5 On Genomes and Cosmologies

Antoine Danchin

Such a provocative title was selected for this chapter because the status of informatics in biology, and especially in genome studies, probably has not yet reached the place where it ought to be. It is this author's hope that readers will be stimulated not only to read what follows but to consider whether biology has changed recently since it has been accepted that many genomes have been, or soon will be, amenable to complete sequencing. We will presently possess the total chemical definition (including intermediary metabolism) of many living organisms. Nonetheless, perspective we shall soon have about living organisms will, in many ways, mirror our perspective of stars in cosmology, insofar as we shall be able to propose models to interpret the data we collect but shall forever lack the ability to construct *direct* experiments to test the models. As in cosmology, this does not mean that we cannot say pertinent things about genomes or that we can say nothing: Models can be tested by their internal consistency and by their predictive value, when they predict, for example, that a new object can (or will) be discovered (existential prediction rather than falsification).

Information, a generally poorly defined concept, will be central in our discussion. This concept goes back to the Aristotelian discussion of the Being as a relationship among substance, matter, and form: *Substance* is the being itself, in which one can distinguish a determinable *matter* and a determining *form*. The differentiation between individual entities usually was ascribed to specific features of form, and the study of causes asked for a special process *informing* matter. Thus, information is related not only to the specificity given by form—and DNA sequences are related to a very abstract conception of form—but also, implicitly, to causality. This may explain many of the wrong paths that have been followed when using the corresponding concept of information: In ordinary language, something very different is meant when the word *information* is used: *Information* means some kind of usefulness brought about by a given piece of knowledge on a system. Information is therefore also implicitly associated with the idea of *value*. This observation must be born in mind when considering the various theories of information that can be used to represent living organisms and, in particular, to represent DNA sequences.

Invention of techniques that permit DNA sequencing has brought about a true revolution in our way of considering genomes as a whole. Initially, most studies of DNA sequences dealt with a *local* analysis of the DNA information content. They took into account, however, the availability of homologous structures or functions from different organisms and thus provided some general view of global properties of genomes. This remains the dominant view of informatics applied to genetics. As evidenced in many different periodicals and other publications, and more recently, researchers have come to think of genomes as *global* entities. In the future, computers can be used as experimental tools, generating a new source of investigation of living organisms, their study *in silico* (in contrast to in vivo or in vitro).

A LOCAL VIEW OF GENOMES: ALIGNMENT AND CONSENSUS OF MOLECULAR SEQUENCES

Since protein and nucleotide sequences have been available (and this goes back to the early 1950s for protein sequences), scientists have attempted to compare sequences to one another and to align those sequences that were supposed to match one another after having diverged from a common ancestor. A paragon study of such comparison was the *Atlas of Protein Sequences and Structures* developed by Margaret Dayhoff and her coworkers (Dayhoff, Schwartz, and Orcutt, 1978). After thorough analyses of the phylogenetic divergence in families of globular proteins such as cytochromes or globins, this work permitted construction of a correspondence matrix between amino acids that reflected some of the evolutionary fate of each residue. Hundreds of articles have been written proposing general and less general approaches for creating alignment of sequences and generating the corresponding consensus (e.g., see Altschul and Erickson, 1985; Waterman and Eggert, 1987; Fischel-Ghodsian, Mathiowitz, and Smith, 1990; Gotoh, 1990a, b; Spouge, 1991).

Although it was relatively easy to align similar sequences, it often was impossible to align very divergent sequences unless some other related sequences also were known and were compared in parallel (multiple rather than pairwise alignment). This resulted in the hunt for intelligent algorithms permitting alignments and multialignments and providing the corresponding consensus of multiple sequences. However, it is clear that the production of alignments requires not only production of a matrix for correspondence between nucleotides (usually the identity matrix in this case) or similar amino acid residues (Altschul, 1991; Alexandrov 1992; Landès, Hénaut, and Risler, 1992), but also a choice for the penalties assigned to the gaps present in the sequences when they are aligned (Altschul, 1989; Barton and Sternberg, 1990; Gotoh, 1990b; Blaisdell, 1991; Brouillet, Risler, Hénaut, and Slonimski, 1992; Pearson and Miller, 1992; Rose and Eisenmenger, 1991; Saqi and Sternberg, 1991; Zhu, Sali, and Blundell, 1992).

