Intrinsic information representation in biological systems

Cameron Ray Smith[1], Aviv Bergman[1,2,3]
[1]*Department of Systems and Computational Biology,*
[2]*Dominick P. Purpura Department of Neuroscience,*
[3]*Department of Pathology, Albert Einstein College of Medicine,*
1301 Morris Park Ave, Bronx, NY 10461, USA
(Dated: October 2, 2012)

One of the most important aspects of advancing theory in biology is the development of a constructed language that can be explicitly and precisely defined and supports communication among scientists, between scientists and computers, and even within individual scientists themselves. Such a language should itself be evolvable and support flexible abstractions that enable the unification and compression of redundant concepts. It would be helpful if this language could be formally specified prior to or in concert with the development of a computational implementation that may support complementary organization of semi-autonomously linked and computable data. We propose the adaptation of a language originally developed in a mathematical context that apparently meets these criteria. We argue for the particular choice we suggest, not from a mathematical perspective but, by providing an example of the way in which this language enables precise qualitative reasoning and flexible methods of abstraction with respect to integrated biological information. Enabling lossless compression and representation of synthetic knowledge about biological systems, in addition to raw biological data, is necessary for understanding in the context of limited cognitive and other computational resources.

Contents

I. INTRODUCTION

The development of a formal language for modeling biological systems was suggested by Woodger in collaboration with the developmental biologist Waddington and the logician Tarski as early as 1937 [1–4]. At that time it was perhaps difficult to understand how such a language could be put to use. Today we have tools that could enable the use of such a language: namely 1) computational machines to deal with the details of repeated transformations within the language and 2) large accessible repositories of data. Though we have access to necessary infrastructure, we lack the type of language suggested by Woodger and others throughout the course of the 20th century as we have continued to rely on natural language heuristics to communicate and reason about biological systems.

Category theory [5–10] is a language that has been suggested since Woodger to provide a framework for representing and reasoning about biological systems [11–13]. What is immediately useful about this language from the perspective of biology is that it presents as primitive the notion of transformation or interaction between objects. In fact, from this point of view, a defining characteristic of any entity (e.g. a protein, cell, organism, or population) is structural: the set of relationships between it and other entities under consideration. Intrinsic properties are taken into account implicitly in this framework since the possible set of relationships between any object and any other is constrained by the nature of its intrinsic properties.

What is less obvious at first meeting is the way in which category theory, especially with regard to its interface with logic and geometry [7, 14], enables the precise definition of what might be viewed as a precise framework for so-called *emergent properties* of systems of interacting objects. This expression is general enough that it can be equally well applied to the consideration of relationships between any levels of biological organization and may be generalizable to arbitrary physical contexts. This unity deriving from the judicious definition of underlying categorical concepts along the way is precisely the type of abstraction that we argue is necessary to enable compression without loss of biological information and thus the synthesis of existing biological knowledge.

Here we define the concepts from category theory necessary to understand the way in which interacting objects at one level of organization (e.g. molecules) can produce phenomena that would themselves be identifiable as derived objects (e.g. cells) that justify the very conceptualization of a *level of organization* in the first instance. Here we focus exclusively on defining the boundary conditions relating levels (e.g. molecules and cells or cells and multicellular organisms) of organization, which are necessary to understand in the course of defining a dynamical system that could model the *evolution* of such levels of organization. What results is a refinement of the concept of levels of organization examples of which are used so far only heuristically as guides to pre-existing intuition.

II. BIOLOGICAL INFORMATION

A. Biological system-environment duality

Distinction between biological systems or some components thereof and the environments within which they are embedded is implied in models of such systems. This distinction is useful in many contexts, but the boundary between a biological system and its environment is dependent upon the level of resolution of the *model and modeler* independent of its relationship to properties of the biological system itself [15]. In this light, it is desirable to develop a model of biological systems that supports variation of this boundary without requiring any reconfiguration of the model. Constructing a framework for such models requires the determination, unification, and incorporation of abstract features of biological systems that are invariant across levels of organization from molecules to cells, organisms, populations, communities, ecosystems and more fine-grained levels of resolution that likely lie between these broadly and imprecisely defined perspectives one can take with regard to representing biological systems.

B. Transformation of biological information at system-environment boundaries

The use of categorical adjunctions to model information representation in biological systems has been discussed in some detail [11, 16, 17], but these efforts have yet to be incorporated into more detailed theories. We begin here by developing the prerequisite definitions to explain why the transformation of biological information is naturally represented as a pair of adjoint functors in the context of category theory.

