newly evolved gene sequence that codes for a protein. The number of possible genome sequences of a given length that can support this cell code decreases because fewer distinct genomes can include the gene sequence for the new protein (i.e. total sequences becomes less than X^L, electronic supplementary material, figure S2). However, the number of distinguishable arrangements of all molecules can either increase or decrease. An increase can occur because addition of the new DNA sequence, the transcribed RNA and the translated protein to the contents of the cell could all lead to new interactions with pre-existing molecules (i.e. E, S and P could all increase, resulting in a larger value for $\sum_i E_i \sum_j S_j \sum_k P_k$). A decrease can occur because these new molecules could constrain the arrangement through regulatory interactions (see entity–sensor–property: insights). Furthermore, the magnitude of changes in $\sum_i E_i \sum_j S_j \sum_k P_k$ depends on the nature of the new gene product (e.g. expression or repression of many gene sequences by a transcriptional activator or repressor, respectively, could lead to large changes). Studies on the origin and evolution of information storage could illuminate trends in the partition of heritable information between different molecular stores and lead to general principles (for related views emphasizing arrangement, see [21,22], genome sequence, see [23,24], anatomy, see [25] and energy, see [26]). For example, the complexity of cell codes, and thus organisms, may have increased through restriction of the genome sequence along with expansion of the arrangement of molecules as sources of neutral or adaptive variation.

5. Entity–sensor–property: extensions

Several processes in living systems could limit or expand the number of arrangements in the bottleneck stage ($\sum_i E_i \sum_j S_j \sum_k P_k$). Processes that can change the information content of cell codes by decreasing (e.g. self-organization and self-assembly), increasing (e.g. chemical modification) or variably changing (e.g. compartmentalization) entities, sensors and/or properties are being actively analysed. Living

systems could manipulate heritable information through the regulation of all such processes.

5.1. Impact of self-assembly and self-organization

Order can arise through the spontaneous association of molecules in living systems. Two forms of such spontaneous order have been recognized: (i) self-assembly, which refers to the formation of static structures that are relatively stable (e.g. viruses, flagella) [27]; and (ii) self-organization, which refers to the formation of dynamic structures that appear stable (e.g. cytoskeleton, endocytic compartments) [28]. Both forms of order, however, depend on the immediate molecular environment. Therefore, changing the surroundings of a 'self-assembled' or 'self-organized' structure can result in alternative configurations that may be distinguishable by evolved sensors. For example, cells can use an adaptor protein to modulate the size of vesicles that form through self-assembly [29] and cells can respond to pressure by reversibly disassembling the mitotic spindle that is maintained through self-organization [30,31]. In this way, living systems can store and retrieve information from self-assembled and self-organized collections of molecules.

5.2. Impact of chemical modifications

Modifications of nucleic acids (^{5m}C, ^{5hm}C, ^{m6}A, etc.) or proteins (phosphorylation, methylation, glycosylation, etc.) result in new entities with properties that could potentially be measured by sensors. Modified bases on the genome could increase the number of possible spatial arrangements of the genome and its binding partners (i.e. E, S and P in equation (4.1)), and could also increase sequence information (i.e. X in equation (4.1)) *if* the modification alters base-pairing. Modifications on RNA or proteins on the other hand could either increase or decrease E, S, and P, but always reduce the maximal number of genomes of a given length that could support such a modification because each possible genome would be constrained to include the gene sequence for the enzyme that catalyses the modification (i.e. total sequences become less than X^L). Similar considerations hold for modifications of all other molecules in the bottleneck stage.

5.3. Impact of compartmentalization

Living systems dynamically manipulate which entities come together into organized units and which outputs from these units are subsequently measured. When different subcellular compartments form, the same entity or sensor could be present in two or more different compartments. If two such pools of the same entity are sensed separately during the life cycle of an organism, the total number of possible configurations is effectively increased. Alternatively, many different entities could be encapsulated into one compartment. If only a few aggregate properties of the compartment are sensed during the life cycle of an organism (e.g. droplet sizes of phase-separated aggregates such as RNA granules [32] or numbers of organelles such as mitochondria), the number of distinguishable configurations are effectively reduced.

These different ways of changing entities, sensors and properties highlight the multiscale nature of living systems and suggest the utility of different entity–sensor–property frameworks at different scales and across scales.

Entities : $E_1 = $ A,a,Ꞁ,ᴀ ; $E_2 = $ B,b,Ꞝ,ᴊ ; $E_3 = $ C,c,Ꞓ,ᴄ

Sensors : $S_1 = $ measures lines ; $S_2 = $ measures curves

Property : $P = $ number

measurable property values:

$$E_1S_1P = 3,0,4,5 \quad E_2S_1P = 1,1,5,4 \quad E_3S_1P = 0,0,3,3$$
$$E_1S_2P = 0,2,0,0 \quad E_2S_2P = 2,1,0,0 \quad E_3S_2P = 1,1,0,0$$

distinguishable states:

$$\begin{bmatrix} |E_1S_1P| & |E_2S_1P| & |E_3S_1P| \\ |E_1S_2P| & |E_2S_2P| & |E_3S_2P| \end{bmatrix} = \begin{bmatrix} 4 & 3 & 2 \\ 2 & 3 & 2 \end{bmatrix} = 288$$

Figure 2. Distinguishable states in a toy model of possible cell codes with a given 'genome' and 'environment'. Three entities (E_1, E_2, E_3), two sensors (S_1, S_2) and one sensed property (P) are considered. The measurable property values of each entity by each sensor are enumerated (E_1S_1P, E_1S_2P,...). Each distinguishable set of property values for all entities defines a distinguishable state. Therefore, the number of distinct elements in a set of the measured values (i.e. $|E_1S_1P|$, $|E_1S_2P|$,...) can be used to calculate the total number of distinguishable states in the system ($4 \times 3 \times 2 \times 2 \times 3 \times 2 = 288$), which is less than the number expected if every value of every entity were distinguishable ($4 \times 4 \times 4 \times 4 \times 4 \times 4 = 4096$).

