CHAPTER THIRTEEN

Genes Encode Information for Phenotypic Traits

Sahotra Sarkar

13.1 Introduction

According to the *Oxford English Dictionary*, the term "information" (though spelt "informacion") was first introduced by Chaucer in 1386 to describe an item of training or instruction. In 1387, it was used to describe the act of "informing." As a description of knowledge communicated by some item, it goes back to 1450. However, attempts to quantify the amount of information contained in some item only date back to R. A. Fisher in 1925. In a seminal paper on the theory of statistical estimation, Fisher argued that "the intrinsic accuracy of an error curve may . . . be conceived as the amount of information in a single observation belonging to such a distribution" (1925, p. 709). The role of information was to allow discrimination between consistent estimators of some parameter; the amount of "information" gained from a single observation is a measure of the efficiency of an estimator. Suppose that the parameter to be estimated is the mean height of a human population; potential estimators can be other statistical "averages" such as the median and the mode. Fisher's theory of information became part of the standard theory of statistical estimation, but is otherwise disconnected from scientific uses of "information." The fact that the first successful quantitative theory of "information" is irrelevant in scientific contexts outside its own narrow domain underscores an important feature of the story that will be told here: "information" is used in a bewildering variety of ways in the sciences, some of which are at odds with each other. Consequently, any account of the role of informational thinking in a science must pay careful attention to exactly what sense of "information" is intended in that context.

Shortly after Fisher, and independently of him, in 1928 R. V. L. Hartley provided a quantitative analysis of the amount of information that can be transmitted over a system such as a telegraph. During a decade in which telecommunication came to be at the forefront of technological innovation, the theoretical framework it used proved

to be influential. (Garson (2002) shows how this technological context led to the introduction of informational concepts in neurobiology during roughly the same period.) Hartley recognized that, "as commonly used, information is a very elastic term," and proceeded "to set up for it a more specific meaning" (1928, p. 356). Relying essentially on a linear symbolic system of information transmission (for instance, by a natural language), Hartley argued that, for a given message, "inasmuch as the precision of the information depends upon what other symbol sequences might have been chosen it would seem reasonable to hope to find in the number of these sequences the desired quantitative measure of information" (p. 536). Suppose that a telegraphic message is n symbols long with the symbols drawn from an alphabet of size s. Through an ingenious argument, similar to the one used by Shannon (see below), Hartley showed that the appropriate measure for the number of these sequences is $n \log s$. He identified this measure with the amount of information contained in the message. (Even earlier, H. Nyquist (1924) had recognized that the logarithm function is the appropriate mathematical function to be used in this context. Nyquist's "intelligence" corresponds very closely with the modern use of "information.")

Using the same framework as Hartley, in 1948, C. E. Shannon developed an elaborate and elegant mathematical theory of communication that came to be called "information theory" and constitutes one of the more important developments of applied mathematics in the twentieth century. The theory of communication will be briefly analyzed in section 13.2 below, with an emphasis on its relevance to genetics. The assessment of relevance will be negative. When, for instance, it is said that the hemoglobin-S gene contains information for the sickle cell trait, communication-theoretic information cannot capture such usage. (Throughout this chapter, "gene" will be used to refer to a segment of DNA with some known function.) To take another example, the fact that the information contained in a certain gene may result in polydactyly (having an extra finger) in humans also cannot be accommodated by communication-theoretic information. The main problem is that, at best, communication-theoretic information provides a measure of the amount of information in a message but does not provide an account of the content of a message, its specificity, what makes it *that* message. The theory of communication never had any such pretension. As Shannon bluntly put it at the beginning of his paper: "These semantic aspects of communication are irrelevant to the engineering problem" (1948, p. 379).

Capturing *specificity* is critical to genetic information. Specificity was one of the major themes of twentieth-century biology. During the first three decades of that century, it became clear that the molecular interactions that occurred within living organisms were highly "specific" in the sense that particular molecules interacted with exactly one, or at most a very few, reagents. Enzymes acted specifically on their substrates. Mammals produced antibodies that were highly specific to antigens. In genetics, the ultimate exemplar of specificity was the "one gene–one enzyme" hypothesis of the 1940s, which served as one of the most important theoretical principles of early molecular biology. By the end of the 1930s, a highly successful theory of specificity, one that remains central to molecular biology today, had emerged. Due primarily to L. Pauling (see, e.g., Pauling, 1940), though with many antecedents, this theory claimed: (i) that the behavior of biological macromolecules was determined by their

shape or "conformation"; and (ii) what mediated biological interactions was a precise "lock-and-key" fit between shapes of molecules. Thus the substrate of an enzyme had to fit into its active site. Antibodies recognized the shape of their antigens. In the 1940s, when the three-dimensional structure of not even a single protein had been experimentally determined, the conformational theory of specificity was still speculative. The demonstration of its approximate truth in the late 1950s and 1960s was one of early molecular biology's most significant triumphs.

