Biological Codes and Topological Causation*

Benjamin Jantzen, David Danks†‡

---

*

† To contact the authors, please write to: Benjamin Jantzen, or David Danks; Department of Philosophy, Baker Hall 135, Carnegie Mellon University, Pittsburgh, PA 15213; e-mail: jantzen@cmu.edu, or ddanks@cmu.edu.

‡ Portions of this paper first appeared in the first author's M.A. thesis. Thanks to Peter Spirtes and two anonymous referees for helpful comments and suggestions. Frederick Eberhardt, Douglas Perkins, and Richard Yeh also provided valuable discussions and comments on earlier drafts.

Abstract

Various causal details of the genetic process of translation have been singled out to account for its privileged status as a “code”. We explicate the biological uses of coding talk by characterizing a class of special causal processes in which topological properties are the causally relevant ones. This class contains both the process of translation and communication theoretic coding processes as special cases. We propose a formalism in terms of graphs for expressing our theory of biological codes and discuss its utility in understanding biological systems.

1. Introduction

Molecular biology is pervaded by talk of coding and information (Sarkar 1996; Maynard Smith 2000). For example, the process of translation—the causal process leading from RNA to primary protein structure—is almost always discussed explicitly in terms of a "code," often because of the special causal role played by RNA in the process. At the same time, the particular content of that talk is often deeply ambiguous. In the case of translation, a variety of details have been singled out to account for its privileged status. Maynard Smith (2000), for instance, suggests that translation is distinguished from other causally specific processes by the "arbitrariness" of the correspondence between codons and amino acids (though see (Godfrey-Smith 2000a)). In a similar vein, Godfrey-Smith (2000a; 2000b) argues that translation is unusual because of two aspects of the causal relation. First, the mechanisms by which an mRNA causes a protein to be produced are insensitive to any "direct" (Godfrey-Smith 2000a, 204) chemical affinities between codons and amino acids; any codon could (in a "chemical possibility" sense) be paired with any amino acid given an appropriate tRNA mediator. Second, the specification of protein primary structure from the sequence of mRNA monomers is in accord with a "combinatorial rule" (Godfrey-Smith 2000a, 204) that functions roughly as a locality constraint on the RNA $\rightarrow$ protein mapping: the primary structure of a segment of protein depends only on the structure of a corresponding small segment of RNA. More recently, Šustar (2007) has emphasized this locality condition, claiming that the causal role of DNA in protein production is privileged because local perturbations of a DNA molecule—e.g., a single base substitution—produce strictly local perturbations in the primary structure of the protein it causes.

All of these proposed conditions are, in important ways, vague. Moreover, even if they can be made sufficiently precise, it is unclear whether they offer anything more than a redescription of translation; a theory of coding should not simply be a restatement of the canonical case of biological coding. And even if one or all of these criteria serve to distinguish on independent grounds translation from processes such as glycolysis, no account is offered as to why such properties imply that translation is a *code*. Our central goal here is to argue for a particular explication of (at least) the biological uses of coding talk that is not subject to these concerns. In particular, we provide a precise characterization of a class of special causal processes that contains communication theoretic coding processes as a special case, as well as the canonical example of coding in molecular biology.

2. Topological Causation

We focus on biochemical processes that potentially warrant identification with codes, and in particular, on processes that result in the production of some macromolecule. In general, to say that some property *P* of an object is *causally relevant* in a process is to say that a potential manipulation of the value of *P* results in a different effect produced after introduction of the object at the start of the causal process under consideration. This account of causal relevance is closely related to the interventionist characterization of direct causation found in Woodward (2003), but is intended only to pick out one feature of causation, not provide a definition of it (see also (Eberhardt and Scheines 2006; Woodward and Hitchcock 2003)). The causally relevant properties for biochemical and

molecular biological processes typically include such features as molecular mass, charge distribution, and geometry.

Code-like processes, however, involve what we call *topological causation*: a casual process in which objects' topological[1] features—loosely, the neighborhood relations amongst connected subunits—are the causally relevant properties. In processes involving topological causation, the outcome is influenced by details about adjacencies of distinct subunits, but not necessarily by other features of the molecules. In topological causation, a specific intervention on the connectedness of distinct molecular subunits in the reactant molecule produces a profound change in what, if any, products result. In contrast, independently varying (within bounds) the high-level geometrical properties (e.g. shape) or physical properties (e.g. mass) of the reactant results in little or no change in the outcome of the process. Of course, most biological processes are not based on topological causation: in enzyme catalyzed hydrolysis, for example, the details of enzyme mass and charge distribution determine which substrate it interacts with and which bonds are

[1] This term does not refer to the "rubber-sheet geometry" of algebraic topology, which is concerned with the invariant properties of donuts and coffee cups and other continuous topological spaces. By 'topological properties' we mean to refer strictly to a relation of adjacency that holds amongst a finite, discrete set of elements, particularly things like atoms in a molecule. As early as 1955, Rashevsky was explicitly using "topological" in this sense when speaking of molecular structure. This usage is currently standard in the field of chemical graph theory (Pogliani 2000), which is also called "molecular topology" (Diudea, Gutman, and Jantschi 2001, 1).

cleaved; the "lock and key" model of enzyme action explicitly emphasizes the causal primacy of geometry; and so on. We argue, however, that causal dependence upon topology is exactly the relevant feature for a process to be potentially a code in (something like) the classical communication theory (CT) sense of the term.