6. Entity–sensor–property: insights

To appreciate some implications of the framework, consider a toy model where the genome sequence and the environment are held constant (figure 2).

Let the remaining contents of a 'cell' include three entities (E_1, E_2, E_3—three English letters) that can be at four different states (two fonts with upper and lower cases) and be sensed by two sensors (S_1 measuring lines and S_2 measuring curves). Each state is analogous to different *experimentally* measurable values for a property of molecules in a cell (e.g. concentration, localization, shape, charge, etc.). Consider the entity E_1 in state 'A' made of three straight lines. A sensor that measures lines could measure one of numerous possible properties: thickness of lines, colour of lines, length of lines, etc. For simplicity, let number be the only property P sensed by both S_1 and S_2. For example, the value of the property sensed by S_1 of E_1 in state 'A' is 3 and that sensed by S_2 of E_3 in state 'c' is 1 (see figure 2 for all values).

6.1. Sensors can limit information storage

To calculate the relevant information stored in a system, we need to enumerate the number of different states of the entities sensed by the system (figure 2).

While each sensor can sense one property of each entity, different states of an entity may not always be distinguishable by a sensor (e.g. S_1 will measure the states 'C' and 'c' of E_3 as 0 and S_2 will measure both as 1). Such indistinguishability is evident in living systems as the requirement for threshold levels of a signal for a detectable response. Thus, 'thresholding' by the sensor results in a reduction in the total number of states of the system and thus the storable information (figure 2). For example, this system of E, S and P can only distinguish 288 states (approx. 8.17 bits), but a naive estimate based on the ability to distinguish all states of all entities yields 4096 states (12 bits). Thus, the number of entropic states of a system depends on the available sensors and their sensitivity.

6.2. Distinct states may be equivalent

Selection can impose external constraints on the form and function of a living system. For example, the environment of the system in the toy model might require 'cells' with consistent fonts and case for survival. This would result in the survival of cells with 'ABC', 'abc', etc., as the values for each of the three entities (electronic supplementary material, figure S3). However, because there are two possible cases (upper versus lower), there would be two distinguishable states that are effectively equivalent for survival in this environment. Such situations could result in unregulated redundancy such that similar functions are performed by different molecules in random sets of cells [33]. Over evolutionary timescales, this type of unregulated redundancy could result in organisms with similar form and function but different underlying molecular mechanisms [34]. These considerations also hold when such equivalency is imposed by sensors that fail to distinguish different entities. For example, a channel protein responding to changes in membrane potential would measure changes in different ions as equivalent as long as the end result was a similar change in potential [35].

6.3. Regulation reduces sensed states

The different states of each entity could be classified as high (upper case) or low (lower case) to simplify the analysis of regulation in the system (figure 3a). This simplification is similar to Boolean networks that have been used to explore the impact of regulation [36]. All additions of either activation or inhibition as a regulatory interaction between two entities reduce the number of distinguishable states in the system (figure 3b). This reduction occurs because any regulatory interaction between two entities couples changes in those entities. As a result, two entities that were previously free to vary independently become either directly or inversely correlated, leading to an overall reduction in the number of possible states. Different regulatory architectures can lead to different states with equivalent capacity for information storage. Specifically, 12 different single regulatory interactions in the toy model lead to only three different storage capacities—96, 216 or 256 states (figure 3b). Adding two regulatory interactions results in all 36 different regulatory architectures having only 96 distinguishable states (electronic supplementary material, figure S4). These results suggest a preliminary conclusion: regulation reduces the ability of systems to store information in the arrangement of molecules.

6.4. Reducing states may promote robustness

Robustness is the ability of living systems to remain similar despite *some* variation introduced by environmental or internal conditions [37]. In other words, some changes either do not alter anything about a robust system or can alter some entities but nevertheless do not substantially affect the system. The differences in the number of states in cells with different regulatory architectures (figure 3b) suggest a relationship between regulation and robustness of cell types. Unlike in the toy model, in living systems, all sensors are made from entities (equation (4.1)). Therefore, cell types with fewer states could be more robust because they are only capable of sensing, and thus responding to, fewer perturbations. Conversely, cell types with many states could be less robust because they are capable of sensing and

Figure 3. Regulation reduces the number of states that can be sensed by the system. (a) Two states for each entity—high (upper case) or low (lower case)—were considered for exploring the impact of regulation in the toy model. (b) Consequences of introducing regulatory constraints. Inhibition (bar) or activation (arrow) between all pairs of entities were considered. Matrices of distinguishable values (product of all = number of states) for each cell with regulatory interactions between E_1 and E_2 (top), E_1 and E_3 (middle) or E_2 and E_3 (bottom) are shown. Different regulatory constraints result in differential reduction in the number of states of the system.

responding to many perturbations. Changes in regulatory architectures could therefore be used to generate cell types that are differently responsive to external signals, which may have implications for the observed robustness of development [38]. To achieve such robust development, entities need to be assembled into cell codes such that the same sequence of events unfolds despite some perturbations. Storing entities as perturbation-resistant assemblies or combining entities that fail under some conditions with entities that fail under other conditions (redundancy) are possible ways to ensure robust cell codes and subsequent development. An additional possibility suggested by these observations is reducing the number of sensors through increased regulation such that some perturbations are simply not sensed.

