Biol Philos
DOI 10.1007/s10539-009-9180-z

The transmission sense of information

Carl T. Bergstrom · Martin Rosvall

Received: 29 October 2008 / Accepted: 26 August 2009
© US Government 2009

Abstract Biologists rely heavily on the language of information, coding, and transmission that is commonplace in the field of information theory developed by Claude Shannon, but there is open debate about whether such language is anything more than facile metaphor. Philosophers of biology have argued that when biologists talk about information in genes and in evolution, they are not talking about the sort of information that Shannon's theory addresses. First, philosophers have suggested that Shannon's theory is only useful for developing a shallow notion of correlation, the so-called "causal sense" of information. Second, they typically argue that in genetics and evolutionary biology, information language is used in a "semantic sense," whereas semantics are deliberately omitted from Shannon's theory. Neither critique is well-founded. Here we propose an alternative to the causal and semantic senses of information: a *transmission sense of information*, in which an object X conveys information if the function of X is to reduce, by virtue of its sequence properties, uncertainty on the part of an agent who observes X. The transmission sense not only captures much of what biologists intend when they talk about information in genes, but also brings Shannon's theory back to the fore. By taking the viewpoint of a communications engineer and focusing on the decision problem of how information is to be packaged for transport, this approach resolves several problems that have plagued the information concept in biology, and highlights a

C. T. Bergstrom · M. Rosvall (✉)
Department of Biology, University of Washington, Seattle, WA 98195-1800, USA
e-mail: rosvall@u.washington.edu

C. T. Bergstrom
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
e-mail: cbergst@u.washington.edu
URL: http://octavia.zoology.washington.edu/

Published online: 01 October 2009

number of important features of the way that information is encoded, stored, and transmitted as genetic sequence.

Keywords Information · Evolution · Shannon theory · Natural selection · Entropy · Mutual information

Introduction

Biologists think in terms of information at every level of investigation. Signaling pathways transduce information, cells process information, animal signals convey information. Information flows in ecosystems, information is encoded in the DNA, information is carried by nerve impulses. In some domains the utility of the information concept goes unchallenged: when a brain scientist says that nerves transmit information, nobody balks. But when geneticists or evolutionary biologists use information language in their day-to-day work, a few biologists and many philosophers become anxious about whether this language can be justified as anything more than facile metaphor (Sterelny and Griffths 1999; Sterelny 2000; Godfrey-Smith 2000a; Griffths 2001; Griesemer 2005; Godfrey-Smith 2008). Why do the neurobiologists get a free pass while evolutionary geneticists get called on the carpet? When neurobiologists talk about information, they have two things going for them. First, there is a straightforward analogy between electrical impulses in neural systems and the classic communications theory picture of source, channel, and receiver (Shannon 1948). Second, information theory has obvious "legs" in neurobiology: for decades, neurobiologists have profitably used the theoretical apparatus of information theory to understand their study systems. Geneticists are not so fortunate. For them, the analogy to communication theory is less obvious. Efforts to make this analogy explicit seem forced at best, and the most successful uses of information-theoretic reasoning within the field of genetics rarely make explicit their information-theoretic foundations or make use of information-theoretic language (Crick et al. 1957; Kimura 1961; Felsenstein 1971; Freeland and Hurst 1998).

As a consequence, philosophers have concluded that the mathematical theory of communication pioneered by Claude Shannon in 1948 (hereafter "Shannon theory") is inadequate to ground the notion of information in genetics and evolutionary biology. First, philosophers have unfairly suggested that Shannon theory is only useful for developing a shallow notion of correlation, the so-called "causal sense" of information. Second, they typically argue that in genetics and evolutionary biology, information language is used in a "semantic sense"—and of course semantics are deliberately omitted from Shannon theory.

Neither critique is well-founded. In this paper we begin by summarizing the causal and semantic views of information. We then propose an alternative—a transmission sense of information—that not only captures much of what biologists intend when they talk about information in genes, but also brings Shannon theory back to the fore.

Causal view of information

Inspired by Dretske (1983), several authors (Sterelny and Griffiths 1999; Griffiths 2001; Godfrey-Smith 2008; Griffiths and Gray 1994) have explored Shannon theory as a grounding for information language in biology. They derive roughly the following picture: The key currency in information theory is the entropy $H(X)$ of a random variable X. The entropy is a measure of uncertainty in the realization of X. If X takes on value X_i with probability p_i, the entropy $H(X) = \sum_i p_i \log p_i$. The key statistic in information theory is the mutual information $I(X;Y)$ between two random variables X and Y. The mutual information, defined as $I(X;Y) = H(X) - H(X|Y)$, measures how much we learn about the value of X by knowing Y. Information is conveyed in Grice's sense of natural meaning (Grice 1957): whenever Y is correlated with X, we can say that Y carries information about X. There is no deep notion of meaning or coding here. "[W]hen a biologist introduces information in this sense to a description of gene action or other processes, she is not introducing some new and special *kind* of relation or property", Godfrey-Smith writes, "She is just adopting a particular quantitative framework for describing ordinary correlations." (Godfrey-Smith 2008). In this *causal sense* of information, genes carry information about phenotypes just as smoke carries information about fire, nothing more. If biologists are using information only as a shorthand for talking about correlations, this is a shallow use of the information concept compared to what we see in engineering and communications theory!