As an illustration of topological causation, consider the case of translation. If the transcript to be translated is any moderately large mRNA molecule, then it will typically fold into a more compact and complex structure than a simple chain. One particularly common structure, formed through intramolecular base pairing, is the RNA "hairpin." In a hairpin, two distant segments of a single RNA strand stick together, resulting in the formation of a loop between them. Under normal cell conditions, the stability of a typical hairpin structure depends on the ambient cytosolic salt concentration. *In vitro*, the same is true; salt concentration determines the conformation of the RNA and one can force either the hairpin or chain-like geometry to dominate. In the case of *in vitro* translation of mRNA transcripts using a cell-free system[2], salt concentration may be changed (within limits) without adversely affecting translation accuracy (though translation rates may suffer).

Suppose then that we have an mRNA transcript, M, which causes the production of some protein P under conditions in which our *in vitro* translation system functions "normally" (produces consistent polypeptide products, etc.). Suppose further that the transcript M

[2] A cell-free system is simply a solution of the components of the translation system maintained in a test-tube rather than a living cell.

maintains either a hairpin or chain-like conformation depending on salt concentration. We can now test the assertion that the set of neighborhood relations of distinct codons (and not molecular geometry) is the causally relevant factor determining the precise protein product of the process. If we make an appropriate[3] change in a single base of M (by substitution) we can, for a moderately sized hairpin, create a new molecule M' that is effectively identical in geometry to the original M. However, the polypeptide product will be P', a protein differing from P by a single amino acid. An intervention on neighborhood relations alone produces a different outcome. Now suppose that we instead raise the salt concentration to a point at which M is no longer in a hairpin conformation but the other components of the translation apparatus are still capable of functioning. The resulting effect is still production of P; an intervention on geometry alone has no impact on the terminus of the causal chain.

Though we speak of intervening on high-level geometry without altering topology, topological properties are not independent of low-level geometric and physical properties. What is considered a subunit or an adjacency relation will depend on the underlying physics and chemistry of a given system[4]. For instance, in the hairpin example, codons were taken to be coherent subunits, stable objects which can stand in the

[3] Because the genetic code is degenerate, not all changes will be causally efficacious.

[4] This dependence of topology on low-level physical and geometric properties may in fact be an instance of full-fledged supervenience in some cases or domains. Whether or not it is actually a supervenience relation is irrelevant to both the notion of biological code we develop here, and the question of whether such codes exist.

relation of adjacency to one another; the adjacency relation was based on the presence of a covalent bond between the sugars of distinct codons, and not on hydrogen bonding between the bases. A codon has a particular charge, geometry, etc. that allows it to be distinguished as a subunit, and so the topology of the hairpin depends in some sense on the physical and chemical properties of the things identified as subunits. But the choice of which molecular components constitute subunits and which chemico-physical relations count as neighborhood relations is crucially not arbitrary in many domains. Given the typical energies at which translation occurs, hydrogen bonds are much shorter-lived than covalent bonds, making the latter natural candidates for an adjacency relation. Similarly, atoms and—as we argue below—certain groups of atoms constitute stable and chemically distinct objects that are in some sense natural subunits. This situation is in contrast to the conventionality of the low-level characteristics of subunits and adjacency relations in human communication codes. Morse Code, for instance, is a correspondence between patterns of dots and dashes and characters of the English alphabet, represented by real-world objects, such as long and short flashes of light or pulses of high voltage on an electrical cable. It is entirely arbitrary what we choose to constitute a short pulse or a high voltage, so long as we are consistent; similarly, whether a gap counts as an adjacent pulse or an entirely new message is also arbitrary. For biochemical systems that instantiate codes, on the other hand, the choice is made "by nature," and so there are objectively better or worse ways to identify subunits and the associated topology.

The notion of topological causation corresponds to a basic intuition about CT codes: real-world instantiations of encoding and decoding depend only on topology, rather than other

features of the signal. The transmission and reception of a message in Morse Code over electrical cable depends on two causal processes: (i) the production of a series of pulses of high voltage encoding some printed English text and (ii) the production of a printed English text corresponding to the received series of pulses. Each of these processes, as they are carried out in the world, involve topological causation, since the actual lengths of light pulses, or the exact values of the high and low voltages, are not causally relevant in determining which English text is produced in decoding. The arrangement—the collection of neighborhood relations amongst dots and dashes, and the existence of particular objects to fill those roles—is all that matters. Similarly, insofar as the text of this paper plays some causal role in eliciting particular thoughts or concepts, only the neighborhood relations amongst characters are relevant. The text color, chemical properties of the ink, or actual print size are all causally irrelevant (at least to a first approximation).

Topological causation also provides a coherent explanation for the special features of translation highlighted by Godfrey-Smith (2000a). Specifically, he says that the causal process of translation appears to be a code because "…genetic specification of protein primary structure is done via a combinatorial rule and via mechanisms that are insensitive to any direct chemical affinities between codons and the corresponding amino acids" (2000a, 204). The appeal to a "combinatorial rule" can be read as the positive assertion that combinations of distinct subunits—the topological properties of an mRNA transcript—are causally relevant to determining which protein gets produced. The

requirement that the mechanisms involved are insensitive to the chemical affinities of codons implies the *absence* of causal impact for non-topological properties.