7. Two-base genomes could be part of efficient living systems

Our current ability to exquisitely edit genomes and transcriptomes [39,40] is a limited manipulation of living systems in that the outcome of the edit is entirely determined by how the living system interprets the change. In other words, we can make changes to a sequence and read out what the living system does with the changed sequence, but we cannot yet make changes that instruct a living system to perform arbitrary tasks. Such expanded manipulation could require ways of increasing the complexity of the stored heritable information. As suggested by equation (4.1), this increase could be achieved by either increasing storage in the genome sequence or by increasing storage in the arrangement of molecules. Increases in the storage capacity of a genome by increasing the number of different bases will require concomitant increases in the complexity of the machinery for accurate reading and writing of the genome.
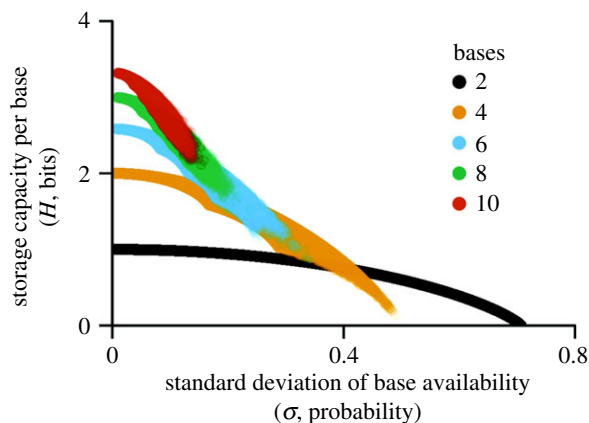
Figure 4. Increases in the types of base pairs in a genome increase the maximal information that can be stored per base but decrease the mean variation in base availability that can support the high information storage. Plot showing how maximal information (H) of a base pair varies with the standard deviation (σ) of base probability. For each n-base system ($n = 2, 4, 6, 8, 10$), the results for 1 million base probabilities drawn from a uniform distribution are plotted.

For example, a 16-base genome of length L has four times the capacity of a two-base genome ($4L$ bits versus L bits). However, such a genome would require machinery for discerning eight times as many kinds of bases. Furthermore, the range of availabilities of bases that can support the enhanced information storage decreases with an increase in the number of different bases (figure 4).

Perhaps the simplest route to the synthesis of living systems with arbitrary capacity to store heritable information would be to use a longer two-base genome that is equivalent to the natural four-base genome (e.g. needs five-base codons for encoding at least 20 amino acids, $2^5 > 20 > 2^4$), but can be supported by simpler machinery in the cell to read and write the genome. Testing the practicability of this speculation requires systematically changing the chemistry of the genome *and* the cell while preserving overall storage capacity.

8. Discussion

By jointly considering all information transmitted from one generation to the next using molecules, I have developed an expanded view of heredity (see electronic supplementary material for other applications). Heritable information stored outside the genome sequence is limited by mutual constraints with the sequence, by regulatory architectures and by what a living system can sense about itself and its environment.

8.1. Strengths and limitations of framework

Entities and sensors in a cell were parsed based on their roles at a particular time in development—the bottleneck stage. However, the roles of entities and sensors are potentially interconvertible over time. A sensor could become an unresponsive entity for a while and an entity could become a responsive sensor when it encounters another appropriate entity. Such changes in roles are likely part of the changes during development that lead to the assembly of cell codes at the start of each generation. Given this time-bound nature of entities and sensors, what is the *duration* of a bottleneck stage? This question is currently very difficult to answer

and poses a practical problem for unambiguously defining the cell code of an organism. Nevertheless, the stability of cell types suggests that functionally important states are preserved for significant periods through homeostasis.

The framework presented here does not account for the stochastic and noisy nature of all interactions within a cell. For example, there are fundamental limits to control that result from information loss [41] and the physical limits of biochemical signalling [42,43]. Unlike in man-made communication systems, the presence of numerous simultaneous signalling pathways in living systems—including as-yet unknown pathways—makes it unclear whether any observed variation in one signalling pathway should be characterized *a priori* as interference from another signalling pathway or as noise. Nevertheless, developing an understanding of heredity in terms of genome sequence, entities, sensors and properties is a first step towards future extensions of the framework that could address these issues.

Some past frameworks for analysing living systems provide conceptual structures for explaining their evolution and behaviour but do not inform their construction or origin. Models that analyse evolutionary outcomes regardless of the material basis of genotype and phenotype (e.g. [44]) are useful guides for the analysis of organisms at the population level but not for the construction of organisms from molecules sought here. Phylogeny, architecture and adaptation have been combined to understand trends in the evolution of form [45], but such models are currently not fine-grained enough to enable construction. The productive analysis of complex systems by partitioning a system into abstract nodes and edges to view particular aspects of living systems as networks [46] has generated intuitions and approaches that could be extended to the framework presented here. Such extension beyond abstract networks is necessary to enable the construction of living systems because typical abstractions do not incorporate all relevant properties of cellular contents. The explicit consideration of relevant properties for all entities that are measured by sensors in the framework presented here could help in accruing knowledge in a form that is useful for the construction of living systems and for the realization of a practical systems biology [47,48].

8.2. Synthesis of living systems

Building something using its constituent parts is a good way to discover the flaws in our understanding of how it is put together. For example, it is currently unclear if *perfect* self-replication ever occurs in living systems. Perhaps the perpetuation of life is always associated with having entities that are not recreated with a period of one generation but rather with longer or shorter periods. For example, when the noisy and variable behaviour of a synthetic oscillating circuit in *E. coli* [49] was improved to obtain synchronous long-term oscillations [50], the period of oscillation increased to 14 generations. Such possibilities can be explored by allowing different generation times for the precise recreation of some entities and arrangements in the cell code. The similarity in form and function of parent and progeny, however, suggests that the cell codes recreated with a period of one generation are at least nearly equivalent for specifying development in each generation.

Evolved cell codes are unlikely to be efficient stores of heritable information because of the historical measures and counter-measures through which evolution proceeds [51,52].

The efficient storage of the mutual information between two variables can be achieved using a compressed bottleneck variable [53,54]. If there was selection for effectively packing maximal information into the bottleneck stage in living systems, the entities and arrangements of evolved cell codes could similarly be efficient stores of the mutual information between the past and the future. All such efficient cell codes might have similar characteristics as observed in cellular automata in which the capacity to support computation emerges (captured in the λ parameter in [55]). Despite the possibility of such overall optimization, it is unclear if living systems can evolve to maximally optimize information storage and/or transmission for a particular trait. In fact, it might be difficult to define what the 'optimum' is for a process because the presence of many homeostatic mechanisms in cells, including transgenerational homeostasis [6], requires opposing processes that could limit optimality. Experimental approaches that attempt to generate minimal bacterial cells [56] need to be extended to different organisms to discover how the complexity of organisms scales with their cell codes.