Not only is this sense shallow, it fails to capture the directional flow of information from genotype to phenotype that is the central dogma of molecular biology (Crick 1970). If, by "G has information about P," we mean only the mutual information $I(G;P) > 0$, then we are not acknowledging the direction of information flow from genotype to phenotype (Godfrey-Smith 2000a, 2008; Griffiths 2001). The reason is that the mutual information I is by definition a symmetric quantity; $I(G;P) = I(P;G)$. While mutual information is a key component of the deeper applications of Shannon theory that we will discuss later, for now let us consider it simply as a statistical quantity. Mathematically, the amount of information that knowing the genotype G provides about the phenotype P is always exactly equal to the amount of information that knowing the phenotype P provides about the genotype G (Fig. 1).

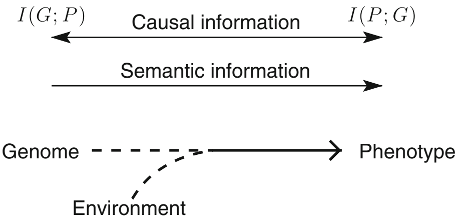


Fig. 1 Information theory restricted to a descriptive statistics for correlations. Information flows in both directions between genotype and phenotype, $I(P;G) = I(G;P)$, and, according to the parity thesis, there is nothing that privileges genes over environment. In a semantic notion of genetic information, genes represent phenotypes but phenotypes do not represent genes

Here it is helpful to consider a simple example. Suppose that the genotype to phenotype map is degenerate, such that there are $n\ k$ possible genotypes but only n possible phenotypes. Now it seems surprising that phenotype would tell us as much about genotype as genotype can tell us about phenotype. After all, knowing genotype predicts phenotype exactly, but knowing phenotype leaves us uncertain about which of k possible genotypes is responsible. The resolution to this puzzle is that mutual information measures how much an observation reduces our uncertainty, *not* how much residual uncertainty we face after making the observation. We can see this clearly from our example. If we observe the phenotype, this reduces the number of possible genotypes from a huge number (nk) to a much smaller number (k). If we observe genotype, this reduces the number of possible phenotypes from n to 1. Mutual information doesn't measure the fact that after making our observations there are k possibilities in one case and only 1 in the other; mutual information measures the fact that in both cases the observation reduces uncertainty n-fold. Thus assuming that all outcomes are equally likely, the entropy of genotype $H(G) = \log n\ k$, the entropy of phenotype $H(P) = \log n$, and the mutual information is $I(G;P) = \log k$.

An additional critique of the mutual information approach is that it fails to capture the sense of privilege that biologists often ascribe to the informational molecule DNA over other contributors to phenotype. So far as causal covariance is concerned, both genes G and environment E influence phenotype P—and in principle we can equally well compute either $I(G;P)$ or $I(E;P)$. So it seems that Shannon theory has no way of singling out DNA as an information-bearing entity.

This criticism is formalized as the parity thesis, and is crafted around an important result in information theory that the roles of source and channel conditions are exchangeable (Griffths and Gray 1994). Typically when one sits in front of the television, the football broadcast is the signal. The weather, a crow landing on the television antenna, interference from a neighbor's microwave—these are sources of noise, the channel conditions for our transmission. But a television repairman has an opposite view. He doesn't want to watch the game, he wants to learn about what is altering the transmission from the station. So he tunes your set to a station broadcasting a test pattern. For the repairman, this test pattern provides channel conditions to read off the signal of how the transmission is being altered. As Sterelny and Griffiths (1999) point out, "The sender/channel distinction is a fact about our interests, not a fact about the physical world." The parity thesis applies this logic to genes and environment. In the parity view, whether it is genome or environment that carries information must be a fact about our interests, not a fact about the world.

The problem with these arguments is that they adopt a few tools from Shannon theory, but neglect its *raison d'être*: the underlying decision problem of how to package information for transport. Before delving deeper into Shannon theory, we will take a brief detour to summarize the semantic sense of information in biology.

Semantic view of information

In addition to the limitations enumerated above, the causal view of biological information fails to highlight the intentional, representational nature of genes and other biological objects (Sterelny and Griffths 1999; Godfrey-Smith 2008; Griffths and Gray 1994; Shea 2007). When biologists talk about genes as informational molecules, this argument goes, it is not because they are correlated with other things (e.g. amino acid sequence or phenotype), but rather because they *represent* other things. This semantic sense of information in which "genes semantically specify their normal products" (Godfrey-Smith 2007) cannot be captured using Shannon theory, which is by design silent on semantic matters.

But what is it that genes are supposed to represent? Much of the conventional language of molecular biology suggests phenotypes as an obvious candidate, and this is the approach that Maynard Smith takes in a target article that triggered much of the recent debate over the information concept in biology (Maynard Smith 2000). At first glance, this view has several things to recommend it. A semantic notion of genetic information captures the directionality discussed above: genes represent phenotypes but phenotypes do not represent genes (Fig. 1). One could also try to argue that the semantic view privileges genes in that we can say that genes have a representational message about phenotype, but environment does not. Finally, it allows for misrepresentation or false representation, whereas causal information does not (Griffths 2001).[1]

Does this mean that the problem is solved? No—Griffths (2001) and Godfrey-Smith (2008) argue that semantic information remains vulnerable to the parity thesis. Moreover, Godfrey-Smith (1999, 2008) and Griffiths (2001) note that the reach of the semantic information concept within genetics is very shallow: legitimate talk of semantic representation can go no further than genes coding for amino acids. Beyond this point, the mapping from genotype forward is context-dependent and hopelessly entangled in a mesh of causal interactions. Thus, these authors conclude, the relation from genes to phenotype cannot be a representational one. Accordingly, it seems as if the semantic view of information has been pushed as far as it will go, and yet we are left without a fully satisfactory account of the information concept in biology. Let us therefore return to Shannon's information theory, but move beyond the causal sense.