Definition C. A *category* $\mathcal{C}$ is:

1. A set of objects $\mathrm{Ob}(\mathcal{C})$.
2. For each pair $x, y \in \mathrm{Ob}(\mathcal{C})$ a set of morphisms $\mathrm{Mor}_{\mathcal{C}}(x, y)$.
3. For each triple $x, y, z \in \mathrm{Ob}(\mathcal{C})$ a composition map $\mathrm{Mor}_{\mathcal{C}}(y, z) \times \mathrm{Mor}_{\mathcal{C}}(x, y) \to \mathrm{Mor}_{\mathcal{C}}(x, z)$, denoted $(\phi, \psi) \mapsto \phi \circ \psi$.

Such that these constraints are satisfied:

1. For every element $x \in \mathrm{Ob}(\mathcal{C})$ there exists a morphism $\mathrm{id}_x \in \mathrm{Mor}_{\mathcal{C}}(x, x)$ such that $\mathrm{id}_x \circ \phi = \phi$ and $\psi \circ \mathrm{id}_x = \psi$.
2. Composition is associative, i.e., $(\phi \circ \psi) \circ \chi = \phi \circ (\psi \circ \chi)$.

Definition D. A *functor* $F : \mathcal{A} \to \mathcal{B}$ between two categories $\mathcal{A}, \mathcal{B}$ is:

1. A map $F : \mathrm{Ob}(\mathcal{A}) \to \mathrm{Ob}(\mathcal{B})$.
2. For every $x, y \in \mathrm{Ob}(\mathcal{A})$ a map $F : \mathrm{Mor}_{\mathcal{A}}(x, y) \to \mathrm{Mor}_{\mathcal{B}}(F(x), F(y))$, denoted $\phi \mapsto F(\phi)$.

These data should be compatible with composition and identity morphisms in the following manner: $F(\phi \circ \psi) = F(\phi) \circ F(\psi)$ for a composable pair (ϕ, ψ) of morphisms of $\mathcal{A}$ and $F(\mathrm{id}_x) = \mathrm{id}_{F(x)}$.

Definition E. Let $F, G : \mathcal{A} \to \mathcal{B}$ be functors. A *natural transformation*, or a *morphism of functors* $t : F \to G$, is a collection $\{t_x\}_{x \in \mathrm{Ob}(\mathcal{A})}$ such that

1. $t_x : F(x) \to G(x)$ is a morphism in the category $\mathcal{B}$, and
2. for every morphism $\phi : x \to y$ of $\mathcal{A}$ the following diagram is commutative

$$\begin{array}{ccc} F(x) & \xrightarrow{t_x} & G(x) \\ {\scriptstyle F(\phi)}\downarrow & & \downarrow{\scriptstyle G(\phi)} \\ F(y) & \xrightarrow{t_y} & G(y) \end{array}$$

We can define a category having functors as objects and natural transformations as morphisms, which is called a functor category, by recognizing that every functor F comes with the *identity* transformation $\mathrm{id}_F : F \to F$. In addition, given a morphism of functors $t : F \to G$ and a morphism of functors $s : E \to F$ then the *composition* $t \circ s$ is defined by the rule

$$(t \circ s)_x = t_x \circ s_x : E(x) \to G(x)$$

for $x \in \mathrm{Ob}(\mathcal{A})$. This is a morphism of functors from E to G. Thus, given categories $\mathcal{A}$ and $\mathcal{B}$ we obtain the category of functors between $\mathcal{A}$ and $\mathcal{B}$.

Definition F. An *equivalence of categories* $F : \mathcal{A} \to \mathcal{B}$ is a functor such that there exists a functor $G : \mathcal{B} \to \mathcal{A}$ such that the compositions $F \circ G$ and $G \circ F$ are isomorphic to the identity functors $\mathrm{id}_{\mathcal{B}}$, respectively $\mathrm{id}_{\mathcal{A}}$. In this case we say that G is a *quasi-inverse* to F.

Definition G. Let $\mathcal{C}$, $\mathcal{D}$ be categories. Let $u : \mathcal{C} \to \mathcal{D}$ and $v : \mathcal{D} \to \mathcal{C}$ be functors. We say that u is a *left adjoint* of v, or that v is a *right adjoint* to u if there are bijections

$$\mathrm{Mor}_{\mathcal{D}}(u(X), Y) \longrightarrow \mathrm{Mor}_{\mathcal{C}}(X, v(Y))$$

functorial in $X \in \mathrm{Ob}(\mathcal{C})$, and $Y \in \mathrm{Ob}(\mathcal{D})$.