Both instantiated CT codes (e.g., Morse code) and some biological processes (e.g. translation) exhibit features of topological causation, but similarity of features is obviously insufficient to show that they are identical. One must further show that the relevant features of those biological processes can be systematically formalized and expressed in terms of the mathematical relation of CT coding. Since the relevant features are topological, any formal representation of those processes and molecules as "codes" must be capable of representing the full set of neighborhood relations within any given molecule. Graphs, and undirected graphs more specifically, provide the most appropriate representations of topological properties. An undirected graph has vertices and edges, where the edges are unordered pairs of vertices that represent, in the most general sense, neighborhood relations amongst objects (vertices). Typically, vertices in an undirected graph are unmarked. Because we are interested in representing neighborhood relations amongst subunits of many different and discernible types (such as atoms of particular elements or codons of a particular sort), we augment the graphs with a labeling function that maps vertices to particular types or "marks." These undirected graphs with marking functions are, we argue, sufficient and minimal abstractions for representing the properties that are relevant in topological causation.[5]

[5] All formal details about the framework, including precise details about this type of graph, can be found in the appendices.

The simplest schema for representing molecules using this framework is: atoms map to vertices marked by an appropriate label indicating the element, and covalent bonds map to edges. The motivation for this identification is simple: atoms are, so far as biochemistry is concerned, the smallest, distinct and indivisible units of a molecule, and chemical bonds are the physical entities that maintain atoms in fixed neighborhood relations. Of course, we are not claiming that this is the only conceivable set of correspondence rules between molecules and graph representations of them. But regardless of the specific rules chosen, the edges of the graph must correspond to the stable patterns of subunit arrangement in the molecule. Stability is largely a matter of degree, and one's choice of threshold for considering a neighborhood relation to be stable will likely depend on which molecular system is under consideration. In some instances, one might consider hydrogen bonds sufficiently stable to warrant an edge, even though their bond life is significantly shorter than that of a covalent bond. Ionic bonds, given their high binding energy, are arguably equivalent to covalent bonds and one might choose to represent them the same way. At the present time, our account of biological codes requires only that we represent all covalent bonds, including double and triple bonds, with single edges. If vertices and edges are depicted as points and lines, this representation is similar to the standard notation in organic chemistry, though we do not use that precise shorthand (and we disregard differences between covalent bonds). An example is shown in Figure 1, where uracil is depicted in both the standard representation of organic chemistry and as the marked graph obtained using our correspondence rules.

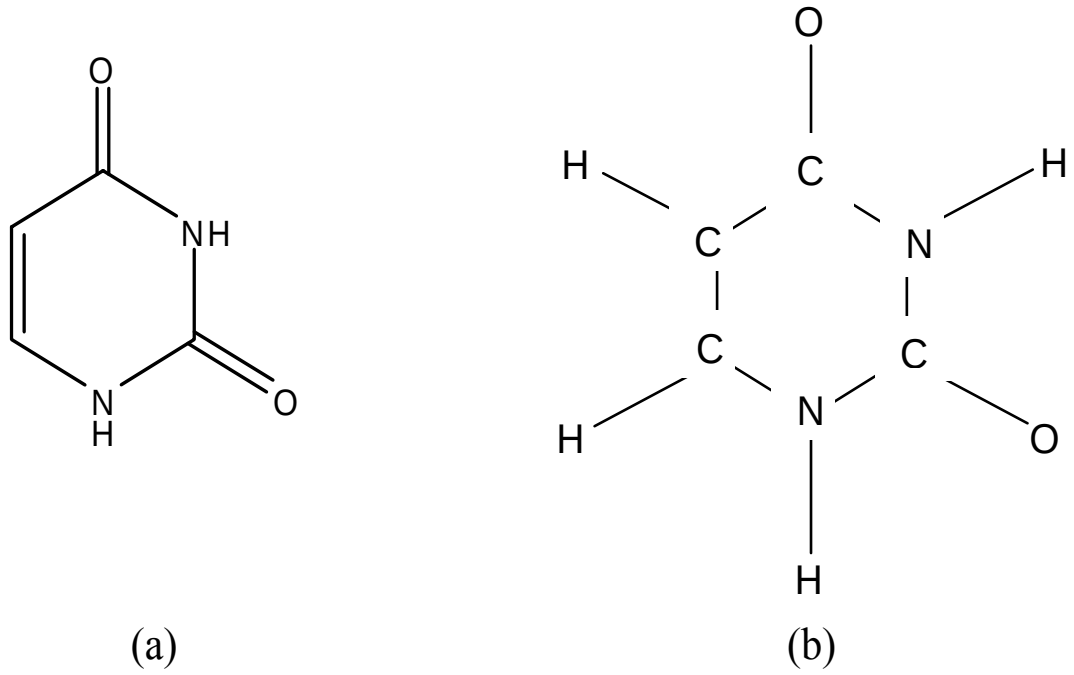

Figure 1 Uracil represented (a) in standard chemical notation and (b) as the corresponding graph using the relevant correspondence rules.