A transmission view of information

As we described above, philosophers of biology largely restrict Shannon theory to a descriptive statistics for correlations. This misses the point. At the core of Shannon theory is the study of how far mathematical objects such as sequences and functions can be compressed without losing their identity, and if compressed further, how much their structure will be distorted. From this foundation in the limits of compression emerges a richly practical theory of coding: information theory is a decision theory of how to package information for transport, efficiently. It is a theory about the structure of those sequences that efficiently transmit information. And it is a theory about the fundamental limits with which that information can be

[1] We think that Stegmann handles the misrepresentation issue even more cleanly with a shallow semantic notion of genes as conveying instructional, as opposed to representational, content (Stegmann 2004).

transmitted (Shannon and Weaver 1949; Pierce 1980; Yeung 2002; Cover and Thomas 2006).

This decision-theoretic view of Shannon theory is missing from the discussions of information in biology. While Shannon theory is no panacea for geneticists and evolutionary biologists (Shannon 1956), we will argue that it provides a justification for information language as applied to genes, and that it also resolves the apparent and unappealing symmetries of (1) mutual information between genes and phenotype, and (2) the parity thesis.

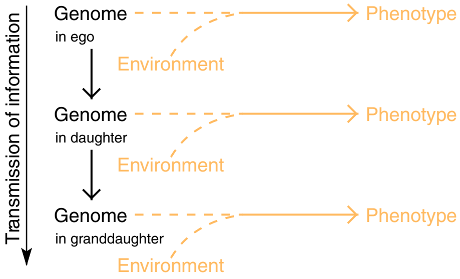
In the original formulation of Shannon theory, information is what an agent packages for transmission through a channel to reduce uncertainty on the part of a receiver. This information is physically instantiated and spatiotemporally bounded. Thus, as Lloyd and Penfield (2003) note, information can be sent either from one *place* to another, or from one *time* to another.[2] Usually when we talk about sending information from one place to another, we posit two separate actors, one of whom sends a message that the other receives; when we talk about sending information from one time to another, we posit a single agent who stores information that she herself can later retrieve. But whether the message goes across space or time, whether there are one or two agents involved, whether we use the language of signal transmission or the language of data storage, mathematically these are exactly the same process. Thus in practice, when you package information and then send it either across the space dimension as a signal or across the time dimension via storage and retrieval, you are *transmitting* information.

Think about what happens when you send a message to your friend by burning a compact disc. Your computer encodes a message onto the digital medium. You send the medium through a channel (e.g. the postal service). Your friend, the receiver, puts the CD into her computer, the computer decodes the message, and she hears the sweet strains of Rick Astley. But it doesn’t matter that you sent the disc through the mail—all of the mathematical operations that underly the information encoding are the same whether you send the CD to a friend or save it for your own later use. To cross space or time, we can encode the same way. Indeed, we use the same error-correcting codes for storage and retrieval on CDs as we do for sending digital images from deep space back to Earth (Cipra 1993).

Taking this view of information and transmission, let us return to the proposed schematic of biological information in Fig. 1. This picture has neither a space dimension or a time dimension; information is not being sent anywhere. Here we simply have a correlation (if one takes a causal view) or a translation (if one takes a semantic view). Thus within the actual Shannon framework, Fig. 1 simply illustrates a decoder. Likewise, notice that the biological processes underlying the schematic in Fig. 1 are not the processes that biologists refer to when they talk about transmission. In biology, *transmission genetics* is the study of inheritance systems, not the study of transcription and translation, and *genetic transmission* is the passing of genes from one generation to another, not the passing of information from genotype to phenotype.

[2] In practice, it takes time to send information from one place to another, but the conventional Shannon framework suppresses this time dimension.

Fig. 2 In biology, genetic transmission occurs vertically (from parent to offspring to grandoffspring). It is upon this axis that the transmission sense of information focuses



In life and in evolution, the transmission of information goes from generation to generation to generation as in Fig. 2. Here is the transmission; we know that genes are transmitted from parent to offspring in order to provide the offspring with information about how to make a living (e.g. metabolize sugars, create cell walls, etc.) in the world. This suggests that we can make sense of a large fraction of the use of information language in biology if we adopt a *transmission view of information*.[3]

Transmission view of information:
An object X conveys information if the function of X is to reduce, by virtue of its sequence properties, uncertainty on the part of an agent who observes X.

As with many aspects of science, the tools and language that we use have a strong influence on the questions that we think to ask—and once we shift to the transmission sense of information, our focus changes. When we view biological information as a semantic relationship, we are drawn to think like developmental biologists, about how information goes from an encoded form in the genotype to its expression in the phenotype. But when we talk about the transmission sense of information, we step out in an orthogonal direction (Fig. 2) and we can now see information as it flows through the process of intergenerational genetic transmission.[4] And once we do that, we can start to think about natural selection, the evolutionary process, and how information gets into the genome in the first place (Szathmary and Maynard Smith 1995).

Symmetry of mutual information

By viewing Shannon information as a result of a decision process instead of as a correlation measure, we can resolve the concern that Godfrey-Smith raises about the bidirectional flow of information in Shannon theory. Godfrey-Smith dismisses the causal sense of information because "information in the Shannon sense 'flows' in both directions, as it involves no more than learning about the state of one variable by attending to another" (Godfrey-Smith 2008). Here Godfrey-Smith is referring to the symmetry of the mutual information measure: $I(X;Y) = I(Y;X)$ (Fig. 1).

[3] A message need not be composed of multiple characters to meet this definition. Even a string of length one is a sequence; thus even a single character conveys information.

[4] Though see Shea (2007) for how a semantic view of information need not be incompatible with a focus on intergenerational processes such as evolution by natural selection.