Starting in the mid-1950s, assumptions about information provided an alternative to the conformational theory of specificity, at least in the relation between DNA and proteins (Lederberg, 1956). This relation is the most important one, because proteins are the principal biological interactors at the molecular level: enzymes, antibodies, molecules such as hemoglobin, molecular channel components, cell membrane receptors, and many (though not most) of the structural molecules of organisms are proteins. Information, as F. H. C. Crick defined it in 1958, was the "*precise* determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein" (1958, p. 153; emphasis in the original). Genetic information lay in the DNA sequence. The relationship between that sequence and the sequence of amino acid residues in a protein was mediated by the genetic "code," an idea that, though originally broached by E. Schrödinger in 1943, also dates from the 1950s. The code explained the specificity one gene–one enzyme relationship elegantly: different DNA sequences encoded different proteins, as can be determined by looking up the genetic code table (Godfrey-Smith explains the relevant biology in chapter 14 of this volume). Whatever the appropriate explication of information for genetics is, it has to come to terms with specificity and the existence of this coding relationship. Communication-theoretic information neither can, nor was intended to, serve that purpose.

Surprisingly, a comprehensive account of a theory of information appropriate for genetics does not exist. In the 1950s there were occasional attempts by philosophers – for instance, by R. Carnap and Y. Bar-Hillel (1952) – to explicate a concept of "semantic" information distinct from communication-theoretic information. However, these attempts were almost always designed to capture the semantic content of linguistic structures and are of no help in the analysis of genetic information. Starting in the mid-1990s, there has been considerable skepticism, at least among philosophers, about the role of "information" in genetics. For some, genetic information is no more than a metaphor masquerading as a theoretical concept (Sarkar, 1996a,b; Griffiths, 2001). According to these criticisms, even the most charitable attitude toward the use of "information" in genetics can only provide a defense of its use in the 1960s, in the context of prokaryotic genetics (i.e., the genetics of organisms without compartmentalized nuclei in their cells). Once the "unexpected complexity of eukaryotic genetics" (Watson, Tooze, and Kurtz, 1983, ch. 7) – that is, the genetics of organisms with compartmentalized nuclei in their cells – has to be accommodated, the loose use of "information" inherited from prokaryotic genetics is at least misleading (Sarkar, 1996a). Either informational talk should be abandoned altogether or an attempt must be made to provide a formal explication of "information" that shows that it can be used consistently in this context and, moreover, is useful.

Section 13.3 gives a sketch of one such attempted explication. A category of "semiotic" information is introduced to explicate such notions as coding. Semiotic infor-

mation incorporates *specificity* and depends on the possibility of *arbitrary* choices in the assignment of symbols to what they symbolize as, for instance, exemplified in the genetic code. Semiotic information is not a semantic concept. There is no reason to suppose that any concept of *biological* information must be "semantic" in the sense that philosophers use that term. Biological interactions, at this level, are about the rate and accuracy of macromolecular interactions. They are not about meaning, intentionality, and the like; any demand that such notions be explicated in an account of biological information is no more than a signifier for a philosophical agenda inherited from manifestly nonbiological contexts, in particular from the philosophy of language and mind. It only raises spurious problems for the philosophy of biology.

Section 13.3 also applies this framework to genetics at the macromolecular level of DNA and proteins. It concludes that there is a sense in which it is appropriate and instructive to use an informational framework for genetics at this level. However, proteins are often far removed from the traits that are usually studied in organismic biology; for instance, the shape, size, and behavior of organisms. Section 13.4 explores the extent to which informational accounts carry over to the level of such traits. Much depends on how "trait" is construed, and there is considerable leeway about its definition. Given a relatively inclusive construal of "trait," section 13.4 concludes that, to the extent that a molecular etiology can at all be attributed to a trait, a carefully circumscribed informational account remains appropriate.

Finally, section 13.5 cautions against any overly ambitious interpretation of the claims defended earlier in this chapter. They do not support even a mild form of genetic reductionism (that genes alone provide the etiology of traits), let alone determinism. They do not support the view that DNA alone must be the repository of biological information. Perhaps, most importantly, they do not support the view that the etiology of traits can be fully understood in informational terms from a predominantly genetic basis.

13.2 The Theory of Communication

Shannon conceived of a communication system as consisting of six components:

1 An information source that produces a raw "message" to be transmitted.
2 A transmitter that transforms or "encodes" this message into a form appropriate for transmission through the channel.
3 The channel through which the encoded message or "signal" is transmitted to the receiver.
4 The receiver that translates or "decodes" the received signal back into what is supposed to be the original message.
5 The destination or intended recipient of the message.
6 Sources of noise that act on the channel and potentially distort the signal or encoded message (obviously this is an optional and undesirable component, but one that is unavoidable in practice).

The most important aspect of this characterization is that it is abstracted away from any particular protocol for coding as well as any specific medium of transmission.

From the point of view of the theory of communication, information is conceived of as the choice of one message from a set of possible messages with a definite probability associated with the choice: the lower this probability, the higher is the information associated with the choice, because a greater uncertainty is removed by that choice. The central problems that the theory of communication sets out to solve include the efficiency (or relative accuracy) with which information can be transmitted through a channel in the presence of noise, and how the rate and efficiency of transmission are related to the rate at which messages can be encoded at the transmitter and to the capacity of the channel.