Importantly, the justification of the undirected graph representation for topological properties of complex molecules makes no reference to coding. Rather, there are independent reasons to suppose that such graphs are faithful representations of relevant neighborhood relations in biological molecules. Atomic nuclei are the smallest stable units capable of chemical recombination. They are neither created nor destroyed in biochemical processes and are, so far as physics is concerned, the smallest objects capable of stable neighborhood relations in a macromolecule.[6] Similarly, covalent and

[6] Subatomic particles, e.g. neutrons and protons, are found in non-localized states in which neighborhood relations are not well-defined.

ionic bonds are arguably the basic physical entities that maintain these neighborhood relations between atoms. The motivation for these correspondence rules is thus strictly physical and invokes only chemical and physical concepts. Coding relations are not required by our representation scheme, but arise (if at all) only as a matter of contingent fact.

3. The Proposed Formalism

Any formal explication of 'biological code' should provide for the unproblematic uses of the term in theoretical and experimental contexts. At the very least, any proposal must imply that the "genetic code" in translation—the correspondence between triplets of RNA and amino acids—is an instance of biological coding. A suitable formalism should also enable inquiry into whether various non-genetic systems in molecular biology instantiate codes. The definition of 'biological code' should not be simply an *ad hoc* re-description of translation; if it is not at least possible that other, non-genetic causal systems are codes, then the term 'biological code' provides no content beyond the term 'translation.' More generally, this requirement that any formal definition potentially apply to previously unconsidered systems ensures that it can serve a theoretical role in explanation and prediction. This section presents a formal theory of coding that is (a) based on marked, undirected graphs, and (b) satisfies these desiderata for such a term, including a basis in the concept of topological causation. Moreover, as shown later, the classical communication theory of coding emerges as a special case of this framework.

In general, coding is a relation between sets (possibly of infinite cardinality) of objects. Our particular focus is on sets of graphs that can be defined in terms of (i) potential marked vertices; and (ii) recursive formation rules.[7] A set of graphs *S* is *complete* with respect to a set of formation rules and marked vertex types iff *S* contains all graphs that can be assembled from those vertices using the formation rules (possibly infinitely many applications of formation rules). We define a *graphical code* to be a mapping—either one-one or many-one—from a complete set of graphs G_A to a not-necessarily-complete set of graphs G_B. Such a relation between G_A and G_B is a code because each graph "source message" in G_A corresponds to exactly one graph "code message" in G_B. The correspondence rules between biological macromolecules and their graph representations mean that this extended formalism for coding can be immediately applied to biological processes. The resulting definition of coding for biological processes can be stated informally as follows:

> A *biological code* is any relation between two sets of macromolecules *A* and *B* that satisfies three conditions:
>
> 1. Elements of *A induce* elements of the set *B* under normal cell conditions.
> 2. G_A (the set of graphs representing the members of *A*) is *complete* relative to formation rules corresponding to actual biological processes.

[7] All technical details concerning coding—including definitions of completeness, formation rules, and so on—are provided in Appendix A. A precise statement of natural source alphabets (see below) is located in Appendix B.

3. There is a graphical code between G_A and G_B such that if $g_a \in G_A$ maps to $g_b \in G_B$, then $a \in A$ must induce $b \in B$.

This definition introduces an additional term—'induce' in conditions (1) and (3)—as shorthand for a particular type of causal arrangement. Suppose that the introduction of some structure $a \in A$ initiates a causal chain resulting in the production or presence of any structure in B under normal cell conditions. We say that *a induces* $b \in B$ just when a has this causal role, and it *always* causes the presence of the particular member b. Condition (1) in the definition of biological code is thus not as strong as it might seem, since most molecular cell processes can be described in this way. There are, however, various stochastic processes that do not have deterministic outcomes. For instance, in microtubule formation, introduction of the molecules of α- and β-tubulin initiates the production of a microtubule of stochastically determined length. Such processes are excluded by this definition from being biological codes, precisely because there is no definite "message" for which the initiating structure can be said to code. A salient example in which condition (3) fails to hold may be the relation between DNA sequences and mature mRNA transcripts. Because there are often multiple ways in which a given precursor mRNA transcript may be spliced, there are generally multiple mature mRNA transcripts corresponding to a single DNA sequence. The mapping from the set of graphs representing DNA molecules to the set representing mature mRNA transcripts is generally one-many or even many-many.

Clearly, translation is a relation of biological coding between the set of possible mature mRNA transcripts—those that have already been spliced—and the set of polypeptide products. The process of translation satisfies condition (1), since introduction of a particular mRNA transcript into the proper portion of a functioning cell always causes the production of the same polypeptide, if it causes production of anything at all. There are natural correspondence rules for representing the set of all possible mRNA transcripts (including arbitrarily long ones) by a set of graphs, and that set of graphs is complete with respect to the collection of marked vertices representing atoms, and with respect to a set of simple formation rules that model the biochemical facts of transcript production. Finally, there is a many-one mapping from this complete set of graphs to the set of graphs representing polypeptide products that captures the biological mapping from mRNA transcripts to polypeptides. Use of coding talk in the canonical example—the "genetic code"—is thus justified according to this definition.