What is the mutual information actually measuring when we apply this equality in a communication context? An example helps. Peter and Paul are concerned about the state of the world W. Suppose that Peter observes a correlate X of the random variable W. We want him to communicate his observation to Paul, and he does so using a signal Y. The mutual information $I(X;Y)$ tells us how effectively he conveys what he sees, *on average*, given the statistical distribution of possible X values, the properties of the channel across which the signal is sent, etc. Specifically, $I(X;Y) = H(X) - H(X|Y)$ measures how much Paul learns by knowing Y about what Peter saw, X, again on average. Because $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) = I(Y;X)$, Peter knows exactly as much about what Paul learns as Paul learns about what Peter saw. But $I(Y;X)$ is usually irrelevant when we think about the decision problem of communicating. In this context we want Peter to get a message about the world to Paul, and we rarely care how much Peter knows afterwards about what Paul has learned.

This directionality is manifested within Shannon theory by the *data processing inequality* (Fig. 3; Yeung 2002; Cover and Thomas, 2006). This theorem states that the act of processing data, deterministically or stochastically, can never generate additional information. A corollary pertains to communication: along a communication chain, information about the original source can be lost but never gained. In the scenario described above, both the observation step $W \rightarrow X$ and the communication step $X \rightarrow Y$ are steps in a Markov chain. For any Markov chain $W \rightarrow X \rightarrow Y$, the data processing inequality states that $I(W;X) \geq I(W;Y)$. In our example, the data processing inequality reveals that communication between Peter and Paul is not symmetric. Paul may know as much about what Peter sent as Peter knows about what Paul received, but Paul does not in general know as much about what matters—the state of the world—as Peter does. Shannon theory is not symmetric with respect to the direction of communication.

The parity thesis

In the first part of this paper, we described the parity thesis. While there is a good case to be made for parity between genes and environment when we restrict our view to the horizontal development of phenotype from genotype, that parity is shattered when we look along the vertical axis of intergenerational transmission.

Look at Fig. 2, which corresponds to a neo-Darwinian view of evolution. In this model of the biological world, the transmission concept cleanly separates genes

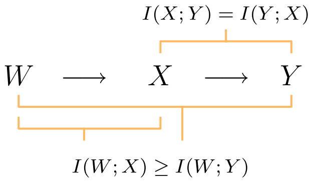


Fig. 3 Despite the symmetry of the mutual information $I(X;Y) = I(Y;X)$, the data processing inequality reveals the directional flow of information in Shannon's scheme.

from environments. The former are transmitted across generations, the latter are not. Moreover, taking a teleofunctional view as Sterelny et al. (1996) do for replicators in general, the hypothesis that genes are *for* transmission across generations is richly supported by the physical structure of the DNA. Genes are made out of DNA, a molecule that is exquisitely fashioned so as to (1) encode lots of sequence information in a small space, (2) be incredibly easy to replicate, (3) be arbitrarily and infinitely extensible in what it can say, and (4) be structurally very stable and inert (Lewontin 1992). In fact, DNA is perhaps the most impressive known substance with respect to (1) and (2). No machine can look at a protein and run off a copy; DNA is exquisitely adapted so that a relatively simple machine, the DNA polymerase, does this at great speed and with high fidelity. Think about how amazing it is that PCR works. It is as if you could throw a hard drive in a water bath with a few enzymes and a few raw materials, run the temperature through a few cycles, and pull out millions of identical hard drives. DNA practically screams, "I am for storage and transmission!"

One might object that Fig. 2 conveys an over-simplified view of the world. This is true. A more sophisticated view of the evolutionary process allows for additional channels of intergenerational transmission and information flow: environments can be constructed and inherited (Odling-Smee et al. 2003). Non-genetic biological structures such as membranes and centrosomes are inherited (Griffths and Knight 1998) and can even be argued to carry some information (Griesemer 2005). Methylation provides an extensive layer of markup on top of nucleic acid sequence. Developmental switches actively transduce environmental information into epigenetically heritable forms (Griffths 2001).

But such an objection misses our point. Our aim with the transmission sense of information is not to single out uniquely the genes as having some special property that we deny to all other biological structures, but rather to identify those components of biological systems that have information storage, transmission, and retrieval as their primary function. Methylation tags are obvious members of this information-bearing class: they carry information across generations in the transmission sense, and this appears to be their primary function. Extrinsic features of the environment such as ambient temperature are obviously not members of this class: they are not transmitted across generations, they carry information only in the causal sense and information transmission is not their role under any reasonable teleofunctional explanation. Biological structures such as membranes and centrosomes may appear as some sort of middle ground, but we note that (1) their primary function is not an informational one, and (2) their bandwidth is extremely restricted compared to that of DNA sequence. Birds' nests (Sterelny et al. 1996) could be seen as an environmental analogue to these intracellular structures, whereas libraries start to push toward genes and methylation tags in their informational capacity. Developmental switches (Griffths 2001) are another interesting case; these transduce environmental information but, in addition to their bandwidth limitations, they appear to have a more limited intergenerational transmission role. Genes may not be unique in their ability to convey information across generations—but at the same time a transmission view makes it clear that not all components of the

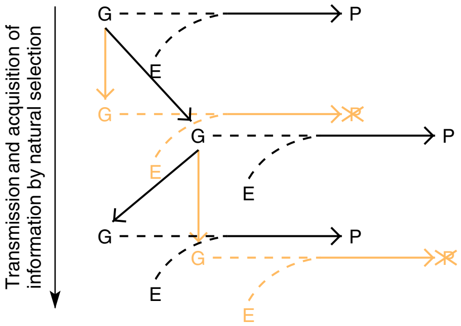

Transmission and acquisition of information by natural selection
G
P
E
G
G
P
E
E
G
P
G
E
E

Fig. 4 Transmission and natural selection. With the parent sending variant messages to each offspring and natural selection acting on the phenotype, information can accumulate in the genome

developmental matrix (Griffths 2001; Griffths and Gray 1994) enjoy parity in their informational capacities.