To solve these problems requires a quantitative measure of information. Suppose that a message consists of a sequence of symbols chosen from a basic symbol set (often called an "alphabet"). For instance, it can be a DNA sequence, with the basic symbol set being $\{A, C, G, T\}$. In a sequence of length n, let p_i be the probability of occurrence of the ith symbol in that sequence. Then the information content of the message is $\mathrm{H} = -\Sigma_{i=1}^{n} p_i \log p_i$. (In what follows, this formula will be called the "Shannon measure of information.") Consider the sequence "*ACCTCGATTC*." Then, at the first position, $\mathrm{H}_1 = p_1 \log p_1$, where $p_1 = \frac{2}{10} = 0.2$ because "*A*" occurs twice in the sequence of ten symbols. $\mathrm{H} = \Sigma_{i=1}^{10} \mathrm{H}_i$, computed in this way, is the information content of the entire sequence.

Shannon justified this choice for the measure of information by showing that $-K\Sigma_{i=1}^{n} p_i \log p_i$, where K is a constant, is the only function that satisfies the following three reasonable conditions: (i) the information function is continuous in all the p_i; (ii) if all the p_i are identical and equal to $1/n$, then the function is a monotonically increasing function of n; and (iii) if the choice involved in producing the message can be decomposed into several successive choices, then the information associated with the full choice is a weighted sum of each of the successive choices, with the weights equal to the probability of each of them. K is fixed by a choice of units. If the logarithm used is of base 2, then K is equal to 1.

Formally, the Shannon measure of information is identical to the formula for entropy in statistical physics. This is not surprising since entropy in statistical physics is a measure of disorder in a system and Shannon's information measures the amount of uncertainty that is removed. Over the years, some have held that this identity reveals some deep feature of the universe; among them are those who hold that physical entropy provides a handle on biological evolution (e.g., Brooks and Wiley, 1988). Most practitioners of information theory have wisely eschewed such grandiose ambitions (e.g., Pierce, 1962), noting that the identity may be no greater significance than, for instance, the fact that the same bell curve (the normal distribution) captures the frequency of measurements in a wide variety of circumstances, from the distribution of molecular velocities in a gas to the distribution of heights in a human population. However, the entropic features of the Shannon measure have been effectively used by T. D. Schneider (see, e.g., Schneider, 1999) and others to identify functional regions of DNA sequences: basically, because of natural selection, these regions vary less than others and relatively invariant sequences with high probabilities and, therefore, low information content is expected to be found in these regions.

Shannon's work came, as did the advent of molecular biology, at the dawn of the computer era when few concepts were as fashionable as that of information.

The 1950s saw many attempts to apply information theory to proteins and DNA; these were an unmitigated failure (Sarkar, 1996a). Within evolutionary genetics, M. Kimura (1961) produced one intriguing result in 1961: the Shannon measure of information can be used to calculate the amount of "information" that is accumulated in the genomes of organisms through natural selection. A systematic analysis of Kimura's calculation was given by G. C. Williams in 1966. Williams first defined the gene as "that which segregates and recombines with appreciable frequency" (1966, p. 24) and then assumed, offering no argument, that this definition is equivalent to the gene being "any hereditary information for which there is a favorable or unfavorable selection bias equal to several or many times its rate of endogenous change [mutation]" (p. 25). From this, Williams concluded that the gene is a "cybernetic abstraction" (p. 33). Williams' book lent credence to the view that the gene, already viewed informationally by molecular biologists, can also be so viewed in evolutionary contexts.

Nevertheless, interpreting genetics using communication-theoretic information presents insurmountable difficulties even though the most popular objection raised against it is faulty:

1 The popular objection just mentioned distinguishes between "semantic" and "causal" information (Sterelny and Griffiths, 1999; Maynard Smith, 2000; Sterelny, 2000; Griffiths, 2001; see also Godfrey-Smith's chapter 14 in this volume). The former is supposed to require recourse to concepts such as intentionality. As mentioned in the last section, and as will be further demonstrated in the deflationary account of information given in the next section, such resources are not necessary for biological information. (It is also not clear why a notion of information based on intentionality should be regarded as "semantic" if that term is supposed to mean what logicians usually take it to mean.) Flow of the latter kind of information is supposed to lead to covariance (in the sense of statistical correlation) between a transmitter and receiver. Communication-theoretic information is supposed to be one type of such "causal" information. (It is also misleading to call this type of interaction "causal," since it presumes nothing more than mere correlation.) With this distinction in mind, the objection to the use of "causal" information in genetics goes as follows: in genetics, the transmitter is supposed to consist of the genes (or the cellular machinery directly interpreting the DNA – this choice does not make a difference), the receiver is the trait, and the cytoplasmic and other environmental factors that mediate the interactions between the genes and the trait constitute the channel. Now, if environmental conditions are held constant, the facts of genetics are such that there will indeed be a correlation between genes and traits. However, one can just as well hold the genetic factors constant, treat the environment as the transmitter, and find a correlation between environmental factors and traits. The kernel of truth in this argument is that, depending on what is held constant, there will be a correlation between genetic or environmental factors and traits. In fact, in most circumstances, even if neither is held entirely constant, there will be correlations between both genetic and environmental factors and traits. All that follows is that, if correlation suffices as a condition for information transfer, both genes and environments carry information for traits. Thus, if genes are to be informationally privileged, more than correlation

is required. However, mathematical communication or information theory is irrelevant to this argument. The trappings of Shannon's model of a communication system are extraneously added to a relatively straightforward point about genetic and environmental correlations, and do no cognitive work. This argument does not establish any argument against the use of communication-theoretic information to explicate genetic information.

