When Information Theory Is No Longer Theory

Root Gorelick
Department of Biology and School of Mathematics and Statistics
Carleton University, Ottawa, Ontario, Canada
Root_Gorelick@carleton.ca

Biological Theory 1:3 was a thematic issue on information theory. Inter alia, it covered the potential centrality of information theory (Callebaut and Collier 2006; Pfeifer 2006), whether information theory provides an approach that is reductionist or synthetic (i.e., top-down or bottom-up; Harms 2006), and several applications (e.g., Harms 2006; Pfeifer 2006). There is little doubt regarding the utility of information theory in biology, as can be seen from the applications of mutual entropy for measuring breadth of ecological niches (Colwell and Futuyma 1971), biodiversity of metacommunities (Gorelick 2006), division of labor of social insects (Gorelick et al. 2004), efficiency of the genetic code (see below and Yockey 1992), linkage disequilibrium (Liu and Lin 2005), and several other applications alluded to in the special issue. Here, I take a step back from such practicalities and describe how information theory devolves into a methodological tool, rather than a theory, when used in biological theory. I also caution that before applying any mathematical tool in biology, including those from information theory, we need to carefully specify the domain and range of mathematical functions.

In part, the controversy over the role of information theory in biological theory (Callebaut and Collier 2006) stems from lack of any cogent definition of theory. Philosophers of science have struggled with defining scientific theory (Suppes 1967; Nagel 1971; Lloyd 1988). They describe a logical calculus underlying theory, but uniformly agree that there is something more to theory. “Theory is not just a body of facts or a set of personal opinions. It involves explanations and hypotheses that are based on available knowledge and experience. It is also dependent on conjecture and insight about how to interpret those facts and experiences and their significance” (Bunch 1979: 8). Philosophers of science also largely agree that there exists a fuzzy demarcation between theory and observation, a fuzziness that is even more evident in the biological and social sciences than in the physical sciences (Suppes 1967; Nagel 1971).

Information theory is part of mathematics, despite it largely having been developed by theoretical electrical engineers (Shannon 1948; Cover and Thomas 1991). Regardless of general definitions of theory and the nebulous nature of these definitions, there are important differences between biological theories and mathematical theories. One can formulate axiomatic underpinnings for both, but only biological theories counterpose theory and observation. Mathematical theories rely solely on internal self-consistency, with no comparison between theory and observation. Consequently, it is no surprise that various authors in the issue of *Biological Theory* on information theory see a disjunction between information theory and biological theory. Instead of trying to decide whether information theory is central in biological theory, maybe we should take the pragmatic and less ambitious route chosen by Pfeifer (2006) and simply use *results* from information theory as convenient methodological tools to guide our biological theory.

If biologists invoke theorems from information theory, and not just information-theoretic methods, would this constitute genuine biological theory? For example, should we consider Schneider’s (2000) direct invocation of Shannon’s theorems as a genuine example of biological theory? If a necessary criterion for a something to be considered theory is for that something to involve hypotheses, then the answer is no.

February 21, 2007; revised and accepted May 27, 2007

Whether or not applied to biology, mathematical theorems are not hypotheses. Indeed, there are no hypotheses in mathematics. From a Popperian perspective, unless we are willing to make the leap of deeming mathematics to be science—which I believe is unjustified—we cannot construe use of mathematical theorems to be a necessary aspect of a scientific theory. Thus it is crucial that philosophers of science formulate necessary and sufficient criteria for something to be defined as theory. Bunch's (1979) definition from the social sciences is only one possible way we may define theory, but her definition seems utterly sensible when applied to biological or physical sciences.

Information theory provides neither a bottom-up nor a top-down approach to unifying biology (cf. Harms 2006). Information theory is merely a methodological tool than can be applied in whatever ways benefit us by suitable conceptualization of the domain, range, and mapping between them. Just because mutual entropy maps a matrix onto a scalar does not make it any more top-down than bottom-up. Information theory is simply a branch of mathematics. And, like all of mathematics, it provides us with tools by which we can make scientific theories testable. However, in at least one sense, information theory plays a central role in biological theory: Mathematics bridges theory and observation by allowing us to rigorously test the hypotheses that are the output of scientific theories.