The parity thesis typically is linked to Shannon's information theory via the claim that "The source/channel distinction is imposed on a natural causal system by the observer." (Griffiths 2001, p.398) What is signal, and what is noise—Sterelny and Griffths (1999) take this to be merely a reflection of our interests. So must we impose our own notions of what makes an appropriate reference frame in order to single out certain components of the developmental matrix as signal and others as noise? If we want to know how the information necessary for life was compiled by natural selection, the answer is no. In this case, we are not the ones who pick the reference frame, *natural selection* is. Because natural selection operates on heritable variation, it acts upon some components of the developmental matrix differently than others.[5] For biologists, therefore, the source-channel distinction is imposed not by the observer but rather by the process of natural selection from which life arose and diversified.

To better understand the role of natural selection, it helps to expand Fig. 2 somewhat. (For simplicity we retain our focus on the genes as transmitted elements, but one could extend this to include other heritable structures). In Fig. 2, we highlight the fact that it is the genes, and not the environment, that are transmitted from generation to generation. In Fig. 4, we highlight the fact that *not all* genes are transmitted to the next generation. It is by the means of variation in the genes and selection on the phenotypes with which they are correlated that information can built up in the genome over time (Felsenstein 1971).

Causal information versus transmitted information

One motivation for replacing causal sense-views of information with semantic-sense views is that the causal sense of information appears to cast too broad of a net. Any physical system with correlations among its components carries causal information—but in their use of the information concept, biologists appear to mean something stronger than the notion of natural meaning that has smoke in the sky carrying information about a fire below (Godfrey-Smith 2008). If we substitute a

[5] Similarly, Shea (2007) uses this fact to derive teleosemantic meaning in his account of biological information.

transmission view of information for a semantic view, will we be driven back to this overly-broad notion of information? Not at all. Like naturalized views of semantics, the transmission notion of information rests upon function: to say that X carries information, we require that the function of X be to reduce uncertainty on the part of a receiver.

The failure to consider function when talking about information sometimes generates confusion among practicing biologists. After all, there are correlations everywhere in biological systems; measuring them is what we do as biologists, and we often talk about these correlations as information. This language is understandable; indeed, these correlations provide *us* with information about biological systems. But this is merely causal-sense information (Godfrey-Smith 1999). As Godfrey-Smith explains, when a systematist uses gene sequences to make inferences about population history, “there is no more to this kind of information than correlation or ‘natural meaning’; the genes are not *trying to tell us* about their past.” (Godfrey-Smith 1999). In other words, these correlations do not convey teleosemantic information. Nor can these correlations be considered information in the transmission sense.

To expand upon this distinction, an example from population genetics is helpful. Voight and colleagues (Voight et al. 2006) developed a method for inferring positive selection at polymorphic loci in the human genome. Their key insight is that, with enough sequence data from sufficiently many members of the population, we can pick out regions of the genome that have unusually long haplotypes of low diversity. Such extended haplotype blocks tend to surround an allele that has recently risen in frequency due to strong selection, because there has not been enough time for recombination to break down the association between the favored allele and the genetic background in which it arose. Using this method, Voight and colleagues find strong evidence for recent selection among Europeans in the lactase gene LCT, which is important for metabolism of lactose beyond early childhood. The favored allele results from a single nucleotide change 14 kb upstream of the lactase gene on the nearly 1 Mb haplotype. Positive selection has presumably occurred because the ability to process lactase throughout life became advantageous with the invention of animal agriculture approximately 10,000 years ago.

What does this have to do with information? There is information about the history of selection in the population-level correlations. Voight et al. found an extended haplotype length surrounding the LCT+ allele relative to that around the LCT- allele. But notice that we have to observe the genotypes of multiple individuals in order to determine that one allele at the LCT locus is surrounded by longer haplotype blocks than is the other. Once we have made observations of multiple genomes, we as external observers can conclude something about the history of selection on the population. But this information is not available at the level of a single individual. A single individual cannot look at its own genome and notice a longer (or shorter) haplotype block around any given focal locus—these haplotype blocks are defined with respect to the genotypes of others in the population. A single individual can only look at its own genome and see a sequence of base pairs. This sequence of base pairs is what is transmitted; it is what has the function of reducing uncertainty on the part of the agent who observes it.[6] These individual gene sequences are the entities that have an informational function in biology (though bioinformaticians have not always recognized this distinction (Adami 2002)). The population-level statistics that geneticists use to infer history are informative, but they are not information in the transmission sense. They are merely the smoke that is cast off by the fire of natural selection.

Coding without appeal to semantics

The transmission sense of information allows us to separate claims about how information is transmitted from claims about what information means.[7] Indeed, we can study how information is transmitted without having any knowledge of the "codebook" for how to interpret the message, or even what the information represents.