The use of dynamic programming can provide a first set of aligned pairs (Smith, 1988). Many programs, derived from or created independently from the Needleman and Wunsch algorithm, including use of dedicated hardware, have been proposed (Waterman, 1983; Waterman and Byers, 1985; Gribskov et al., 1987; Taylor and Orengo, 1989; Fischel-Ghodsian et al., 1990; Smith and Smith, 1990; Rose and Eisenmenger, 1991; Saqi and Sternberg, 1991; Pearson and Miller, 1992; Fagin, Watt, and Gross, 1993; Subbiah, Laurents, and Levitt, 1993; Vogt and Argos, 1993). Unfortunately, these programs cannot be used for sets larger than three sequences because of combinatorial explosion.

In contrast, looking for common words in a composite sequence made of the concatenation of all sequences to align, followed by a manual fine alignment in which the unformalized knowledge of the biologist is used, can yield interesting results (see Schuler, Altschul, and Lipman, 1991; Zuker, 1991; and Pearson and Miller, 1992, for comparison of different approaches and inclusion of statistical error assessment). A typical example of such combination is the Smarties software[1] designed by Viari and coworkers for multialignment of proteins, using an adaptation of the algorithm of Karp, Miller, and Rosenberg (1972), in which one combines an algorithm searching for matching patterns with "anchor" sites that correspond to matching patterns chosen by the biologist and aligned forcefully. In this process, matrices meant to represent similarities between amino acids can be used for protein alignment. It is interesting to use matrices that take into account the constraints on amino acid residues that have been measured from comparison of known three-dimensional structures (Risler, Delorme, Delacroix, and Hénaut, 1988; Miyazawa and Jernigan, 1993); multiple alignments can also implement knowledge from such structures (Holm and Sander, 1993; Johnson, Overington, and Blundell, 1993; Orengo, Flores, Jones, Taylor, and Thornton, 1993; Pongor et al., 1993; Subbiah et al., 1993). Nonetheless, it must be stressed that a single matrix is used throughout the sequences, meaning that one deals with a composite view of the equivalence rules between amino acids at each position in the sequence (Saqi and Sternberg, 1991). This cannot represent what happens in reality because at some positions for example, it is the aromatic feature that is important (meaning that F [phenylalanine], Y [tyrosine], and W [arginine] are equivalent), whereas at other positions the fact that the residue is big could be critical (meaning that F, Y, R, H [histidine], and W [tryptophan] are equivalent), so that one should use an equivalence matrix for each position in the sequence. Clearly, new research program still need to be developed in this field, despite the large number of articles already published (cf. Day and McMorris, 1993).

Learning Techniques, Perceptron, and Neural Networks

When significant numbers of sequences are compared with one another, one may wish to extract a significant pattern not as a consensus sequence but

rather as a matrix of probability reflecting the frequency of a given motif at a given place. This is reminiscent of pattern recognition problems involving, for example, learning techniques or techniques used in visual pattern recognition (Nussinov and Wolfson, 1991; Fischer et al., 1992; Bachar et al., 1993). It is not unexpected, therefore, that methods involving neural networks or the precursor of this new field of informatics, the Perceptron, have been used for generating pertinent patterns. Stormo and coworkers (1982b) were pioneers the field, using the Perceptron, a classical learning method in pattern recognition research, for creating matrices that permitted identification of the ribosome binding sites in *Escherichia coli* messenger RNAs, and they have been followed by many other authors studying both DNA and protein sequences (Schneider, Stormo, Gold, and Ehrenfeucht, 1986; Alexandrov and Mironov, 1987; Bohr et al., 1988; Nakata, Kanehisa, and Maizel, 1988; Stormo, 1988; Stormo and Hartzell, 1989; Hirst and Sternberg, 1991; Farber, Lapedes, and Sirotkin, 1992; Prestridge and Stormo, 1993; Snyder and Stormo, 1993). The main drawback of the Perceptron approaches, which unfortunately is present also in many neural network learning techniques, is that they involve formal "synapses" that have a transmitting strength which evolves in a *quantitative* fashion as a function of their actual use (Stormo et al., 1982b; Geourjon and Deleage, 1993; Gracy, Chiche, and Sallantin, 1993; Han and Kim, 1993; Hirosawa, Hoshida, Ishikawa, and Toya, 1993; Jones, Orengo, Taylor, and Thornton, 1993; Landès, Hénaut, and Risler, 1993; Miyazawa and Jernigan, 1993; Orengo et al., 1993; Orengo and Taylor, 1993; Orengo and Thornton, 1993; Panjukov, 1993; Pongor et al., 1993; Lawrence et al., 1993). This means that, except in the cases where there is obvious and strong similarity between objects in the training set, the synapse strength fluctuates. As a consequence, when the set size increases, the strength of each synapse generally goes through an optimum value and then slowly recedes toward a more or less average value as more exceptions invade the training set, thus losing discriminative power. Accordingly, as the training set of examples increases in size, the actual efficiency of the consensus matrix created by the learning procedure loses accuracy. Ancestors of neural networks have been proposed that evolve in such a way that the effective transmitting capacity of a synapse goes irreversibly to zero when its value falls below a certain threshold value, thereby freezing the learning state at an optimal value (Danchin, 1979). This feature has recently been implemented in neural networks by Horton and Kanehisa (1992). It would therefore be important to compare and develop new approaches involving neural networks, with emphasis on the *stability* of their learning capacity as a function of the training sets.