2 The main reason why communication-theoretic information does not help in explicating genetic information is that it does not address the critical point that genetic information explains biological specificity. This is most clearly seen by looking carefully at what the various components Shannon's model of a communication system do. Three of these are relevant here: the transmitter, the channel, and the receiver. There is a symmetry between the transmitter and the receiver; the former encodes a message to be sent, the latter decodes a received message. Specific relations are only at play in Shannon's model during encoding and decoding. In most cases, though not all, encoding and decoding consists of something like translation using a dictionary. The transmitter obtains from the source a message in some syntactic form. The encoding process uses this as an input and produces as output some sequence of entities (another syntactic form) that is physically amenable for propagation through the channel. This process may consist of using the value of some syntactic object in the input (for instance, the intensity of light of a particular frequency) to select an appropriate output; however, most often, it consists of using a symbolic look-up table. In either case, there is a specific relationship between input and output. (In information theory, this is the well-studied "coding" problem.) A similar story holds for decoding. The critical point is that the Shannon measure of information plays no role in this process. That measure only applies to what happens along the channel. If a highly improbable syntactic object (the signal) is the input and, because of noise, a less improbable entity is the output, information has been lost. Specificity has nothing to do with this.

3 Even leaving aside specificity, note that high communication-theoretic information content is the mark of a message of low probability. Yet, due to natural selection, DNA sequences that most significantly affect the functioning of an organism, and are thus most functionally informative, become increasingly frequent in a population. Thus, communication-theoretic information is negatively correlated with functional information. This conclusion raises a peculiar possibility. Communication-theoretic information may be a guide to functional information if, as a measure of functional information, its inverse is used. Now if functional information is taken to be a measure of genetic information (which is reasonable since, during evolution, functional information is primarily transmitted through DNA), communication-theoretic information provides access to genetic information through this inversion. Nevertheless, this attempt to rescue communication-theoretic information for genetics fails because those genes that are not selected for are no longer carriers of information. These include not only genes that may be selected against but nevertheless persist in a population (for instance, because of heterozygote advantage or tight linkage – that is, being very close on a chromosome to a gene that is selected for) but also the many genes that are simply selectively neutral. Understanding the role of information in genetics will require a different approach altogether.

13.3 Semiotic Information

Informational talk pervades contemporary genetics with the gene as the locus of information. DNA is supposed to carry information for proteins, possibly for traits. The so-called Central Dogma of Molecular Biology states: "once 'information' has passed into protein *it cannot get out again*. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible" (Crick, 1958; emphasis in the original). J. Maynard Smith (2000) has claimed that DNA is the sole purveyor of biological information in organisms. Routinely, talk of information is intertwined with linguistic metaphors, from both natural and artificial languages: there is a genetic *code*, because a triplet of DNA (or RNA) nucleotides codes for each amino acid residue in proteins (polypeptide chains); there are alternative *reading frames* – DNA is *transcribed* into RNA, RNA is *translated* into protein, RNA is *edited*, and so on. The use of such talk is so pervasive that it almost seems impossible that, short of pathological convolution, the experimental results of genetics can even be communicated without these resources.

What is largely forgotten is that there is good reason to believe that such talk of information in genetics may not be necessary: "information" was introduced in genetics only in 1953, just before the DNA double-helix model (Ephrussi et al., 1953). "Information" was supposed to introduce some frugality in a field that had become unusually profligate in terminological innovation (for instance, "transformation" and "transduction") to describe genetic interactions in bacteria. The main reason why the informational framework became central to the new molecular biology of the 1950s and 1960s was the characterization of the relationship between DNA and proteins as a universal genetic code, "universal" in the sense that it is the same for all species. A triplet of nucleotide bases specifies a single amino acid residue. Since there are four nucleotide bases (*A*, *C*, *G*, and *T*) in DNA, there are 64 possible triplets. Since there are only 20 standard amino-acid residues in proteins, the code must be degenerate (or partially redundant): more than one triplet must code for the same amino-acid residue. Three factors make the informational interpretation of this relationship illuminating: (a) the relationship can be viewed as a symbolic one, with each DNA triplet being a symbol for an amino acid residue; (b) the relationship is combinatorial, with different combinations of nucleotides potentially specifying different residues; and, most importantly, (c) the relationship is arbitrary in an important sense. Functional considerations may explain some features of the code – why, for instance, an arbitrary mutation tends to take a hydrophilic amino-acid residue to another such residue – but it does not explain why the code is specifically what it is. The physical mechanisms of translation do not help either. The genetic code is *arbitrary*. Along with specificity, this arbitrariness is what makes an informational account of genetics useful.

To explain how this account works will now require an explication of semiotic information (Sarkar 2000, forthcoming). This explication will proceed in two stages. After the structure of an information system is defined, conditions will first be laid down to establish specificity; secondly, more conditions will be imposed to capture the type of arbitrariness also required of semiotic information.