In many biological studies, we are in exactly this position. Again an example—this time drawn from neurobiology—is helpful. In a study of the fly visual system, de Ruyter van Steveninck et al. (1997) presented flies with a moving grating as a visual stimulus, and made single-cell recordings of the spike train from the H1 visual neuron. This neuron is sensitive to movement, but it is not known how movement information is encoded into the spike train, nor even what aspects of movement are being represented. Nonetheless, de Ruyter van Steveninck and colleagues were able to determine how much information this neuron is able to encode. The investigators exposed a fly to the stimulus, and measured the (average) entropy of the spike train. This is the so-called total entropy for the neuron's output. They then looked at what happens if you play the same stimulus back repeatedly: how much does the resultant spike train vary from previous trials? This is the so-called noise entropy. The information that the spike train carries about the stimulus, i.e., the mutual information between spike train and stimulus, is simply the difference in these two quantities. Using this approach, the authors were able to show that this single insect neuron conveys approximately 80 bits of information per second about a dynamic stimulus. Thus an individual visual neuron achieves a bit-rate that is roughly 7 times the bit rate of a skilled touch typist![8] More importantly, the researchers were able to compare the response of this neuron to static stimuli

[6] Although causal-sense information is transmitted from the population at time t to the population at time t+1 in the population frequencies of haplotypes, this is not transmission-sense information because the function of these population-level haplotype assemblages is not to reduce uncertainty on the part of future populations.

[7] The source-channel separation theorem (Cover and Thomas 2006, Chapter 7, p.218) proves that in any physical communication system for error-free transmission over a noisy channel, one can entirely decouple the process of tuning the code to the nature of the specific channel from not only the semantic reference of the signal but, indeed, from all statistics of the message source. This follows because the theorem states that one can achieve channel capacity with separate source and channel coders—and in this setup, the source coder can always be configured so as to return output that maximizes the entropy given the symbol set.

[8] Using Shannon's 1950 upper bound on bits per letter and his estimate of letters per word in the English language (Shannon 1950), we can estimate the bit rate of a touch typist as $\frac{120\,\text{words}}{\text{minute}}\frac{1\,\text{minute}}{60\,\text{seconds}}\frac{4.5\,\text{letters}}{\text{word}}\frac{1.3\,\text{bits}}{\text{letter}} = 11.7$ bits/second.

with the response of the neuron to natural patterns of motion. They found that, for natural patterns, the neuron is able to attain the high bandwidth that it does by "establishing precise temporal relations between individual action potentials and events in the sensory stimulus." By doing so, the neuron's response to the dynamic stimulus greatly surpasses the bit rate that could be obtained if the stimulus were encoded by a simple matching of spike rate to stimulus intensity. Subsequent investigators have used related methods to show that evolved sensory systems are tuned to natural stimuli, to study the properties of neural adaptation and history dependence, and to examine temporal sensitivity—all without knowing the way in which the signals that they study are actually encoded.

de Ruyter van Steveninck et al. (1997) sum up the power of being able to study information without appeal to semantics: "This characterization of ... information transmission is independent of any assumptions about [or knowledge of!] which features of the stimulus are being encoded or about which features of the spike train are most important in the code".

This brings us back to Shannon theory. When information theorists think about coding, they are not thinking about semantic properties. All of the semantic properties are stuffed into the codebook, the interface between source structure and channel structure, which to information theorists is as interesting as a phonebook is to sociologists. When an information theorist says "Tell me how data stream A codes for message set B," she is not asking you to read her the codebook. She is asking you to tell her about compression, channel capacity, distortion structure, redundancy, and so forth. That is, she wants to know how the structure of the code reflects the statistical properties of the data source and the channel with respect to the decision problem of effectively packaging information for transport.

With these things in focus, we can now look at the concept of arbitrariness, what it means, and why this concept is critically important in biological coding.

Information theory and arbitrariness

In arguing that DNA is an informational molecule, Maynard Smith (Maynard Smith 2000) appeals to Jacques Monod's concept of *gratuité* (Monod 1971), and a number of additional authors have further explored this thread (Godfrey-Smith 2000a, b; Stegmann 2005). For Monod, *gratuité* was an important component of the logical structure of his theory of gene regulation. *Gratuité* is the notion that, in principle, regulatory proteins can cause any inducer or repressor to influence the expression of any region of DNA. There need be no direct chemical relation between the structure of an inducer and the nucleic acid sequence on which it operates. Maynard Smith observes that we can see something like *gratuité* in the structure of the genetic code as well: there is an *arbitrary* association between codons and the amino acids that they specify.

Yet as they grapple with this idea of an arbitrary code, these authors confront the fact that the genetic code is not a random assignment of codons to amino acids, but rather a one-in-a-million evolved schema for associating these molecules: the genetic code is structured so as to smooth the mutational landscape (Sonneborn

1965) and ensure that common translational errors generate amino acid replacements between chemically similar amino acids (Freeland and Hurst 1998; Woese 1965; Haig and Hurst 1991). So what can these authors mean when they say that the code is arbitrary? Maynard Smith, Godfrey-Smith, and others are not suggesting that the structure of the code is random or contingent on random historical processes, as in Crick’s frozen accident hypothesis. Rather, they are making a semiotic claim. Arbitrariness refers to the fact that “[m]olecular symbols in biology are symbolic,” as opposed to indexical or iconic (Maynard Smith 2000). In the case of the genetic code, this means that the association between a codon and its corresponding amino acid is not driven by the immediate steric interactions between the codon and the amino acid, but instead is mediated by an extensive tRNA structure that in principle could have coupled this codon to any other amino acid instead.

This fact is enormously important to the function of the biological code—not as a matter of the semiotic classifications that fascinated Charles Pierce, but rather to solve the sort of decision problem that motivated Claude Shannon. From the symbolic relation between code and product, there arise the degrees of freedom that a communication engineer requires to tune the code to the statistical properties of source and channel. To see how this works we will visit an example from the early history of telecommunications.