Definition H. Given a category $\mathcal{C}$ the *opposite category* $\mathcal{C}^{opp}$ is the category with the same objects as $\mathcal{C}$ but all morphisms reversed.

Definition I. Let $\mathcal{C}$, $\mathcal{S}$ be categories. A *contravariant* functor F from $\mathcal{C}$ to $\mathcal{S}$ is a functor $\mathcal{C}^{opp} \to \mathcal{S}$.

Definition J. Let $\mathcal{C}$ be a category.

1. A *presheaf of sets on* $\mathcal{C}$ or simply a *presheaf* is a contravariant functor F from $\mathcal{C}$ to *Sets*.

2. The category of presheaves is denoted $PSh(\mathcal{C})$.

Example K. Functor of points. For any $U \in \mathrm{Ob}(\mathcal{C})$ there is a contravariant functor

$$\begin{array}{rccc} h_U & : & \mathcal{C} & \longrightarrow & Sets \\ & & X & \longmapsto & \mathrm{Mor}_{\mathcal{C}}(X, U) \end{array}$$

which takes an object X to the set $\mathrm{Mor}_{\mathcal{C}}(X, U)$. In other words h_U is a presheaf. Given a morphism $f : X \to Y$ the corresponding map $h_U(f) : \mathrm{Mor}_{\mathcal{C}}(Y, U) \to \mathrm{Mor}_{\mathcal{C}}(X, U)$ takes ϕ to $\phi \circ f$. We will always denote this presheaf $h_U : \mathcal{C}^{opp} \to Sets$. It is called the *representable presheaf* associated to U If $\mathcal{C}$ is the category of schemes this functor is sometimes referred to as the *functor of points* of U.

L. Intrinsic measurement processes of biological systems

M. Topological features of biological information

N. Variable filtration of biological information

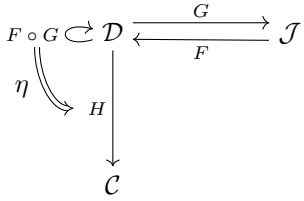
O. Deconstructing and synthesizing biological information

P. Enabling conditions of hierarchical transformations

III. BIOLOGICAL EVOLUTION

A. Niche construction and multilevel selection

Selection is part of any evolutionary process. Niche construction, which is complementary to selection, has been introduced more recently [18, 19]. An organism experiences a situation similar to that represented in diagram A.1.



(A.1)

The molecular interaction network that underlies an arbitrary type of organism is represented by a category $\mathcal{D}$ and the environment is represented by a category $\mathcal{C}$. Due to the dependence of selection on the environment, those organisms capable of survival are those that most closely approximate a natural transformation $\eta : F \circ G \Rightarrow H$. Why might this be the case?

To answer this question requires consideration of the necessity of the category $\mathcal{J}$ in this conceptual framework. From the perspective of any organism, we can roughly think of $\mathcal{D}$ as "internal" and $\mathcal{C}$ as "external" in a sense that can be likened but not yet clearly identified with "known" and "unknown" respectively. The selection process can be thought of as requiring the organism to establish a representation of properties of $\mathcal{C}$ so as to enable its existence to remain complementary, and thus stable, relative to its environment. The molecular network underlying the various biochemical functions of the organism has no direct access to the processes that determine the causal structure of the environment. In this case we consider the organism to have direct access only to the structure contained in $\mathcal{D}$, but nevertheless being implicitly assigned, as a result of its relationship to the environment, the seemingly impossible task of establishing the structure of $\mathcal{C}$. If $\mathcal{C}$ were internalized by the organism then relationships between $\mathcal{D}$ and $\mathcal{C}$ that preserve

structure, specifically the functors labelled $H : \mathcal{D} \to \mathcal{C}$, could also be directly internalized. Since $\mathcal{C}$ is external to $\mathcal{D}$, the organism must find another way of determining the relationship between its internal structure and that which is external, even if local.

One way to address this apparently paradoxical situation is to consider the relationship between endofunctors on $\mathcal{D}$ and those like H. In the case of A.1, $\mathcal{J}$ is considered to be a subcategory of $\mathcal{D}$ that contains all the domains (in the sense of complete partial orders) of $\mathcal{D}$. This scenario allows functors $F : \mathcal{J} \to \mathcal{D}$ to index diagrams in $\mathcal{D}$.

See [20] for information on multilevel selection.

B. The structure of molecular networks

1. Endo-genetics, genetics, epi-genetics, and n-epi-genetics

C. Hierarchical organization and evolution

See [21], [22].