A formal information system consists of a relation, ι (the information relation), between two sets, A (for instance, a set of DNA sequences) and B (for instance, a set of polypeptide sequences) and will be symbolized as $\langle A, B, \iota \rangle$. The relation ι holds between A and B because it holds between their elements. The following simplifying assumptions will be made: (i) ι partitions A into a set of equivalence classes, with all member of an equivalence class being informationally equivalent to each other. (In the case of DNA and protein, an equivalence class consists of all the triplets that specify the same amino acid residue); and (ii) A and ι exhaust B – that is, for every element of B, there is some element in A that is related to it by ι. One consequence of this assumption is that ι partitions B into an equivalence class, with each equivalence class of B corresponding to one of A.

With this background, for an information system to allow for specificity, the most important condition on $\langle A, B, \iota \rangle$ is:

(I1) *Differential specificity*: suppose that a and a' belong to different equivalence classes of A. Then, if $\iota(a, b)$ and $\iota(a', b')$ hold, then b and b' must be different elements of B.

Condition (I1) suffices to capture the most basic concept of a specific informational relation holding between A and B.[1] An additional condition will sometimes be imposed to characterize a stronger concept of specificity:

(I2) *Reverse differential specificity*: if $i(a, b)$ and $i(a', b')$ hold, and b and b' are different elements of B, then a and a' belong to different equivalence classes in A.

If every equivalence class in A specifies exactly one element of B through ι, condition (I2) is automatically satisfied provided that (I1) is satisfied. That A and B co-vary (i.e., there are statistical correlations between the occurrences of elements of A and B) is a trivial consequence of either condition. The justification for these conditions is that they capture what is customarily meant by information in any context: for instance, they capture the sense in which the present positions and momenta of the planets carry information about their future positions.

By definition, if only condition (I1) holds, then A will carry *specific information* for B; if both conditions (I1) and (I2) hold, A *alone* carries specific information for B. In the case of prokaryotic genetics, both conditions hold: DNA alone carries information for proteins. Some care has to be taken at this stage and the discussion must move beyond formal specification to the empirical interpretation of ι in this (genetic) context. The claim that DNA alone carries information for proteins cannot be interpreted as saying that the presence of a particular DNA sequence will result in the production of the corresponding protein no matter what the cellular environment does. Such a claim is manifestly false. In certain cellular contexts, the presence of that DNA sequence will not result in the production of any protein at all. Rather, the claim must be construed counterfactually, if the presence of this DNA sequence were to lead to

1 Because ι may hold between a single a and several b, this relation is not transitive. A may carry information for B, B for C, but A may not carry information for C. As will be seen later, this failure of transitivity results in a distinction between encoding traits and encoding information for traits.

the production of a protein, then ι describes which protein it leads to. In general, the production of the relevant protein requires an enabling history of environmental conditions. This history is not unique. Rather, there is a set M of "standard" histories that result in the formation of protein from DNA. A complete formal account of biological information must specify the structure of this set. This is beyond what is possible given the current state of empirical knowledge: M cannot be fully specified even for the most studied bacterium, *Escherichia coli*. However, because the relevant informational relation is being construed counterfactually in the way indicated, the inability to specify M does not prevent the use of that relation. (It does mean, though, that the information content of the DNA, by itself, does not suffice as an etiology for that protein.) The structure of the relation between M and the protein set will be a lot more complicated than $\langle A, B, \iota\rangle$, where A is the DNA set and B the protein set; this is already one sense in which genes are privileged.

So far, only specificity has been analyzed; it remains to provide an account of arbitrariness. This requires the satisfaction of two conditions that are conceptually identical to those imposed by Shannon on communication systems in order to achieve a suitable level of abstraction. It shows an important commonality between communication-theoretic and semiotic information:

(A1) *Medium independence*: $\langle A, B, \iota\rangle$ can be syntactically represented in any way, with no preferred representation, so long as there is an isomorphism between the sets corresponding to A, B, and ι in the different representations.[2]

An equation representing the gravitational interaction between the earth and the sun is a syntactic representation of that relation. So is the actual earth and sun, along with the gravitational interaction. The latter representation is preferred because the former is a representation of the latter in a way in which the latter is not a representation of the former. Medium independence for information denies the existence of any such asymmetry. From the informational point of view, a physical string of DNA is no more preferred as a representation of the informational content of a gene than a string of *A*s, *C*s, *G*s, and *T*s on this sheet of paper. (The most useful analogy to bear in mind is that of digital computation, where the various representations, whether it be in one code or other or as electrical signals in a circuit, are all epistemologically on a par with each other.) This is also the sense of medium independence required in Shannon's account of communication. The second criterion is:

(A2) *Template assignment freedom*: let A_i, $i = 1, \ldots, n$, partition A into n equivalence classes, and let $\langle A_1, A_2, \ldots, A_n\rangle$ be a sequence of these classes that are related by ι to the sequence $\langle B_1, B_2, \ldots, B_n\rangle$ of classes of B. Then the underlying mechanisms allow for any permutation $\langle A_{\sigma(1)}, A_{\sigma(2)}, \ldots, A_{\sigma(n)}\rangle$ (where $\sigma(1)\ \sigma(2) \ldots \sigma(n)$ is a permutation of $12 \ldots n$) to be mapped by ι to $\langle B_1, B_2, \ldots, B_n\rangle$.