Other learning techniques rest on the explicit construction of grammars—that is, sets of structures and rules permitting the complete description of a sequence or a network of genes (see chapter 9). They can be divided roughly into two classes, learning by assimilation (i.e., by measuring *conformity* to features of a training set) and learning by discrimination (i.e., by constructing

descriptors of the *differences* between the sets of examples supposed to differ from one another). An example of the first type is the Calm software, used for identifying the calcium-binding sites in calmodulin, for instance (Haiech and Sallantin, 1985), whereas an example of the second type is the Plage software used in identifying the relevant features of signal peptides in *E. coli* (Gascuel and Danchin, 1986; Gascuel, 1993). Both types depend on a grammar that is employed to generate descriptors used on the training sets. In general, grammars are essential to describe biological phenomena, and their efficiency rests heavily on the actual biological knowledge of their conceptors, providing therefore an interesting way to incorporate biological knowledge into automated procedures. For this reason, one probably should consider always as the most pertinent ones the grammars devoted to a specific question. A case in point is the work of Collado-Vides (1991a, b, 1992), who constructed a grammar for the identification of regulated promoters in *E. coli* (see also chapter 9).

Sequence acquisition has been exponentially growing for 10 years. In parallel, the power of computing has also increased in an exponential fashion, and it has been possible to foresee treatment of biological data that would have been impossible just a few years ago (see Barton, 1992). Because DNA and protein sequences can be seen, to first approximation, as linear texts of an alphabetical writing, they pertain to the theories of information that have precisely the study of such texts as privileged objects. In what follows, we attempt to develop ideas showing that the alphabetical metaphor, and the corresponding treatment of information, is likely to have a very deep meaning in terms of genomes. This will require a progression in the conceptualization of information and, because biologists are not always familiar with the underlying mathematics, rough principles are provided so that such researchers can follow the reasons for working to improve our concepts of information.

Shannon and Weaver's Model for the Information Seen During DNA Replication

When considering living organisms, one often casually refers to the *information* they carry, namely in their genomes' DNA sequence. Information is a "prospective character" (Myhill, 1952), which refers to the identification or localization of control processes involving small amounts of energy that result in large changes in processes involving large quantities of matter or energy. A very elementary, but useful and often-used, concept of information has been derived by Shannon and, independently, by Weaver, who published their findings together (Shannon and Weaver, 1949). These authors, by comparing the sequence before and after transmission, defined the information of a collection of digit sequences transmitted through an electromagnetic wave channel submitted to some noise (Shannon, 1949). The information concept they invented is a very crude view of a message, explicitly considering

not its meaning but the local accuracy of its transmission. Genomes can be analyzed using this view, especially if one considers only the information carried out during the process of replication, leaving aside the information involved in gene expression. This concept of information can shed light on interesting properties, but one should always bear in mind the fact that the actual information content of a sequence must be much richer (it has a meaning), unless one considers only the information seen by the replication machinery. Later we shall see other ways of considering the information content of sequences, cases in which the global signification is better taken into account.