Making efficient living systems of arbitrary complexity requires a holistic approach to information handling. The joint consideration of all heritable information presented in this article suggests that a genome with two different kinds of bases might function as an efficient replicating store when combined with the simplest possible cycling stores (figure 4). Thus far, experimental approaches to fundamentally change heritable information have focused on increasing the storage capacity of the genome. A 50% increase in the storage capacity of DNA sequence can be achieved by doubling the number of different bases in DNA [57].

Furthermore, an organism that uses a four-base genome can be modified with two additional DNA bases to successfully store [58] and retrieve [59] information. By contrast, we cannot yet engineer such increases in the information stored by the arrangement of molecules because our knowledge of this store of heritable information is in its infancy. The theoretical and practical limits of varying *all* heritable information deserve exploration to understand the evolution of natural, modified and synthetic living systems.

Endnotes

[1]Individual units in these machines are called 'cells', but are referred to as 'parts' in this article to avoid confusion with biological cells.
[2]For simplicity, the term molecule is used to refer to everything found in a living system that is chemically isolatable such as ions, atoms and chemically bonded collections of atoms, and is extensible to all factors that remain to be discovered.

References

1. Tkacik G, Bialek W. 2016 Information processing in living systems. *Annu. Rev. Condens. Matter Phys.* **7**, 89–117. (doi:10.1146/annurev-conmatphys-031214-014803)

2. Tanaka M. 1984 A physical characterization of biological information and communication system model of ecosystems. *J. Theor. Biol.* **110**, 89–117. (doi:10.1016/S0022-5193(84)80148-4)

3. Nurse P. 2008 Life, logic and information. *Nature* **454**, 424–426. (doi:10.1038/454424a)

4. Danchin A. 2009 Information of the chassis and information of the program in synthetic cells. *Syst. Synth. Biol.* **3**, 125–134. (doi:10.1007/s11693-009-9036-5)

5. Brenner S. 2010 Sequences and consequences. *Phil. Trans. R. Soc. B* **365**, 207–212. (doi:10.1098/rstb.2009.0221)

6. Jose AM. 2018 Replicating and cycling stores of information perpetuate life. *Bioessays* **40**, 1700161. (doi:10.1002/bies.201700161)

7. Battail G. 2010 Heredity as an encoded communication process. *IEEE Trans. Inf. Theory* **56**, 678–687. (doi:10.1109/TIT.2009.2037044)

8. Shannon C. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656. (doi:10.1002/j.1538-7305.1948.tb00917.x)

9. Cover TM, Thomas JA. 2016 *Elements of information theory*. Hoboken, NJ: Wiley-Interscience.

10. Sipper M. 1998 Fifty years of research on self-replication: an overview. *Artif. Life* **4**, 237–257. (doi:10.1162/106454698568576)

11. Pesavento U. 1995 An implementation of von Neumann's self-reproducing machine. *Artif. Life* **2**, 337–354. (doi:10.1162/artl.1995.2.4.337)

12. Langton CG. 1984 Self-reproduction in cellular automata. *Physica D* **10**, 135–144. (doi:10.1016/0167-2789(84)90256-2)

13. Rust MJ, Golden SS, O'Shea EK. 2011 Light-driven changes in energy metabolism directly entrain the cyanobacterial circadian oscillator. *Science* **331**, 220–223. (doi:10.1126/science.1197243)

14. Mori T, Binder B, Johnson CH. 1996 Circadian gating of cell division in cyanobacteria growing with average doubling times of less than 24 h. *Proc. R. Soc. Lond. B* **93**, 10 183–10 188. (doi:10.1073/pnas.93.19.10183)

15. Tsai RYL, McKay RDG. 2005 A multistep, GTP-driven mechanism controlling the dynamic cycling of nucleostemin. *J. Cell Biol.* **168**, 179–184. (doi:10.1083/jcb.200409053)

16. Ragkousi K, Marr K, McKinney S, Ellington L, Gibson MC. 2017 Cell-cycle-coupled oscillations in apical polarity and intercellular contact maintain order in embryonic epithelia. *Curr. Biol.* **27**, 1381–1386. (doi:10.1016/j.cub.2017.03.064)

17. Hastings MH, Maywood ES, Brancaccio M. 2018 Generation of circadian rhythms in the suprachiasmatic nucleus. *Nat. Rev. Neurosci.* **19**, 453–469. (doi:10.1038/s41583-018-0026-z)

18. Joyce GF. 2012 Bit by bit: the Darwinian basis of life. *PLoS Biol.* **10**, e1001323. (doi:10.1371/journal.pbio.1001323)

19. Morowitz HJ. 1955 Some order–disorder considerations in living systems. *Bull. Math. Biophys.* **17**, 81–86. (doi:10.1007/BF02477985)

20. Yamanaka S. 2012 Induced pluripotent stem cells: past, present, and future. *Cell Stem Cell* **10**, 678–684. (doi:10.1016/j.stem.2012.05.005)

21. Walker SI, Davies PCW. 2013 The algorithmic origins of life. *J. R. Soc. Interface* **10**, 20120869. (doi:10.1098/rsif.2012.0869)

22. Fields C, Levin M. 2018 Multiscale memory and bioelectric error correction in the cytoplasm–cytoskeleton membrane system. *Wiley Interdiscip. Rev.: Syst. Biol. Med.* **10**, e1410. (doi:10.1002/wsbm.1410)

23. Lynch M. 2007 The frailty of the adaptive hypothesis for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA* **104**, 8597–8604. (doi:10.1073/pnas.0702207104)

24. Száthmáry E, Smith JM. 1995 The major evolutionary transitions. *Nature* **374**, 227–232. (doi:10.1038/374227a0)

25. Levin M. 2020 The biophysics of regenerative repair suggests new perspectives on biological causation. *Bioessays* **42**, 1900146. (doi:10.1002/bies.201900146)