This condition looks more complicated than it is; as in the case of Shannon's communication systems, it shows that any particular protocol of coding is arbitrary.

2 Here, ι is being interpreted extensionally for expository simplicity.

(However, as emphasized in the last section, coding and decoding are not part of information transfer in the theory of communication.) In the case of DNA (the template), all that this condition means is that any set of triplets coding for a particular amino acid residue can be reassigned to some other residue. However, there is a subtle and important problem here: such a reassignment seems to violate a type of physical determinism that no one questions in biological organisms, namely that the chemical system leading from a piece of DNA to a particular protein is deterministic. (Indeed, conditions [I1] and [I2] of specificity require determinism of this sort.) Condition (A2) must be interpreted in the context of genetics as saying that all these different template assignments were evolutionarily possible.[3] The customary view, that the genetic code is an evolutionarily frozen accident, supports such an interpretation. By definition, if conditions (I1), (A1), and (A2) hold, then ι is a coding relation, and *A encodes B*. Thus, DNA encodes proteins; for prokaryotes, DNA alone encodes proteins.

The formal characterization given above lays down adequacy conditions for any ι that embodies semiotic information. It does not specify what ι is; that is, which *a*'s are related to which *b*'s. The latter is an empirical question: coding and similar relations in molecular biology are empirical, not conceptual, relations. This is a philosophically important point. That DNA encodes proteins is an empirical claim. Whether conditions (I1), (I2), (A1), or (A2) hold is an empirical question. Under this interpretation, these conditions must be interpreted as empirical generalizations with the usual *ceteris paribus* clauses that exclude environmental histories not belonging to *M*. The relevant evidence is of the kind that is typically considered for chemical reactions: for the coding relation, subject to statistical uncertainties, the specified chemical relationships must be shown to hold unconditionally. Now suppose that, according to ι, a string of DNA, *s*, codes for a polypeptide, σ. Now suppose that, as an experimental result, it is found that some σ' different from σ is produced in the presence of *s*. There must have been an *error*, for instance, in transcription or translation: a *mistake* has been made. All this means is that an anomalous result, beyond what is permitted due to standard statistical uncertainties, is obtained. It suggests the operation of intervening factors that violate the *ceteris paribus* clauses; that is, the reactions took place in an environmental history that does not belong to *M*. Thus, the question of what constitutes a mistake is settled by recourse to experiment. There is nothing mysterious about genetic information, nothing that requires recourse to conceptual resources from beyond the ordinary biology of macromolecules.

As was noted, the informational account just given allows DNA alone to encode proteins' prokaryotic genetics. Genetically, prokaryotes are exceedingly simple. Every bit of DNA in a prokaryotic genome either codes for a protein or participates in regulating the transcription of DNA. For the coding regions, it is straightforward to translate the DNA sequence into the corresponding protein. Consequently, the informational interpretation seems particularly perspicuous in this context. This is the

3 However, condition (A2) is stronger than what is strictly required for arbitrariness, and may be stronger than what is biologically warranted. For arbitrariness, all that is required is that there is a large number of possible alternative assignments. Biologically, it may be the case that some assignments are impossible because of developmental and historical constraints.

picture that breaks down in the eukaryotic context. Eukaryotic genetics presents formidable complexities including, but not limited to, the following (for details, see Sarkar, 1996a):

(a) the nonuniversality of the standard genetic code – some organisms use a slightly different code, and the mitochondrial code of eukaryotic cells is also slightly different;
(b) frameshift mutations, which are sometimes used to produce a variant amino-acid chain from a DNA sequence;
(c) large segments of DNA between functional genes that apparently have no function at all and are sometimes called "junk DNA";
(d) similarly, segments of DNA within genes that are not translated into protein, these being called introns, while the coding regions are called exons – after transcription, the portions of RNA that correspond to the introns are "spliced" out and not translated into protein at the ribosomes; but
(e) there is alternative splicing by which the same RNA transcript produced at the DNA (often called "pre-mRNA") gets spliced in a variety of ways to produce several proteins – in humans, it is believed that as many as a third of the genes lead to alternative splicing; and
(f) there are yet other kinds of RNA "editing" by which bases are added, removed, or replaced in mRNA, sometimes to such an extent that it becomes hard to say that the corresponding gene encodes the protein that is produced.

In the present context, points (a), (b), (d), (e), and (f) are the most important. They all show that a single sequence of DNA may give rise to a variety of amino-acid sequences even within a standard history from *M*. Thus, in the relation between eukaryotic DNA and proteins, condition (I2) fails. Nevertheless, because condition (I1) remains satisfied, eukaryotic DNA still carries specific information for proteins. However, this formal success should not be taken as an endorsement of the utility of the informational interpretation of genetics. As stated, condition (I1) does not put any constraint on the internal structure of the sets *A* (in this case, the set of DNA sequences) or *B* (in this case, the set of protein sequences). If both sets are highly heterogeneous, there may be little that the informational interpretation contributes. In the case of the DNA and protein sets, heterogeneity in the former only arises because of the degeneracy of the genetic code. This heterogeneity has not marred the utility of the code. For the protein set, heterogeneity arises because a given DNA sequence can lead to different proteins with varied functional roles. However, leaving aside the case of extensive RNA editing, it seems to be the case that these proteins are related enough for the heterogeneity not to destroy the utility of the informational interpretation.