Shannon and Weaver's theory of information derives from a subset of the general theory of measurable spaces (Benzécri, 1984). Given a measurable space Ω (space of events) and a a part of Ω, if ω is an event, chosen at random, one would like to say that if one knows that ω belongs to a, this means that we have some information on ω. Thus, information brought about by asserting $\omega \in a$, is noted $H(a)$. H (Shannon's information) is defined in probabilistic terms as follows: $H(a) = f(p_a)$, where p_a is the probability that the information on a is true (the smaller a, the easier the location of ω). First, f is supposed to be positive and decreasing and such that a random assertion having equal probabilities to be true or false gives a unit of information: $f(1/2) = 1$ (this is a natural condition for normalization). Second, f is such that if A and B are two partitions of W and if the assertions $\omega \in A$ and $\omega \in B$ are independent, then $H(A \otimes B) = H(A) + H(B)$. Finally, one adds the natural condition that $f(1) = 0$, meaning that if an event is known to be true for certain (ω belongs to Ω is trivial), this is equivalent to give no information on it. A consequence of this definition of information is that if Ω is split into a number of n of equivalent separate subsets Ω_n, one has, after the final property of H:

$$H(W_n) = n \times (1/n) \times f(1/n) = f(1/n)$$

This can be used as a starting point to show that function f is:

$$\forall p \in (0, 1) \qquad f(p) = -\log_2(p)$$

More generally, let $i \in I$ be a random variable of probability law p_I (where $p_I = \{p_i | i \in I\}$; then Shannon's information becomes:

$$H(p_i) = -\sum \{p_i \log_2 p_i | i \in I\}$$

It can be seen that such definition of information is highly schematic and far from being able to describe the properties of what one's common sense would consider information. Indeed, a more elaborate view would indicate that to consider an event in isolation is meaningless in most actual cases: One needs, for instance, to possess a means to recover the considered information. Also, when one considers life, the present approach of information does not take semantics into account, as would be necessary if one considers gene expression.

Another way to consider the information carried in a message, when it corresponds to an instance of a family of messages considered as similar, leads to the same equation. If p_r is the probability of a message when received, and p_s the probability of this message when it is sent, then the information corresponding to the quality of the message is: $H = \log_2(p_r/p_s)$. If one does not take into account the loss of information during transmission (message transmitted without error), then p_r is equal to unity. If there is a large number of similar messages, the corresponding information is $H = -\sum p_s \log_2 p_s$. One notes here that the equation obtained has still the same form as in the preceding description.

In the case of DNA, what is taken into account is the probability of a message that, in the usual description of genomes, reduces to the actualization of a succession of letters (chosen among A, T, C, or G) at a given position of the genomic text. The specific information analyzed corresponds to the difference between an actual collection of DNA sequences and a random sequence having the same base composition. An evident improvement would be to consider not the one-letter successions of the program but the relevant words (sequences of letters of definite length). However, this assumes that one knows the ones that are relevant! It would appear, therefore, that analysis of the genomic text using classification methods is a very important prerequisite for an in-depth analysis of its information content.

The mathematical aspect of Shannon's information is developed to display the logarithmic law to which it leads. Indeed, the form of this law is similar to that of another law discovered in the domain of thermophysics. And this coincidence has had unfortunate conceptual consequences, when information was related to life, because the ideas of the latter scientific domain contaminated the ideas of the former one, as we shall now see (see also Yockey, 1992).

Information and Shannon's Entropy

It is quite usual to read that there is *identity* between a form of thermodynamic entropy and the information carried by a given process (but see the recent analysis by Zurek, 1989). In fact, Shannon's information often is named *Shannon's entropy* (but this was not a choice made by Claude Shannon!). It is a consequence of the remark that there is identity between the form of Shannon's law H and a form of entropy described in terms of statistical mechanics, particularly when describing the specific case of the perfect gas. However, identity in the form of a mathematical description of a process is by no means a sufficient reason to identify the underlying principles or the very processes with one another. Moreover several words such as *information, entropy,* and *order* have been used in many different contexts, meaning something very precise in each but usually meaning something very different from the word's significance in another context. Because the ideas of information and order are used often in biology and may

influence the exercise of the corresponding knowledge, it seems useful to explore some of the underlying reasons that have permitted authors sometimes, to identify the negative value of information and entropy.