Let me focus on one specific example of using information theory in the mapping of genotype to phenotype. Harms (2006) notes that the nucleotide triplets CCG and CCA both code for the amino acid glycine. He then asks whether these two nucleotide triplets carry the same information (actually, he asks about their transcribed mRNAs, GGC and GGU, but the distinction is unimportant here). Although this is a valid question, it is not amenable to analysis using Shannon's information theory. Any ordered triple of objects drawn from a four letter alphabet will have identical information content in terms of the number of bits of information or even in terms of the optimal error-correcting code than can be constructed. Implicitly, however, Harms was clearly interested in something different: the *mapping* from nucleotide triplets onto amino acids. Shannon's mutual entropy describes the mapping from an entire domain (all possible nucleotide triplets) to an entire range (all amino acids, including stop codons). It makes no sense to ask whether a given element in the domain—in this instance, one nucleotide triplet—contains more mutual information than another. Mutual entropy instead provides a way to compare one entire genetic code with another. For example, we could ask whether the genetic code that exists in all meiotic genomes contains different mutual entropy than the alternative genetic codes in mitochondria or prokaryotes, where the mRNAs UGA, CUA, AUA can code for different amino acids, including the 21st amino acid selenocysteine. Similarly, we could ask whether there was a change in mutual entropy in the evolution from putative doublet codes to the modern triplet code. We may conjecture that mutual entropy is lowest with a doublet code, higher in alternative mitochondrial codes, and highest in meiotic genomes because of predilections to believe that biological systems grow more complex over evolutionary time. Of course, we would also expect this for purely numerological reasons because the maximum that mutual entropy can be is the logarithm of the possible number of elements in the domain of the mapping. Here, that means 16 for the doublet code and 64 for the triplet code. But even this numerological effect can be normalized by using the ratio of mutual entropy to marginal entropy (Gorelick 2006), as can relative rarity of amino acids (Gorelick et al. 2004). We not only need to ask interesting questions, but also need to invoke the appropriate mathematical tools, and not just metaphors. Mathematics, in general, and information theory, in particular, provide biologists with a suite of tools for testing our theories.

When used in biological theory, information theory is no longer theory, but rather methodology—at least if we consider hypothesis formation to be a necessary aspect of theory. *Information* may be central to biological theory. However, *information theory* cannot be central to biological theory simply because of the different or unspecified definitions of scientific theory and mathematical theory.

Acknowledgments

Thanks to four anonymous reviewers for their useful comments on this manuscript.

References

Bunch C (1979) Feminism and education: Not by degrees. Quest 5: 7–18.

Callebaut W, Collier J (2006) Editorial: Biological information. Biological Theory 1: 221–223.

Colwell RK, Futuyma DJ (1971) Measurement of niche breadth and overlap. Ecology 52: 567–576.

Cover TM, Thomas JA (1991) Elements of Information Theory. New York: Wiley.

Gorelick R (2006) Combining richness and abundance into a single diversity index using matrix analogues of Shannon's and Simpson's indices. Ecography 29: 525–530.

Gorelick R, Bertram SM, Killeen PR, Fewell JH (2004) Normalized mutual entropy in biology: Quantifying division of labor. American Naturalist 164: 677–682.

Harms WF (2006) What is information? Three concepts. Biological Theory 1: 230–242.

Liu ZQ, Lin SL (2005) Multilocus LD measure and tagging SNP selection with generalized mutual information. Genetic Epidemiology 29: 353–364.

Lloyd EA (1988) The Structure and Confirmation of Evolutionary Theory. Princeton: Princeton University Press.

Nagel E (1971) Theory and observation. In: Observation and Theory (Mandelbaum M, ed), 15–43. Baltimore: Johns Hopkins University Press.

Pfeifer J (2006) The use of information theory in biology: Lessons from social insects. Biological Theory 1: 317–330.

Schneider TD (2000) Evolution of biological information. Nucleic Acids Research 28: 2794–2799.

Shannon CE (1948) A mathematical theory of communication. Bell System Technical Journal 27: 379–423, 623–656.

Suppes P (1967) What is a scientific theory? In: Philosophy of Science Today (Morgenbesser S, ed), 55–67. New York: Basic Books.

Yockey HP (1992) Information Theory and Molecular Biology. Cambridge: Cambridge University Press.