The main upshot is this: even in the context of a standard history from *M*, while genes encode proteins and carry specific information for them, this information is not even sufficient to specify a particular protein. Knowledge of other factors, in particular so-called environmental factors, is necessary for such a specification. Whether these other factors should be interpreted as carrying semiotic information depends on whether they satisfy at least conditions (I1), (A1), and (A2). So far, there is no evidence that any of them satisfy (A1) and (A2). This is yet another sense in which genes are privileged.

13.4 Traits and Molecular Biology

Proteins are closely linked to genes: compared to organismic traits such as shape, size, and weight, the chain of reactions leading to proteins from genes is relatively short and simple. It is one thing to say that genes encode proteins, and another to say that they encode information for traits. It depends on how traits are characterized. The first point to note is that "trait" is not a technical concept within biology, with clear criteria distinguishing those organismic features that are traits from those that are not (Sarkar, 1998). In biological practice, any organismic feature that succumbs to systematic study potentially qualifies as a trait. Restricting attention initially to structural traits, many of them can be characterized in molecular terms (using the molecules out of which the structures are constructed). These occur at varied levels of biological organization from the pigmentation of animal skins to receptors on cell membranes. Proteins are intimately involved in all these structures. For many behavioral traits, a single molecule, again usually a protein, suffices as a distinguishing mark for that trait. Examples range from sickle-cell hemoglobin for the sickle-cell trait to huntingtin for Huntington's disease. Thus, for these traits, by encoding proteins, genes trivially encode information for them at the molecular level.

However, at the present state of biological knowledge it is simply not true that, except for a few model organisms, most traits can be characterized with sufficient precision at the molecular level for the last argument to go through. This is where the situation becomes more interesting. The critical question is whether the etiology of these traits will permit explanation at the molecular level, in particular, explanations in which individual proteins are explanatorily relevant. (Note that these explanations may also refer to other types of molecules; all that is required is that proteins have some etiological role. The involvement of these other molecules allows the possibility that the same protein (and gene) may be involved in the genesis of different traits in different environmental histories – this is called *phenotypic plasticity*.) If proteins are so relevant, then there will be covariance between proteins and traits although, in general, there is no reason to suppose that any of the conditions for semiotic information is fulfilled. Proteins will be part of the explanation of the etiology of traits without carrying information for traits. Allowing genes to carry information does not endorse a view that organisms should be regarded as information-processing machines. Informational analysis largely disappears at levels higher than that of proteins.

Nevertheless, because of the covariance mentioned in the last paragraph, by encoding proteins, genes encode information for traits (while not encoding the traits themselves). The critical question is the frequency with which proteins are explanatorily relevant for traits in this way. The answer is "Probably almost always." One of the peculiarities of molecular biology is that the properties of individual molecules, usually protein molecules such as enzymes and antibodies, have tremendous explanatory significance, for instance, in the structural explanations of specificity mentioned in section 13.1. Explanations involving systemic features, such as the topology of reaction networks, are yet to be forthcoming. Unless such explanations, which do not refer to the individual properties of specific molecules, become the norm of organismic biology, its future largely lies at the level of proteins.

This is a reductionist vision of organismic biology, not reduction to DNA or genes alone, but to the entire molecular constitution and processes of living organisms (Sarkar, 1998). Such a reductionism makes many uncomfortable, including some developmental biologists imbued in the long tradition of holism in that field, but there is as yet no good candidate phenomenon that contradicts the basic assumptions of the reductionist program. It is possible that biology at a different level – for instance, in the context of psychological or ecological phenomena – will present complexities that such a reductionism cannot handle. However, at the level of individual organisms and their traits, there is as yet no plausible challenge to reductionism.

13.5 Conclusion

Many of the recent philosophical attacks on the legitimacy of the use of an informational framework for genetics have been motivated by a justified disquiet about facile attributions of genetic etiologies for a wide variety of complex human traits, including behavioral and psychological traits, in both the scientific literature and in the popular media. The account of semiotic information given here does not in any way support such ambitions. While, to the extent that is known today, genes are privileged over other factors as carriers of semiotic information involved in the etiology of traits, genes may not even be the sole purveyors of such information. Moreover, genetic information is but one factor in these etiologies. Traits arise because of the details of the developmental history of organisms, in which genetic information is one resource among others. Thus, even when genes encode sufficient information to specify a unique protein (in prokaryotes), the information in the gene does not provide a sufficient etiology for a trait; at the very least, some history from *M* must be invoked.

Thus, an informational interpretation of genetics does not support any attribution of excessive etiological roles for genes. The two questions are entirely independent: whether genes should be viewed as information-carrying entities, and the relative influence of genes versus nongenetic factors in the etiology of traits. This is why the latter question could be investigated, often successfully, during the first half of the twentieth century, when "information" was yet to find its way into genetics. To fear genetic information because of a fear of genetic determinism or reductionism is irrational.