For over one hundred years, Morse code was the standard protocol for telegraph and radio communication. The code transcribes the English alphabet into codewords composed of short pulses called dots and long pulses called dashes. For example, the letter E is represented by a single dot “.”, the letter T is represented by a single dash “-”, the letter Q by the quartet “- -.-”, and the letter J by “.- - -”. At first glance, the mapping between letters and Morse codewords appears to be arbitrary. They are certainly symbolic rather than iconic or indexical. But there is an important pattern to the way that letters are assigned codes in Morse code.

Instead of assigning codewords sequentially (“.” to A, “-” to B, “..” to C) Samuel Morse exploited the degrees of freedom available for codeword assignments to make an efficient code for fast transmission of English sentences. He could not assign short code words to every single letter—there simply are not enough short code words to go around. Instead, by assigning the shortest code words to the most commonly used letters, Morse created a code in which transmissions would on average be shorter than if he had used sequential codeword assignments.

Morse could not have done this with pictograms. The leeway to associate any message with any code word provides the communications engineer with the degrees of freedom that he needs to tune the semantic and statistical properties of the source messages to the transmission cost and error properties of the channel. In the absence of a full picture of the engineer’s decision problem, the code might look arbitrary. But a well-chosen code is not arbitrary at all; it solves a decision problem for packaging.

Morse exploited the available degrees of freedom in his choice of codes; apparently, natural selection has done the same in evolving a one-in-a-million genetic code. Thinking about these degrees of freedom—along with an important thought experiment—helped us to understand the role of coding in the transmission sense of information. Godfrey-Smith (2000a) imagined a hypothetical world composed of “protein genes” as a way to explore the importance of the coding concept in biology. Since the idea of coding presumably refers to the fact that DNA provides an arbitrary combinatorial representation of amino acid sequence, Godfrey-Smith considers the nature of a world in which the hereditary material neither arbitrary nor representational. He imagine a world of “protein-genes” in which there is no translation, but instead a copying system in which amino acid sequences, assisted by coupling molecules, replicate using previous amino acid sequences as templates. In that world, Godfrey-Smith argues, there is nothing that corresponds to coding, and yet stepping away from the microscope, biology functions much as before. In Godfrey-Smith’s protein-sample world, there is no code, no compression, no redundancy, and information theory can merely be applied as a descriptive statistics for correlations. One could even go so far as to argue that in a protein-genes world, physical structure and not information is inherited across generations. Thus, Godfrey-Smith concludes that “Removing genetic coding from the world need not change much else, and this gives support to my claim that we should only think of coding as part of an explanation of how cells achieve the specific task of putting amino acids in the right order,” rather than something fundamental to the logical structure of biology.

But there is a critical difference between a world with a proper genetic code and a world based upon protein-genes: only the former allows an arbitrary combinatorical mapping between templates and products. From an information-theoretic perspective, this is absolutely critical. The degrees of freedom to construct an arbitrary mapping of this sort turn the problem of code evolution from one of passing physical samples to the next generation into a decision-theoretic problem of how to package information for transport. In the protein-genes world, the fidelity of transmission is at the mercy of the biochemical technology for copying. There are no degrees of freedom for structuring redundancy, minimizing distortion, or conducting the other optimization activities of a communications engineer. In a DNA-based code, the chemically arbitrary assignments of nucleotide triplets to amino acids via tRNAs offer the degrees of freedom to do all of the above. While the precise dynamics of code evolution remain unknown, it appears that natural selection has put these degrees of freedom to good use.

For example, during the early evolution of the genetic code when the translation mechanism was highly inaccurate, there was not only selection to minimize the effect of misreads (Woese 1965), but also selection to minimize the effect of frame-shift errors. Frame-shift errors waste resources by generating potentially toxic nonsense polypeptides. There is therefore a fitness advantage to codes that quickly terminate after a frame-shift by reaching a stop codon (Seligmann and Pollock 2004). Itzkovitz and Alon (2007) have shown that of the 1152 alternative codes that are equivalent to the real code with respect to translational misreads (the real code with independent permutations of the nucleotides and wobble-base constraint), only 8 codes encounter a stop codon earlier after a frame-shift than does the real code. Because in the real code, stop codons overlap with common codons, the length of nonsense peptides are on average 8 codons shorter than in alternative codes.

Even with the genetic code fixed in its present form, natural selection can still make use of the degrees of freedom that are available when selecting which of several synonymous codons to use. For example, there is strong positive correlation both between synonymous codon bias and gene expression level and between tRNA abundance and codon usage in *S. cerevisiae* and *E. coli* (Bennetzen and Hall 1982; Ikemura 1981; Sharp and Li 1987). While the underlying mechanisms for the correlations are not fully understood, simple models of mutation and selection can explain much of the observed variance (Knight et al. 2001; Bulmer 1987).

As another intriguing example, Mitarai et al. (2008) have found that the choice of synonymous codons along the length of a gene operates to prevent "traffic congestion" by ribosomes moving down the mRNA, potentially increasing expression rates and minimizing incomplete translation of proteins. This works as follows. Some codons are translated faster than others. For instance, Sørensen and Pedersen (1991) have shown that the difference in translation rates between the two synonymous glutamate codons GAA and GAG is threefold. Just as the relative position of fast and slow regions of highway influence the rate at which traffic can travel and the amount of traffic congestion that occurs, the relative positions of rapidly and slowly translated codons will influence the rate at which ribosomes can move along an mRNA and the amount of congestion that they experience. If rapidly translated codons were to occur early in a gene and slowly translated ones were to occur late, numerous ribosomes could load on the mRNA, race through the early part of the gene, but then back up against one another when the slower codons were reached toward the gene's end. At best this would cause congestion and slow-down; at worst, incomplete translation as the blocked ribosomes disengage from the mRNA. If instead slowly translated codons occur early and rapidly translated ones occur late, more ribosomes can pass along the mRNA in a given time window, and less congestion arises. Looking at highly-expressed genes in *E. coli*, Mitarai et al. (2008) have found evidence of exactly this pattern; there appears to have been selection for using slower synonymous codons in the beginning of these highly expressed genes.