IV. HIERARCHY IN LOGIC

A. Russell's paradox

A theorem due to Cantor states that given a set X and a map of sets f it is impossible to find a mapping

$$f : X \to 2^X$$

where 2^X represents the set of all subsets of X also referred to as the powerset of X.

B. A type theoretic resolution

C. Categorical type theory

See [23].

V. INFORMATION REPRESENTATION

A. Information theory

See [17] and [24].

B. Domain theory

See [25], [26].

C. Game semantics

See [27]

Acknowledgements

CRS was supported by . AB was supported by .

Appendix A: Category theory for tourists

[1] Joseph Henry Woodger. *The axiomatic method in biology,*. The University press, Cambridge [Eng.], 1937. I

[2] Joseph Henry Woodger. Science without Properties. *The British Journal for the Philosophy of Science*, II(7):193–216, 1951.

[3] Joseph Henry Woodger. From biology to mathematics. *The British Journal for the Philosophy of Science*, III(9):1–21, 1952.

[4] J. H. Woodger. *Biology and Language*. Cambridge University Press, 1952. I

[5] Saunders Mac Lane. *Mathematics: Form and Function*. Springer, 1985. I

[6] Saunders Mac Lane. *Categories for the Working Mathematician (Graduate Texts in Mathematics)*. Springer, 1998.

[7] Saunders Mac Lane and Ieke Moerdijk. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory (Universitext)*. Springer, 1992. I

[8] F. William Lawvere and Stephen H. Schanuel. *Conceptual Mathematics: A First Introduction to Categories*. Cambridge University Press; 1st edition, 1997.

[9] F. William Lawvere and Robert Rosebrugh. *Sets for Mathematics*. Cambridge University Press, 2003.

[10] Steve Awodey. *Category Theory (Oxford Logic Guides)*. Oxford University Press, USA, 2006. I

[11] Joseph A. Goguen and Francisco J. Varela. Systems and distinctions; duality and complementarity. *International Journal of General Systems*, 5(1):31–43, January 1979. I, II B

[12] A C Ehresmann and J.P. Vanbremeersch. *Memory Evolutive Systems; Hierarchy, Emergence, Cognition, Volume 4 (Studies in Multidisciplinarity)*. Elsevier Science, 2007.

[13] Aloisius Louie. *More Than Life Itself: A Synthetic Continuation in Relational Biology (Categories) (Volume 1)*. Ontos Verlag, 2009. I

[14] B. Jacobs. *Categorical Logic and Type Theory*. Elsevier Science, 1998. I

[15] W. Fontana and Leo W Buss. The barrier of objects: From dynamical systems to bounded organization. In J Casti and A Karlqvist, editors, *Boundaries and Barriers*, pages 56–116. Addison-Wesley, Redwood City, MA, 1996. II A

[16] David Ellerman. Adjoints and emergence: applications of a new theory of adjoint functors. *Axiomathes*, 17(1):19–39, March 2007. II B

[17] David Ellerman. Counting distinctions: on the conceptual foundations of Shannons information theory. *Synthese*, 168(1):119–149, March 2008. II B, V A

[18] F. John Odling-Smee, Kevin N. Laland, and Marcus W. Feldman. *Niche Construction: The Neglected Process in Evolution (MPB-37) (Monographs in Population Biology, 37.)*. Princeton University Press, 2003. III A

[19] David C Krakauer, Karen M Page, and Douglas H Erwin. Diversity, dilemmas, and monopolies of niche construction. *The American naturalist*, 173(1):26–40, January 2009. III A

[20] Samir Okasha. *Evolution and the levels of selection*. Oxford University Press, USA, New York, 2006. III A

[21] S J Gould. Tempo and mode in the macroevolutionary reconstruction of Darwinism. *PNAS*, 91(15):6764–71, July 1994. III C

[22] A.J. Arnold and Kurt Fristrup. The theory of evolution by natural selection: a hierarchical expansion. *Paleobiology*, 8(2):113–129, 1982. III C

[23] Roy L. Crole. *Categories for Types (Cambridge Mathematical Textbooks)*. Cambridge University Press, 1994. IV C

[24] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. V A

[25] Samson Abramsky. Information, Processes, and Games. 2008. V B

[26] Samson Abramsky and Achim Jung. Domain Theory. In Samson Abramsky, Dov M. Gabbay, and T. S. E. Maibaum, editors, *Handbook of Logic in Computer Science Volume 3. Semantic Structures*, volume 3, chapter 3, pages 1–168. Oxford University Press, 1995. V B

[27] Paul-André Melliès. Categorical Semantics of Linear Logic. In *Interactive Models of Computation and Program Behavior*, chapter 1. Société Mathématique de France, 2009. V C