26. Annila A, Baverstock K. 2014 Genes without prominence: a reappraisal of the foundations of biology. *J. R. Soc. Interface* **11**, 20131017. (doi:10.1098/rsif.2013.1017)

27. Kushner DJ. 1969 Self-assembly of biological structures. *Bacteriol. Rev.* **33**, 302–345.

28. Nicolis G, Prigogine I. 1977 *Self-organization in nonequilibrium system*. New York, NY: John Wiley & Sons, Inc.

29. Zhang B, Koh YH, Beckstead RB, Budnik V, Ganetzky B, Bellen HJ. 1988 Synaptic vesicle size and number are regulated by a clathrin adaptor protein required for endocytosis. *Neuron* **21**, 1465–1475. (doi:10.1016/S0896-6273(00)80664-9)

30. Salmon ED. 1975 Pressure-induced depolymerization of spindle microtubules. I. Changes in birefringence and spindle length. *J. Cell Biol.* **65**, 603–614. (doi:10.1083/jcb.65.3.603)

31. Salmon ED. 1975 Pressure-induced depolymerization of spindle microtubules. II. Thermodynamics of *in vivo* spindle assembly. *J. Cell Biol.* **66**, 114–127. (doi:10.1083/jcb.66.1.114)

32. Shin Y, Brangwynne CP. 2017 Liquid phase condensation in cell physiology and disease. *Science* **357**, eaaf4382. (doi:10.1126/science.aaf4382)

33. Ravikumar S, Devanapally S, Jose A. 2019 Gene silencing by double-stranded RNA from *C. elegans* neurons reveals functional mosaicism of RNA interference. *Nucleic Acids Res.* **47**, 10 059–10 071. (doi:10.1093/nar/gkz748)

34. True JR, Haag ES. 2001 Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* **3**, 109–119. (doi:10.1046/j.1525-142x.2001.003002109.x)

35. Levin M. 2012 Molecular bioelectricity in developmental biology: new tools and recent discoveries. *Bioessays* **34**, 205–217. (doi:10.1002/bies.201100136)

36. Kauffman SA. 1969 Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467. (doi:10.1016/0022-5193(69)90015-0)

37. Stelling J, Sauer U, Szallasi Z, Doyle III FJ, Doyle J. 2004 Robustness of cellular functions. *Cell* **118**, 675–685. (doi:10.1016/j.cell.2004.09.008)

38. Waddington CH. 1942 Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563–565. (doi:10.1038/150563a0)

39. Knott GJ, Doudna JA. 2018 CRISPR–Cas guides the future of genetic engineering. *Science* **361**, 866–869. (doi:10.1126/science.aat5011)

40. Rees HA, Liu DR. 2018 Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788. (doi:10.1038/s41576-018-0059-1)

41. Lestas I, Vinnicombe G, Paulsson J. 2010 Fundamental limits on the suppression of molecular fluctuations. *Nature* **467**, 174–178. (doi:10.1038/nature09333)

42. Berg HC, Purcell EM. 1977 Physics of chemoreception. *Biophys. J.* **20**, 193–219. (doi:10.1016/S0006-3495(77)85544-6)

43. Bialek W, Setayeshgar S. 2005 Physical limits to biochemical signaling. *Proc. Natl Acad. Sci. USA* **102**, 10 040–10 045. (doi:10.1073/pnas.0504321102)

44. Rivoire O, Leibler S. 2014 A model for the generation and transmission of variations in evolution. *Proc. Natl Acad. Sci. USA* **111**, E1940–E1949. (doi:10.1073/pnas.1323901111)

45. Briggs DEG. 2005 *Evolving form and function: fossils and development*. In *Proc. of a Symposium Honoring Adolf Seilacher for His Contributions to Paleontology, in Celebration of His 80th Birthday*. A Special Publication of the Peabody Museum of Natural History. New Haven, CT: Yale University.

46. Barabási A-L, Oltvai ZN. 2004 Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113. (doi:10.1038/nrg1272)

47. Kitano H. 2002 Systems biology: a brief overview. *Science* **295**, 1662–1664. (doi:10.1126/science.1069492)

48. Kitano H. 2002 Computational systems biology. *Nature* **420**, 206–210. (doi:10.1038/nature01254)

49. Elowitz MB, Leibler S. 2000 A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338. (doi:10.1038/35002125)

50. Potvin-Trottier L, Lord ND, Vinnicombe G, Paulsson J. 2016 Synchronous long-term oscillations in a synthetic gene circuit. *Nature* **538**, 514–517. (doi:10.1038/nature19841)

51. Jacob F. 2018 Evolution and tinkering. *Science* **196**, 770–788.

52. Gould SJ, Lewontin RC. 1979 The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc. R. Soc. Lond. B* **205**, 581–598.

53. Tishby N, Pereira FC, Bialek W. 2000 The information bottleneck method. *arXiv*. arXiv:physics/0004057 [physics.data-an].

54. Kolchinsky A, Wolpert DH. 2018 Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* **8**, 20180041. (doi:10.1098/rsfs.2018.0041)

55. Langton CG. 1990 Computation at the edge of chaos: phase transitions and emergent computation. *Physica D* **42**, 12–37. (doi:10.1016/0167-2789(90)90064-V)

56. Hutchison III CA *et al.* 2016 Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253. (doi:10.1126/science.aad6253)

57. Hoshika S *et al.* 2019 Hachimoji DNA and RNA: a genetic system with eight building blocks. *Science* **363**, 884–887. (doi:10.1126/science.aat0971)

58. Zhang Y, Lamb BM, Feldman AW, Zhou AX, Lavergne T, Li L, Romesberg FE. 2017 A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proc. Natl Acad. Sci. USA* **114**, 1317–1322. (doi:10.1073/pnas.1616443114)

59. Zhang Y, Ptacin JL, Fischer EC, Aerni HR, Caffaro CE, San Jose K, Feldman AW, Turner CR, Romesberg FE. 2017 A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* **551**, 644–647. (doi:10.1038/nature24659)

A framework for parsing heritable information

Antony M Jose

Department of Cell Biology and Molecular Genetics,
University of Maryland,
College Park, MD 20742, USA

Supplementary Material

Text, 4 Figures and Figure Legends, Methods, and References.