The frequent identification between information and entropy is ideological in nature and corresponds to a mistaken identification of disorder and entropy coupled with an unsaid view of the world as spontaneously evolving toward a "bad" ordering. The underlying assumption is that the world is constantly moving toward an unavoidable disorder, a fact against which we should fight (see Schrödinger's *What Is Life?* and my foreword to its French translation [Danchin, 1986]). The starting point of such pessimistic views of the world is the mechanostatistical model of thermodynamics, founded on the very specific and restricted case of the perfect gas. In this model, gas atoms are hard, impenetrable spherical points that interact by pure elastic collisions. The popular description then is that of a container made of two independent chambers containing different gases and separated by a rigid wall. At the start of the process, a small hole is drilled through the wall. It is easy to observe subsequently that the container evolves in such a way that in time there is an equivalent mixture of both gases in both chambers, each gas having a tendency to occupy the volume from which it was absent, until a final equilibrium is achieved, wherein the partial pressure of the gas is equal in both chambers, an equal number of gas atoms coming from each chamber to the other. If one wanted to prevent homogenization, it would be necessary to place a little demon at the hole who had control of a door that could be closed when the wrong gas was going toward the chamber in which it did not belong. One then asserts that the action of the demon corresponds to the knowledge of information on the gases' nature.

This description is considered adequate to justify the identification between information and entropy. Many technical objections to such a primitive description could be given, but one is compelling: In this way of accounting for a system, its entropy depends on its level of description (see also Zurek, 1989). However, *entropy* can also be defined as a macroscopical physical entity linked to the total energy of a system and its temperature. It therefore is independent of the model. Thus, the identification between entropy and disorder is disputable. Let us consider a two-dimensional representation: A blue gas in a square chamber is separated from a yellow gas in a similar chamber. Drilling a hole in wall between them will yield a green rectangle created by mixture of the blue and yellow gases. Is the green rectangle more disordered than the juxtaposition of a blue square with a yellow one? Identification between entropy and information also is misleading, because one can immediately notice here not only that information is a qualification of a state of a given system but that it requires some knowledge of both the system and of the observer. Like order, information is a concept that is *relative* and must take into account several systems.

Finally, when considering the entropy of a system made of many components, one should take into account not only the distribution of the compo-

nents in space and according to their energy states but also the various scales at which they occur: There is some entropy contribution in the building up of a hierarchy, for instance. If one wished to have information on a system, it might be important to consider the structure of the hierarchy. There is also a most important phenomenon: If one considers processes that make a large number of components evolve, there are at least two well-defined evolutions that should be taken into account. One corresponds to systems in which the overall structure of individual levels is preserved, even if the individual components of each level seem to evolve in a stochastic way. This is generally the case of ergodic processes. However, there is another case in which the overall tendency is the mixing up of everything, the overall tendency being that of homogenization. In this case, calculated entropy has a special form, which tells us something about the tendency of the system to lose memory of past events.

Without qualifications about the general structure of the underlying dynamics (ergodic or mixing) or about the observer, the description of information in statistical terms is highly misleading. Shannon's measure of entropy reflects only a mathematical property of statistical analysis, as discussed later (see Benzécri, 1984). For the moment, however, limiting Shannon's information to what its mathematical form assumes it is, this information can be used for a first description of DNA sequences' information content. This first level of local information can be used to identify coding regions in DNA (Shepherd, 1981; Smith, Waterman, and Sadler, 1983; Almagor, 1985; Graniero-Porati and Porati, 1988; Luo, Tsai, and Zhou, 1988; Gusein-Zade and Borodovsky, 1990; States and Botstein, 1991; Farber, Lapedes, and Sirotkin, 1992; Robson and Greaney, 1992). A more evolved use of Shannon's information is also present in the analysis of gene sequences using periodical Markov chain analysis, as demonstrated by Borodovsky and McIninch (1992). A consequence of this type of work is obtaining a large collection of gene sequences from a given organism, which brings us to the use of informatics in the analysis of whole genomes.

LANDSCAPES: A GLOBAL VIEW OF GENOMES

A genome is the result of a historical process that has driven a variety of genes to function collectively. Many biologists using computers have had the following experience: When one scans through a sequence library for similarity with a query sequence, one often ends up with a list of sequences that have in common only the organism to which the initial sequence belongs. This is as if there existed a *style* specific for each organism, independent of the function (an image proposed by P. Słonimski at a public conference is that of columns in a temple, which can be of Corinthian, Doric, or Ionian style but having always the same function). Is this impression true? Several ways of investigating the overall structure of genomes have confirmed and extended this observation (Fickett, Torney, and Wolf, 1992; Karlin and Brendel, 1992).