This definition of biological code places no inherent restrictions on which biological processes might be codes. For instance, glycans are complex molecules composed predominantly of linked sugars, and are involved in an enormous array of molecular cell processes. Causal processes involving glycans may potentially provide instances of complex biological coding. Unlike the topological structure of DNA or RNA, the sugar portions of glycans are typically highly branched (Varki et al. 1999). Furthermore, the effect of introducing a particular glycan into certain signaling reactions appears to be determined by the sequence of simple sugars along multiple branches. Gabius (1997; 2000) has argued that the discrete arrangements of groups of short, nearby branches on

some glycans constitute a code that determines the signal receptor to which the glycan will bind. If the set of possible glycans is complete with respect to the cellular processes that generate them, then such a system would be an excellent candidate for a code involving non-chain-like molecules. This code would share what we have argued are the essential properties of code relations—namely that sets of molecules are related through topological causation—but would involve macromolecules with more complex subunit relations than those in the chain-like molecules of the genetic code. In particular, coding relations involving glycans cannot be captured using representations based on strings of characters.

The notion of topological causation emphasizes the importance of neighborhood relations amongst distinguishable subunits, but does not specify any necessary scale for the subunits. In the particular case of macromolecules, the most natural correspondence rules provide graphs with vertices for atoms, and edges for the stable bonds between atoms. While those rules allow for unambiguous application of the definition of biological code without *ad hoc* decisions about representation, they also sometimes include too much information: namely, the total set of neighborhood relations amongst atoms. In many cases, the *relevant* set of distinguishable subunits is some larger collection of atoms (e.g. the nucleotide monomers in an RNA polymer). We define a *natural alphabet* to be any set of subgraphs that can be used to express the graphs in a graph set in terms of "characters" larger than a vertex. *Natural source alphabets* are the "coarsest" (i.e., with the largest subgraph "characters") natural alphabets in which one can rewrite the graphs of the source set and still preserve the coding relation in which it stands. A natural source

alphabet represents the set of subunits—possibly much larger than atoms—whose large-scale neighborhood relations are the causally relevant ones. Given sets of graphs that stand in a coding relation, we have developed reliable, though typically computationally intractable, algorithms for discovering natural source alphabets.

Natural source alphabets are particularly important for biological systems. Once one set of molecules is known to stand in a coding relation with another, the natural source alphabet provides the causally relevant subunits. Representation in terms of natural source alphabets eliminates irrelevant detail and fixes, in some sense, the proper scale at which to model coding processes. They also form a natural basis on which to construct definitions of information applicable to molecular processes.

The notion of 'graphical code' presented here is a generalization of communication theory coding. It is easy to see qualitatively that standard CT coding is a "special case" of graphical coding, since the graph sets containing only linear, chain-like graphs (graphs for which every vertex has no more than two neighbors) are obviously isomorphic to sets of strings. For this restricted collection of graph sets, 'code' corresponds to[8]:

[8] A typical definition of code in terms of strings is: "Let the set of symbols comprising a given alphabet be called $S = \{s_1, s_2, \ldots, s_q\}$. Then we define a code as a mapping of all possible sequences of symbols of S into sequences of symbols of some other alphabet $X = \{x_1, x_2, \ldots, x_p\}$" (Abramson 1963). This is clearly equivalent to our notion of 'code' for the restricted collection of chain-like graph sets.

> Given a finite alphabet of symbols $S = \{s_1, s_2, \ldots, s_n\}$, a *communication theory code* is a one-one or many-one mapping from a complete set of strings A to a set of strings B.

The notion of 'graphical code' that we have presented here also emerges immediately if one starts with a standard CT definition of coding, but generalizes to use sets of marked vertices instead of alphabets, and sets of marked graphs instead of sets of strings. Our explication of coding is a conservative extension of the standard communication theory notion of 'code' to allow for the possibility of undirected, marked graphs as the messages. It is the most natural understanding of coding between graphical objects, rather than between strings.

4. Two objections

Our appeal to classical CT runs counter to a pervasive sentiment in the philosophical literature: namely, that CT is inadequate for understanding codes in biology. At least in part, we think that this categorical rejection could derive from a mistaken conflation of "codes for" and "carries information about." If these two notions are taken to be equivalent, then the so-called "parity thesis" would ipso facto defeat any CT-based theory of coding. The parity thesis is the claim that one cannot use the concept of CT *information* to pick out a privileged class of bio-molecules. Specifically, Griffiths (2001, 396) suggests that the application of CT to biological systems requires the adoption of a "causal notion of information" derived from Dretske (1981), and so information is passed between two systems whenever a channel exists between them. This is the case "when

the state of one is systematically causally related to the other, so that the state of the sender can be discovered by observing the state of the receiver" (Griffiths 2001, 397). Because causal information makes reference only to the causal relations, there is no non-arbitrary way to distinguish between channel conditions (or noise sources) and the signal: there is parity amongst the set of causally linked systems (i.e., the various biological molecules and cell structures) and so none can be privileged as the information source. Thus (the argument continues), DNA cannot be viewed as special in carrying information because other factors can just as well be referred to in this way; cell conditions could be the "signal" and DNA the "channel conditions/noise." If biological coding is based on or identical to causal information, then we have no way of determining which biological molecules are "signal" (e.g., presumably DNA) and which are "noise."