Supplementary Text

Variations on the single-cell bottleneck

The life cycles of many multicellular organisms include a single-cell stage that limits the transmission of heritable information. However, the amount of information transmitted through this bottleneck stage could vary based on the ecology and developmental strategy of particular organisms.

Growth in a predictable environment that is stable for many generations may reduce the information that needs to be transmitted through the bottleneck stage. The reliable association of microbiomes and other symbionts in each generation could similarly facilitate the reduction of information transmitted through the bottleneck stage. At an extreme, viruses and parasites effectively form joint systems of heredity and development with the organisms they infect.

Developmental strategies can impact the temporal and spatial reach of the information that is transmitted across generations. For example, in a female human fetus, the germ cell precursors that will generate the oocytes have already differentiated, which could facilitate communication from the pregnant mother to the unborn grandchild through shared circulation. Such expansions of the bottleneck stage may increase the complexity of heritable form and function.

While the framework for all heritable information developed in this study is presented for the common case of a single-cell bottleneck between generations, it applies for all alternative configurations of the bottleneck stage.

Cell codes are assembled during development

The form and function of organisms are mostly preserved from one generation to the next. This preservation requires that development begin in successive generations with similar genome sequences and similar arrangements of regulatory molecules. While the genome is copied during every cell division, the arrangement of molecules presumably changes during development but returns to a similar configuration after one life cycle. These recurring arrangements therefore are cycling stores of heritable information that along with the genome sequence form cell codes for the development of organisms in each generation [1].

All cells accumulate molecules using building blocks from the environment in ways that depend on pre-existing molecules within cells. This dependence on prior state means that the current 'phenotype' of a cell is determined by the 'genotype' *and* the pre-existing phenotype of the cell. The observed similarity between organisms of successive generations implies that the bottleneck stage needs to have molecules arranged in such a way that similar temporal sequences and spatial patterns ensue during development in successive generations. In other words, cell codes encode both temporal and spatial order in spatial arrangement during the bottleneck stage.

Different collections and configurations of entities within a cell can arise through differences in physical and chemical processes that include: (1) the temporal sequence of binding or chemical reactions; (2) the relative rates of different reactions; (3) the confinement of reactions; (4) the amplification of biases that arise from intrinsic noise; (5) the addition of external entities; and (6) the destruction or secretion of entities. The arrangement of entities in a cell at any moment is an integrated consequence of the historical values of such differences leading up to that moment. An organism can therefore assemble the information for making a similar organism in the next generation by controlling such processes throughout development such that its bottleneck stage contains a well-configured cell code - a spatial representation of the past ready to shape the future.

Information content of a shape

To attempt calculating the information in a particular shape, we need to make assumptions about the universe of shapes from which that shape is drawn. For example, in a two-dimensional cellular automata environment, we could assume that the sets of parts from which the shape is formed are contiguous (i.e., each part shares at least one side with another part) and that rotations of shapes by multiples of 90 degrees are allowed. Given these assumptions, let $S_1, .., S_n$ be sets of parts that can each form one and only one target shape within it. The total number of objects of all sizes and shapes that could be formed using one such set of S_i parts is given by 2^{S_i}. Let O_i be each such set of 2^{S_i} objects (i.e. $|O_i| = 2^{S_i}$). The number of

all uniquely shaped objects in all such sets combined is given by $|O_1 \cup O_2 ... O_{n-1} \cup O_n| = U$ - the maximal number of unique objects aggregated from universes that could each contain an object with the target shape once and only once. The maximal amount of information in the target shape is therefore given by $H = log_2 U$ bits. Three simple cases illustrate how this number scales with the complexity and size of the shape in the cellular automata environment (figure S1).

For a target shape made of 1 part, $U = 2$ and $H = log_2(2) = 1$ bit. For a target shape made of 2 parts that are next to each other, $U = 4$ and $H = log_2(4) = 2$ bits. However, for a target shape made of 3 parts that are next to each other in a row (or column), $U = \infty$ because an infinite number of parts that zig zag such that they only share a side with one row part and one column part could belong to the universe from which the shapes are drawn. Similarly, for all machines made of $n > 3$ parts in a row, contiguous sets of parts that zig zag and switch direction after fewer than n parts can be constructed to make $U = \infty$. Thus, for the universal constructor made of 6,329 parts that are arranged into a complex shape (figure 1A), the information content is incalculably large.

Configurations of a gene sequence and its regulators

Basic principles underlying the assembly of entire cell codes may be discoverable through reductionist studies on a few units of heredity. These studies would aim to discover how a unit of heredity is configured in one bottleneck stage, how that configuration changes during development, and how the starting configuration is recovered by the next bottleneck stage. More than a century of experimental analyses support the usefulness of analyzing units of heredity [1], which were initially called cell elements [2] and can be thought of as having two parts: (1) a gene sequence transmitted between generations as part of the genome sequence; and (2) gene regulators transmitted between generations as part of an arrangement of molecules. The precise limits of a gene sequence have proven to be difficult to establish [3, 4] and the precise configuration of *all* regulators for a given gene is likely to be similarly difficult to establish. Nevertheless, formulating a unit of heredity that is associated with a gene sequence as a cell element is useful as a practical framework for reductionist studies. Such studies do not need to analyze all entities and their sensors - indeed this is impractical. Instead, subprocesses like transcription, translation, protein localization etc. could be partitioned and then the rest of the entities, their sensors, and sensed properties under study could be analyzed to determine if they are sufficient to account for the observed heritable phenomena. To facilitate such reductionist analyses, a gene sequence and its regulators could be parsed into a provisional entity-sensor-property system and other distantly interacting entities could be considered as part of the 'environment'. After such simplification, the configurations that this provisional entity-sensor-property system can distinguish is given by:

$$c_{tot} = x^l (\sum_{i=1}^{b} e_i (\sum_{j=1}^{s_i} s_j (\sum_{k=1}^{p_j} p_k)))$$

$x =$ number of different types of bases in the gene.