The difference between a protein-genes world and a DNA-based world became clear by taking the perspective of a communications engineer. Throughout the paper, this has been our approach. Correlations, symmetry of mutual information, the parity thesis, arbitrariness, coding—all of these come into focus from a communications viewpoint. We see what makes the genetic code a code, and we get a new perspective on the information language that is part of the everyday working vocabulary of researchers in genetics and evolutionary biology. The transmission sense of information justifies such language as more than shallow metaphor.

References

Adami C (2002) What is complexity? BioEssays 24:1085–1094

Bennetzen J, Hall B (1982) Codon selection in yeast. J Biol Chem 257:3026–3031

Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730

Cipra BA (1993) The ubiquitous reed-solomon codes. SIAM News 26, No. 1

Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, New York

Crick FHC (1970) The central dogma of molecular biology. Nature 227:561–563

Crick FHC, Griffith JS, Orgel LE (1957) Codes without commas. Proc Natl Acad Sci USA 43:416–421

de Ruyter van Steveninck RR, Lewen GD, Strong SP, Koberle R, Bialek W (1997) Reproducibility and variability in neural spike trains. Science 275:1805–1808

Dretske FI (1983) Knowledge and the flow of information. MIT, Cambridge

Felsenstein J (1971) On the biological significance of the cost of gene substitution. Am Nat 105:1–11

Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Godfrey-Smith P (1999) Genes and codes: lessons from the philosophy of mind. In: Hardcastle V (ed) Where biology meets psychology: philosophical essays. MIT, Cambridge, pp 305–331

Godfrey-Smith P (2000a) On the theoretical role of “genetic coding”. Philos Sci 67:26–44

Godfrey-Smith P (2000b) Information, arbitrariness, and selection: comments on maynard smith. Philos Sci 67:202–207. ISSN 00318248

Godfrey-Smith P (2007) Biological information. In: Zalta EN (ed) Stanford encyclopedia of philosophy. Center for the Study of Language and Information, Stanford University

Godfrey-Smith P (2008) Information in biology. In: Hull DL, Ruse M (eds) The philosophy of biology, chap 6. Cambridge University Press, Cambridge, pp 103–119

Grice HP (1957) Meaning. Philos Rev 66:377–388

Griesemer JR (2005) The informational gene and the substantial body: on the generalization of evolutionary theory by abstraction. In: Jones MR, Cartwright N (eds) Idealization XII: correcting the model, idealization and abstraction in the sciences. Rodopi, Amsterdam, pp 59–115

Griffiths PE (2001) Genetic information: a metaphor in search of a theory. Philos Sci 68:394–412

Griffiths PE, Gray RD (1994) Developmental systems and evolutionary explanation. J Philos XCI:277–304

Griffiths PE, Knight RD (1998) What is the developmentalist challenge? Philos Sci 65:276–288

Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic code. J Mol Evol 37:412–417

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389–409

Itzkovitz S, Alon U (2007) The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res 17:405–412

Kimura M (1961) Natural selection as the process of accumulation of genetic information in adaptive evolution. Genet Res 2:127–140

Knight R, Freeland S, Landweber L (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2:r0010.1–r0010.13

Lewontin RC (1992) The dream of the human genome. N Y Rev Books 39

Lloyd S, Penfield P (2003) Information and entropy: mit open courseware

Maynard Smith J (2000) The concept of information in biology. Philos Sci 67:177–194. ISSN 00318248

Mitarai N, Sneppen K, Pedersen S (2008) Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. J Mol Biol 382:236–245

Monod J (1971) Chance and necessity. Collins, London

Odling-Smee FJ, Leland KN, Feldman MW (2003) Niche construction. Princeton University Press, New Jersey

Pierce JR (1980) An introduction to information theory. Dover, New York

Seligmann H, Pollock D (2004) The ambush hypothesis: hidden stop codons prevent off-frame gene reading. DNA Cell Biol 23:701–705

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423

Shannon CE (1950) Prediction and entropy of printed English. Bell Syst Tech J 30:50–64

Shannon CE (1956) The bandwagon. IEEE Trans Inf Theory 2, No. 3

Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Sharp P, Li W (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Shea N (2007) Representation in the genome and in other inheritance systems. Biol Philos 22:313

Sonneborn TM (1965) Degeneracy of the genetic code: extent, nature and genetic implications. In: Bryson V, Vogel JH (eds) Evolving genes and proteins. Academic Press, New York, pp 377–397

Sørensen M, Pedersen S (1991) Absolute in vivo translation rates of individual codons in *Escherichia coli*. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. J Mol Biol 222:265–280

Stegmann UE (2004) The arbitrariness of the genetic code. Biol Philos 19:205–222

Stegmann UE (2005) Genetic information as instructional content. Philos Sci 72:425–443

Sterelny K (2001) The "genetic program" program: a commentary on Maynard Smith on information in biology. Philos Sci 67:195–201

Sterelny K, Griffiths PE (1999) Sex and death: an introduction to philosophy of biology. University of Chicago Press, Chicago

Sterelny K, Smith KC, Dickison M (1996) The extended replicator. Biol Philos 11:377–403

Szathmary E, Maynard Smith J (1995) The major evolutionary transitions. Nature 374:227–232

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:446–458

Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546–1552

Yeung RW (2002) A first course in information theory. Springer, New York