In contrast, we take coding to be conceptually more fundamental than information, and aim (in future work) to use the present concept of biological code as the basis for a theory of biological information. In this, we more closely follow Shannon's original approach in developing CT (Shannon 1948). While it is only implicit in Shannon's definition of a communication system, a channel is partially defined by a set of coding relations: a communication system exists only when a transmitter exists, and a transmitter is defined by the coding relation which it instantiates. Similarly, receivers are defined by the *de*coding relation that they instantiate. In an abstract communication system, noise in a channel is a source of stochastic errors introduced into the signal. Unlike transmitters, sources of noise (and other channel conditions) do not stand in coding relations with the destination. There is thus an asymmetry in the definition of a communication system that,

in all but the most trivial cases, enables us to distinguish the signal from the noise in a principled manner. In Shannon's account, the presence of a coding relation is at least logically independent of whether there is transfer of information, and perhaps even logically prior to it. We agree that parity is a theoretical possibility between two systems that causally influence a third, but deny that all such causal relations fit the abstract schema of a communication system; in practice, we do not get to choose which systems may be viewed as sources and which as channel conditions or noise sources. By requiring that the coding relation hold, some physical systems are privileged as instantiations of communication theoretic transmitters or receivers and parity is broken—the parity thesis does not preclude a substantive application of communication theory, at least with respect to coding. By appealing to the fact that biological coding relations (which are independently discoverable) characterize the channel but not sources of noise or background conditions affecting the channel, our notion of biological coding might be used to provide a theory of biological information that is immune from concerns over parity.

Having established that there are no prima facie inconsistencies in taking a CT approach, it remains to defend our choice of formal representations. The coding relation holds between sets of objects, and so the fact that those objects are graphs does no work in *explaining* what a code is. The properties of a code do not depend on the properties of graphs; rather, graphs are used to express the features of a code. In that case, one might object that the formalism could just as easily be developed using some other type of representation, such as strings, and so nothing is really added by this generalization of CT

coding. Our justification for graphs is two-fold. First, graphs allow for perspicuous and easily defended correspondence rules between molecules and their appropriate representations. Analogous rules for other representational schemes would necessarily be more complex. For instance, one might consider coding relations using representations of DNA as strings of characters, as in Yockey (1992). There is, however, no a priori reason to represent DNA as a string of characters, unless one already knows that DNA is involved in a coding relation with a particular structure. But such a justification is clearly circular, as the coding relation can only be justified by use of the string representation of DNA. Even if there were some independent justification of string representation correspondence rules for DNA, there is no a priori reason to think that all molecules involved in biological coding can be represented as strings of characters using the same rules.

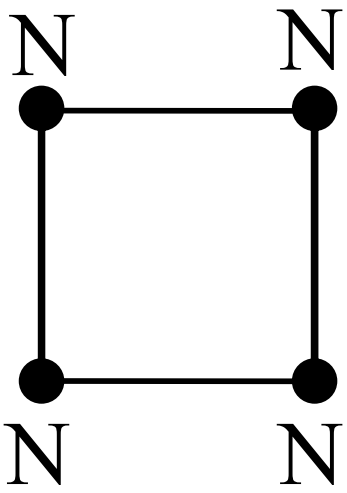

Figure 2 A marked graph containing a cycle.

Second, there is a clear correspondence between properties of graphs and properties of molecules. Graphs allow properties of the structures they represent to be directly "read off." To make this clear, consider a simple example. Suppose the marked graph shown in Figure 2 is a representation of a simple molecule—maybe a nitrogen ring of some sort—and that this graph was constructed in accord with the correspondence rules described above. The graphical property "has a cycle" (i.e., there is a closed path) corresponds directly to the ring-like topology of the molecule being represented. Consider instead the equivalent representation in terms of an adjacency matrix[9]:

0 1 0 1

1 0 1 0

0 1 0 1

1 0 1 0

This matrix may in turn be represented by the string: 0101101001011010. However, there is no straightforward correspondence between properties of this string and properties of the molecule. Of course, properties of the molecule could be extracted from the string, but the relevant transformation is neither transparent nor readily extensible to other molecules. Unlike the graphical representation, properties of the string representation are in part determined or mediated by the intervening representational apparatus (e.g., adjacency matrices).

[9] An element a_{ij} of an adjacency matrix is 1 if there is an edge between the i^{th} and j^{th} vertices of the graph it describes and 0 otherwise.

There are arguably an infinite number of equivalent representations, and so we claim that one should prefer a representational scheme that has both straightforward correspondence rules, and also clear relationships between properties of the representation and of the represented. A representational scheme that permits one to directly read off properties is a much more useful scheme than one that does not. Graphs are the most natural and, we contend, the most useful such representation.

5. Conclusion

A standard claim in molecular biology (and philosophy of biology) is that some of the molecules or molecular processes in living cells form a privileged class that can be described, perhaps heuristically, using 'coding' talk. We have argued that these processes are in fact instances of topological causation: casual processes in which the relevant properties are the neighborhood relations of subunits. By representing molecules with marked, undirected graphs, we conservatively extend classical coding theory to formulate a precise definition of biological code that is not *ad hoc*, that captures the canonical case of translation, and that can substantiate theoretical uses of code terminology in molecular biology. Our theory of biological codes also facilitates the identification of natural source alphabets, the causally relevant subunits involved in a coding process. The implications of natural source alphabets for an information theory of biology are clear: where before only *ad hoc* reference to the sequence of chain-like polymers was used to justify measures of information in macromolecules, now relevant units of information may be discerned in systems that are isomorphic to part of a communication system in a rigorous

sense. This gives both a principled reason why the monomers of DNA should be considered possible states of the source and provides a tool for evaluating information in novel settings. Of course, such an information theory remains to be developed, but the present account gives a principled basis for the language of coding in molecular biology.