$l =$ length of the gene sequence.

$e =$ measured entity (total b in the bottleneck stage: n_b in system, o_b in environment).

$s =$ measuring sensor impacting regulation of the gene sequence (total $= s_i$ for i^{th} entity).

$\quad = f(y)$, where each $y \subseteq \{e_1, e_2, ...e_n\}$, i.e., a configuration of entities, n per life cycle.

$p =$ attainable and measurable values of property (total $= p_j$ for j^{th} sensor of ith entity).

Cell elements that encode the expression patterns of gene sequences would therefore be subsets of c_{tot} that are recreated in successive generations. Progressive application of this framework by considering larger systems successively could provide a principled approach for combining cell elements into cell codes.

Other applications of the Entity-Sensor-Property framework

The three parameters - entities, sensors, and properties - introduced here for measuring cycling stores of information may be applicable to broader non-biological classes of heritable information. As an extreme

example, the persistence or evolution of ideas among groups of people could potentially be analyzed similarly. For ideas (or memes [5]) to be transmitted through a book (entity, E), the book needs to be read by a person (sensor, S) and its meaning (property, P) understood. A reader who writes with or without changing the ideas in the original book is effectively transmitting information across one 'generation'. (Perhaps, the many unread books are akin to the unmeasured molecules that crowd. Such crowding could narrow the focus of the reader on a few books and potentially change the nature of the books that are written.) Collective analysis of many such transmissions through books and other media may provide insights into the origins of a culture or zeitgeist.

The homeostatic preservation of cell codes in successive generations that living systems achieve using their entities, sensors, and properties could inform the design, analysis, and construction of other complex adaptive systems. To apply the Entity-Sensor-Property framework, the information content in a complex system needs to be parsed in terms of these three parameters. Organizations, economies, social networks, ecosystems etc., may all be amenable to such parsing. Indeed, similar sensor-based detection and control has been proposed even for the analysis of human behavior [6]. In cases where the constituent parts of a system are not known or are unknowable [7], simulations exploring a variety of possible entities, sensors, and properties could help constrain hypotheses.

In summary, the framework developed here for heritable information in living systems is applicable across many scales and therefore may be a generally useful lens for viewing other persistent adapting systems.

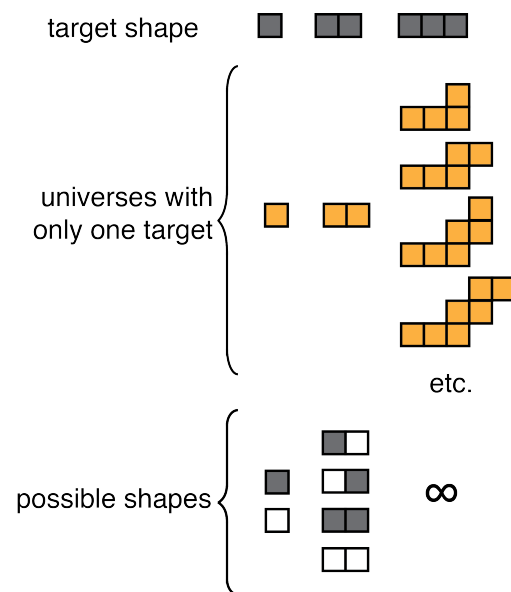# Supplementary Figures and Figure Legends

Figure S1. The maximal information content in non-trivial shapes is incalculably large because the number of universes that can be constructed to contain a target shape once and only once is infinite (see supplemental text for details).
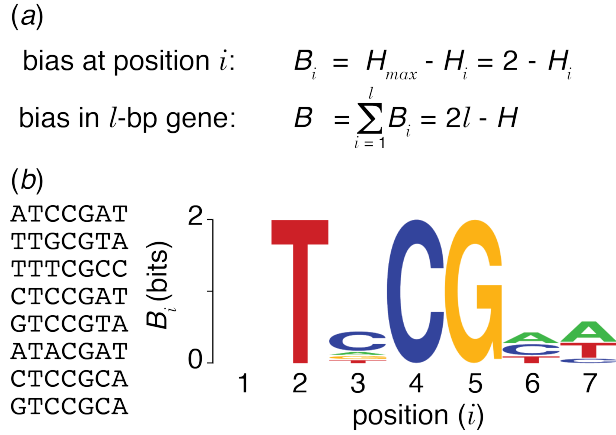
(*a*)

bias at position i: $B_i = H_{max} - H_i = 2 - H_i$

bias in l-bp gene: $B = \sum_{i=1}^{l} B_i = 2l - H$

(*b*)

```
ATCCGAT
TTGCGTA
TTTCGCC
CTCCGAT
GTCCGTA
ATACGAT
CTCCGCA
GTCCGCA
```


Figure S2. Conserved bases among a set of sequences reduce the capacity for storing new information in any one sequence. (A) Sequence bias measures the reduction in the capacity to store information. The bias at any position in a 4-base genome is given by $B_i = H_{max} - H_i = 2 - H_i$, where H_{max} and H_i are the maximal and observed storable information at a position i. For a gene of length l, the total bias is given by $B = \sum_i B_i = l \cdot log_2 4 - \sum_i H_i = 2l - H$. (B) Sequence bias reduces available space for storing new information. *Left*, A set of aligned sequences made of 4 bases with varying degrees of conservation at individual positions. *Right*, Bits of bias at each position (B_i) in the set of sequences depicted as a sequence logo [8] using weblogo [9] without small sample correction. Thus, greater the conservation among a set of genomes from different organisms, fewer the bases available for storing *new* information in the genome that distinguishes each organism.
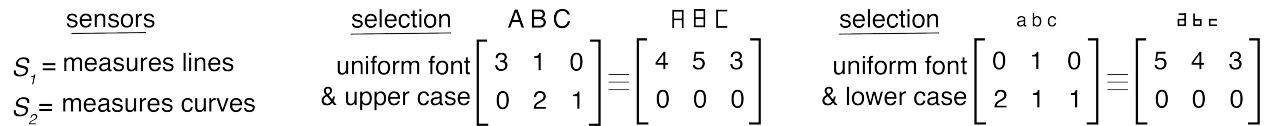