Acknowledgments

Thanks are due to Justin Garson and Jessica Pfeifer for helpful discussions.

Bibliography

Brooks, D. R. and Wiley, E. O. 1988: *Evolution as Entropy: Towards a Unified Theory of Biology*, 2nd edn. Chicago: The University of Chicago Press.

Carnap, R. and Bar-Hillel, Y. 1952: An outline of a theory of semantic information. Technical Report No. 247. Research Laboratory of Electronics, Massachusetts Institute of Technology.

Crick, F. H. C. 1958: On protein synthesis. *Symposia of the Society for Experimental Biology*, 12, 138–63.

Ephrussi, B., Leopold, U., Watson, J. D., and Weigle, J. J. 1953: Terminology in bacterial genetics. *Nature*, 171, 701.

Fisher, R. A. 1925: Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700–25.

Garson, J. 2002: The introduction of information in neurobiology. M. A. thesis, Department of Philosophy, University of Texas at Austin.

Griffiths, P. 2001: Genetic information: a metaphor in search of a theory. *Philosophy of Science*, 67, 26–44.

Hartley, R. V. L. 1928: Transmission of information. *Bell Systems Technical Journal*, 7, 535–63.

Kimura, M. 1961: Natural selection as a process of accumulating genetic information in adaptive evolution. *Genetical Research*, 2, 127–40.

Lederberg, J. 1956: Comments on the gene–enzyme relationship. In Gaebler, O. H. (ed.), *Enzymes: Units of Biological Structure and Function*. New York: Academic Press, 161–9.

Maynard Smith, J. 2000: The concept of information in biology. *Philosophy of Science*, 67, 177–94.

Nyquist, H. 1924: Certain factors affecting telegraph speed. *Bell Systems Technical Journal*, 3, 324–46.

Pauling, L. 1940: A theory of the structure and process of formation of antibodies. *Journal of the American Chemical Society*, 62, 2643–57.

Pierce, J. R. 1962: *Symbols, Signals and Noise*. New York: Harper.

Sarkar, S. 1996a: Biological information: a skeptical look at some central dogmas of molecular biology. In Sarkar, S. (ed.), *The Philosophy and History of Molecular Biology: New Perspectives*. Dordrecht: Kluwer, 187–231.

—— 1996b: Decoding "coding" – information and DNA. *BioScience*, 46, 857–64.

—— 1998: *Genetics and Reductionism*. New York: Cambridge University Press.

—— 2000: Information in genetics and developmental biology: comments on Maynard-Smith. *Philosophy of Science*, 67, 208–13.

—— forthcoming: Biological information: a formal account.

Schneider, T. D. 1999: Measuring molecular information. *Journal of Theoretical Biology*, 201, 87–92.

Shannon, C. E. 1948: A mathematical theory of communication. *Bell Systems Technical Journal*, 27, 379–423, 623–56.

Sterelny, K. 2000: The "genetic program" program: a commentary on Maynard Smith on information in biology. *Philosophy of Science*, 67, 195–202.

—— and Griffiths, P. E. 1999: *Sex and Death: An Introduction to Philosophy of Biology*. Chicago: The University of Chicago Press.

Watson, J. D., Tooze, J., and Kurtz, D. T. 1983: *Recombinant DNA: A Short Course*. New York: W. H. Freeman.

Williams, G. C. 1966: *Adaptation and Natural Selection*. Princeton, NJ: Princeton University Press.

Further reading

Kay, L. 2000: *Who Wrote the Book of Life?: A History of the Genetic Code*. Stanford, CA: Stanford University Press.

Schrödinger, E. 1944: *What is Life? The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press.

Thiéffry, D. and Sarkar, S. 1998: Forty years under the central dogma. *Trends in Biochemical Sciences*, 32, 312–16.

Yockey, H. P. 1992: *Information Theory and Molecular Biology*. Cambridge: Cambridge University Press.

CONTEMPORARY DEBATES IN PHILOSOPHY OF SCIENCE

Edited by

Christopher Hitchcock

Blackwell
Publishing

350 Main Street, Malden, MA 02148-5020, USA
108 Cowley Road, Oxford OX4 1JF, UK
550 Swanston Street, Carlton, Victoria 3053, Australia

First published 2004 by Blackwell Publishing Ltd

Library of Congress Cataloging-in-Publication Data

Contemporary debates in philosophy of science / edited by Christopher Hitchcock.
p. cm. — (Contemporary debates in philosophy ; 2)
Includes bibliographical references and index.
ISBN 1-4051-0151-2 (alk. paper) — ISBN 1-4051-0152-0 (pbk. : alk. paper)
1. Science—Philosophy. I. Hitchcock, Christopher, 1964– II. Series.

Q175.C6917 2004
501—dc22

2003016800

A catalogue record for this title is available from the British Library.

Set in 10/12½ pt Rotis serif
by SNP Best-set Typesetter Ltd., Hong Kong
Printed and bound in the United Kingdom
by MPG Books Ltd, Bodmin, Cornwall

For further information on
Blackwell Publishing, visit our website:
http://www.blackwellpublishing.com