Appendix A: Biological codes

An *undirected graph G* is a pair of finite sets: the vertex set, *V(G)*; and the edge set, *E(G)*. *V(G)* contains *n* vertices. *E(G)* contains *m* unique two-element subsets (unordered pairs) of *V(G)* for which both elements are distinct. A *marked graph* is a graph *G* that has associated with it a marking function L_G:*V(G)* → *M* that maps vertices of the graph onto a set *M* containing a finite number of elements. The elements of *M* are understood to be labels, indicating or marking each vertex in the graph with a type.

An *alphabet* $S = \{s_1, s_2, \ldots, s_n\}$ is a finite set of single marked vertices (which are, of course, graphs). A *set of formation rules* associated with an alphabet *S* is denoted $F_S = \{f_i \mid i = 1, 2, \ldots, m\}$ and consists of a finite set of partial functions that take marked graphs to marked graphs, where each function is a composition of four primitive operations: (1) addition of a marked vertex, (2) addition of an edge, (3) removal of a vertex, and (4) removal of an edge. Additionally, a set of formation rules must satisfy the following conditions:

(1) $\exists i(f_i(\varnothing) \neq \varnothing)$ ($\varnothing$ is the empty graph)

(2) Let $B = \cup_j f_j(\varnothing)$ be the set of "base cases" produced by F_S. Then there exists at least one $f_j \in F_S$ such that for some $\beta \in B$

(2a) $f_j(\beta) = \tau$ and $\tau \notin B$

(2b) $f_j(\beta)$ and β have in common at least one vertex isomorphic to one in *S*

Condition (1) ensures that there are "base cases" generated by the formation rules. Condition (2) implies that at least one of the formation rules is non-trivial, in the sense of both using the input graph, and producing an output graph that is related to the input graph.

Now use the following notation to describe the repeated application of formation rules:

$\Sigma_1 = \cup_j f_j(\varnothing)$

$\Sigma_2 = \Sigma_1 \cup (\cup_{i,j} f_i(\sigma_{1j}))$

$\Sigma_3 = \Sigma_2 \cup (\cup_{i,j} f_i(\sigma_{2j}))$

$\vdots$

where the σ_{nj} are the members of Σ_n. A *complete graph set* Σ with respect to S and F_S is defined as follows:

$\Sigma = \cup_j \Sigma_j$ (complete graph set)

Given the above definitions, we can now provide a formal statement of the notion of biological coding: The set of molecules A is a *code* for the set of molecules B if and only if:

(1) Elements of the set A induce elements of the set B under normal cell conditions. Every $a \in A$ induces some $b \in B$.

(2) Let G_A be the set of graphs representing the elements of A with respect to an alphabet of vertices S. Let F_S be a set of formation rules representing the chemical or physical processes by which elements of A are generated *in vivo*

from the components represented by S. Then G_A is complete with respect to S and F_S.

(3) There is a mapping (one-one or many-one) from G_A to G_B such that if g_a maps to g_b then a induces b where $g_a \in G_A$, $g_b \in G_B$, $a \in A$, $b \in B$ and g_a represents a and g_b represents b.

In the body of the paper, we argued that this definition generalizes the communication theory account of coding, which is defined on strings. Strings can be represented as graphs in which every vertex has degree less than or equal to two, though one must attend to certain details. Strings are inherently directional, in that they have a start and end: the string "aab" is not identical to "baa". However, the chain-like graphs *a*—*a*—*b* and *b*—*a*—*a* are identical. One can straightforwardly represent the directionality of strings by introducing additional elements into the set of marks. For instance, a mark *a** can be used to indicate a "left" vertex of type *a*. If the formation rules and marking function are appropriately designed, then the resulting complete graph set is essentially a complete string set and the classical theory can be recovered.

Appendix B: Natural source alphabets

Consider a subset $S \subseteq V(G)$ of the vertex set of the graph G. The *subgraph induced by S*, denoted $\langle S \rangle$, is the graph consisting of the vertices in S and all of the edges u—v in G such that $u, v \in S$ (Buckley and Lewinter 2003). If G is a marked graph, then the vertices of the induced subgraph retain the same values of the marking function as they did in G.

We can now capture the important idea that a set of subgraphs (and not just vertices) constitutes the appropriate alphabet for a set of graphs. A finite set N of connected, marked graphs is a *natural alphabet* of a set of graphs Σ if and only if:

1. For every $\sigma \in \Sigma$ there exists one and only one partition Q of $V(\sigma)$ such that every $q \in Q$ induces a connected subgraph $\langle q \rangle$ on $V(\sigma)$ that is isomorphic to some $\eta \in$ N.
2. Suppose Q is the partition of $V(\sigma_1)$ that satisfies condition (1) for some $\sigma_1 \in \Sigma$. Let B_1 be the set of connected graphs induced by the members of Q on $V(\sigma_1)$. Note that, since Q satisfies condition (1), $B_1 \subseteq$ N. Let $B = \cup_i B_i$ be the set of all graphs induced on the members of Σ by partitions satisfying (1). Then it must be the case that B = N.