Figure S3. Distinct internal configurations of the toy model may be seen as equivalent by selection. With the two sensors (left) of the toy model, selection for uniform font & upper case (middle) or selection for uniform font & lower case (right) would result in distinct internal states of the cell becoming equivalent. Numbers in matrices are the property values of the contents of the cell in different states (ABC, abc, etc.) as measured by the two sensors.

$$E_1 \alpha E_2 \alpha E_3 \qquad \frac{1}{E_1} \alpha E_2 \alpha E_3 \qquad E_1 \alpha \frac{1}{E_2} \alpha E_3 \qquad E_1 \alpha E_2 \alpha \frac{1}{E_3}$$

$E_1 \rightarrow E_2 \rightarrow E_3$	$E_2 \dashv E_1 \dashv E_3$	$E_1 \dashv E_2 \dashv E_3$	$E_1 \dashv E_3 \dashv E_2$
$E_1 \rightarrow E_3 \rightarrow E_2$	$E_3 \dashv E_1 \dashv E_2$	$E_3 \dashv E_2 \dashv E_1$	$E_2 \dashv E_3 \dashv E_1$
$E_2 \rightarrow E_3 \rightarrow E_1$	$E_1 \dashv E_2 \rightarrow E_3$	$E_2 \dashv E_1 \rightarrow E_3$	$E_3 \dashv E_1 \rightarrow E_2$
$E_2 \rightarrow E_1 \rightarrow E_3$	$E_1 \dashv E_3 \rightarrow E_2$	$E_2 \dashv E_3 \rightarrow E_1$	$E_3 \dashv E_2 \rightarrow E_1$
$E_3 \rightarrow E_1 \rightarrow E_2$	$E_2 \rightarrow E_3 \dashv E_1$	$E_3 \rightarrow E_1 \dashv E_2$	$E_1 \rightarrow E_2 \dashv E_3$
$E_3 \rightarrow E_2 \rightarrow E_1$	$E_3 \rightarrow E_2 \dashv E_1$	$E_1 \rightarrow E_3 \dashv E_2$	$E_2 \rightarrow E_1 \dashv E_3$

$$\begin{bmatrix} 2 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} + \begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix} = 96 \text{ states}$$

Figure S4. Impact of two regulatory constraints on the toy model. Different inhibition (bar) or activation (arrow) interactions between all three entities were considered. Four different relationships between the three entities can arise when two interactions are added (9 regulatory architectures each). In all 36 architectures, there are only two sets of values that all entities can take (e.g., when $E_1 \rightarrow E_2 \rightarrow E_3$, then either all entities have the uppercase values or all entities have the lowercase values). This results in the detection of fewer states by each sensor such that every architecture results in the same two matrices of sensed values. Calculating the sum of the products of the elements in each matrix gives a total of 96 states.

Methods

Calculations of states for the toy model of an Entity-Sensor-Property system were performed by hand. Distributions of H and σ were calculated for one million samples of varying probabilities of base composition using the code below and plotted in R [10].

```
S = 0; H = 0; data = list()
for (j in seq(2,10,2)){
for (i in 1:1000000){
vals = matrix(runif(j, 0, 1), nrow = 1, ncol = j)
probs = sapply(vals, FUN = function(x) x/sum(vals))
S[i] = sd(probs)
H[i] = prod(-1, sum(sapply(probs, FUN = function(x) prod(x, log2(x)))))
}
data[[j]] = data.frame(S,H)
}
```

For figure 1, images were captured from simulations created in the Golly application [11]. Figure 1A depicts an implementation of von Neumann's self-reproducing universal constructor at the starting stage. This machine is a modification of the original design by von Neumann, but is regarded as the first implementation of his vision of a self-replicating universal constructor. Minor correction of the published design and a modification that reduces the tape length by $\approx 13\%$ (script by Tim Hutton) were used in this implementation authored by Renato Nobili and Umberto Pesavento. Figure 1B depicts an implementation of the Langton loop. Instructions are stored in a set of dynamic core parts (light blue) that are surrounded by static sheath parts (grey). Each signal packet consists of the signal (yellow, dark blue, orange, magenta, green) followed by a blank (white). This implementation was done by Eli Bachmutsky.

Supplementary References

1. Jose, A. M. Replicating and cycling stores of information perpetuate life. *BioEssays* **40,** 1700161 (2018).

2. Mendel, G. Experiments on plant hybrids. [Translated in 1901 by William Bateson for Royal Horticultural Society]. *Proceedings of the Naturalist Association, Brno* **IV,** 40–95 (1866).

3. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Research* **17,** 669–681 (2007).

4. Portin, P. & Wilkins, A. The Evolving Definition of the Term "Gene". *Genetics* **205,** 1353–1364 (2017).

5. Dawkins, R. *The selfish gene.* p. 192 (Oxford University Press, Oxford, U.K., 1989).

6. Powers, W. T., Clark, R. K. & McFarland, R. L. A general feedback theory of human behavior. *Perceptual and Motor Skills* **11,** 71–88, 309–323 (1960).

7. Rosen, R. *Essays on life itself* (Columbia University Press, New York, New York, U.S.A., 2000).

8. Schneider, T. D. & Stephens, M. Sequence logos: a new way to display consensus sequence. *Nucleus Acids Research* **18,** 6097–6100 (1990).

9. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Research* **14,** 1188–1190 (2004).

10. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2019). `https://www.R-project.org`.

11. Golly version 3.2. doi:`http://golly.sourceforge.net` (2005-2018).