Suppose that N is a natural alphabet for some graph set Σ. We use Σ_N to denote the set of graphs in Σ *rewritten* in the "language" of N. Specifically, for each $\sigma \in \Sigma$, there is a corresponding $\sigma_N \in \Sigma_N$ that is constructed in the following manner. Let Q be the partition

of $V(\sigma)$ satisfying condition (1) above. That is, Q is the partition of $V(\sigma)$ such that every $q \in Q$ induces a connected subgraph on $V(\sigma)$ that is isomorphic to a member of N. Then for each $q_i \in Q$ there is a vertex $v_i \in V(\sigma_N)$. The vertices of σ_N are marked and so there is associated with Σ_N a marking function L: $V(\sigma_N) \rightarrow M$ where M is a finite set containing one element for every member of the natural alphabet N. Essentially, each vertex of a graph rewritten in the alphabet N is marked with a label indicating which subgraph was replaced by this vertex. The edge set of the new graph is determined via the following rule: if any vertex in q_i shares an edge with any vertex in $q_{j\neq i} \in Q$ then $\{v_i, v_j\} \in E(\sigma_N)$. Constructed in this fashion, $V(\sigma_N)$ and $E(\sigma_N)$ define the graph σ_N ("σ rewritten in N"). There is no requirement that distinct graphs in Σ have distinct corresponding graphs in Σ_N. For example, suppose N contains *a*—*b* (marked by *x* in Σ_N) and the source set Σ contains:

(i) *a*—*b*—*a*—*b*; and (ii) *a*—*b*—*b*—*a*. Both of these graphs correspond to the graph *x*—*x* in Σ_N.

Suppose we have some set Σ (complete with respect to the formation rules in F) that stands in a coding relation with some set Ξ. A natural alphabet N of Σ is a *natural source alphabet* if:

(1) There exists a set of recurrence relations on graphs, F_N, which makes Σ_N complete and which satisfies:

(1.1) $|F| = |F_N|$

(1.2) For all $\sigma_j \in \Sigma$, if $f_i(\sigma_j) = \sigma_{j'}$ for all j and for some $f_i \in F$ then there exists some $f_i^N \in F_N$ such that $f_i^N(\sigma_j^N) = \sigma_{j'}^N$ for all j. Here, σ_j^N is the graph σ_j rewritten in the alphabet N.

(1.3) If, for any $\sigma_j \in \Sigma$, σ_j and $f_i(\sigma_j)$ overlap (share at least one vertex) for all j, then so do σ_j^N and $f_i^N(\sigma_j^N)$.

(2) The average number of vertices per graph in the set Σ_N is as small as possible.

(3) For all $\sigma \in \Sigma$, if σ maps to ξ in the code set, then σ^N maps to ξ and there is *no* σ' such that $\sigma'^N = \sigma^N$ and σ' maps to some $\xi' \neq \xi$.

REFERENCES

Buckley, Fred, and Marty Lewinter (2003), *A friendly introduction to graph theory*. Upper Saddle River, NJ: Prentice Hall.

Diudea, Mircea V., Ivan Gutman, and L. Jantschi (2001), *Molecular topology*. Huntington, NY: Nova Science Publishers.

Dretske, Fred I. (1981), *Knowledge & the flow of information*. Cambridge, MA: MIT Press.

Eberhardt, Frederick, and Richard Scheines (2006), "Interventions and Causal Inference", in, *Proceedings of the 20th Biennial Meeting of the Philosophy of Science Association*.

Gabius, H. J. (2000), "Biological information transfer beyond the genetic code: the sugar code", *Naturwissenschaften* 87 (3):108-121.

Gabius, H. J., and S. Gabius (1997), *Glycosciences : status and perspectives*. London ; New York: Chapman & Hall.

Godfrey-Smith, Peter (2000a), "Information, Arbitrariness, and Selection: Comments on Maynard Smith", *Philosophy of Science* 67 (2):202-207.

——— (2000b), "On the Theoretical Role of 'Genetic Coding'", *Philosophy of Science* 67 (1):26-44.

Griffiths, Paul E. (2001), "Genetic Information: A Metaphor in Search of a Theory", *Philosophy of Science* 68 (3):394-412.

Maynard Smith, John (2000), "The Concept of Information in Biology", *Philosophy of Science* 67 (2):177-194.

Pogliani, L. (2000), "From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors", *Chem. Rev.* 100 (10):3827-3858.

Rashevsky, N. (1955), "Life, information theory, and topology", *Bulletin of Mathematical Biophysics* 17:229-235.

Sarkar, S. (1996), "Decoding "coding" - Information and DNA", *Bioscience* 46 (11):857-864.

Shannon, C. E. (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal* 27 (3):379-423.

Šustar, Predrag (2007), "Crick's notion of genetic information and the 'central dogma' of molecular biology", *Br J Philos Sci* 58 (1):13-24.

Varki, Ajit, Richard Cummings, Jeffrey Esko, Hudson Freeze, Gerald Hart, and Jamey Marth, eds. (1999), *Essentials of glycobiology*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Woodward, James (2003), *Making things happen : a theory of causal explanation*, *Oxford studies in philosophy of science*. New York: Oxford University Press.

Woodward, James, and Christopher Hitchcock (2003), "Explanatory Generalizations, Part I: A Counterfactual Account", *Noûs* 37 (1):1-24.

Yockey, Hubert P. (1992), *Information theory and molecular biology*. Cambridge ; New York, NY: Cambridge University Press.