The indication is that there must exist some sort of mutual interaction which permits the cell to discriminate between self and nonself. It therefore seems interesting to investigate the meaning, at the lowest possible level, of what could be mutual information.

Mutual Information and Standard Statistical Analysis

In Shannon's terms, it is natural to define mutual information between pairs of events. This is a function that is always less than or equal to the information that one possesses on individual corresponding events:

$$\begin{aligned} H(p_{IJ}; p_I \cdot p_J) &= H(p_I) + H(p_{IJ}) - H(p_{IJ}) \\ &= \sum \{p_{ij} \log_2 (p_{ij}/p_i \cdot p_j) | i \in I, j \in J\} \\ &= \sum \{p_i \cdot p_j (p_{ij}/p_i \cdot p_j) \log_2 (p_{ij}/p_i \cdot p_j) | i \in I, j \in J\} \\ &= \sum \{p_i \cdot p_j f(p_{ij}/p_i \cdot p_j) | i \in I, j \in J\} \end{aligned}$$

Now, in usual statistics, there is another means to evaluate correlation between events, using a specific metric known as the χ^2 *test*. This metric is meant to evaluate the mutual distance (this can be interpreted as some mutual information) between two laws. It is designed in such a way that the mutual distance p_{IJ} is equal to $p_I \cdot p_J$ when all the knowledge about the association of the two laws is the result of the knowledge about each one:

$$\begin{aligned} \| p_{IJ} - p_I \cdot p_J \|^2 &= \sum \{(p_{ij} - p_i \cdot p_j)^2 / p_i \cdot p_j | i \in I, j \in J\} \\ &= \sum \{(p_i \cdot p_j [(p_{ij}/p_i \cdot p_j)^2 - 1] | i \in I, j \in J\} \\ &= \sum \{p_i \cdot p_j f(p_{ij}/p_i \cdot p_j) | i \in I, j \in J\} \end{aligned}$$

knowing by definition of the laws p_I (normalization) that the sum of the product $p_i p_j$ as well as that of the mutual law p_{ij} is equal to 1.

One can then remark that the final expression has a general form similar to the form of the Shannon's mutual information law: One has simply to exchange the form of the distance function f involved in the χ^2 paradigm, where it is $x^2 - 1$, for the form in Shannon's law where it is $x \log_2 x$. Both functions have a very similar behavior around $x_0 = 1$: Their value, their first derivative, and their second derivative are equal. In fact, if one proposes to measure mutual information by various plausible laws (mutual "distances"), one is led to the conclusion that the envelope of all such laws is Shannon's information function. This has the consequence that the function $x \log_2 x$ is the natural function that should be used in most cases. However, this conclusion can also be extended, implying that it is natural to find this same function in many situations otherwise unrelated: *When finding that the representation of a phenomenon yields a function of the form* $x \log_2 x$, *one should not infer that the analogy extends to the* nature *of the phenomena described by this same function.*

Methods using χ^2 are numerous (they can even be used in learning techniques for the partition of training samples), but because it is very powerful

yet rarely used in biology, we shall refer to only one such method, factorial correspondence analysis (FCA), which makes use of χ^2 distances for classification of objects without a priori knowledge of the classes (Hill, 1974). The underlying idea is that one can construct pertinent classes where objects do not obligatorily share simultaneously *all* the characters defining the objects in a given class and nevertheless can be said to belong to the same class because they share most of the characters defining a class (Benzécri, 1984). Such an analysis permits one not only to build up classes but also to build up trees describing the relationships between classes. This permits one, under some conditions, to build up hierarchical clusters, reminiscent of phylogenies.

FCA has been used to study codon usage in individual genomes. In the absence of any other knowledge, coding sequences (CDSs) can be described by the usage of codons specifying each amino acid residue of the polypeptide they encode. Accordingly, each CDS is represented as a point in a 61-dimensional space, each dimension corresponding to the relative frequency of each of the 61 codons (Blake and Hinds, 1984; Alff-Steinberger, 1987; Hanai and Wada, 1989; Bulmer, 1990; Lawrence and Hartl, 1991; Sharp, 1991). The set of CDSs is displayed as a cloud of points in the space of codon frequencies. Using the χ^2 distance between each CDS pair, FCA allows calculation of the two-dimensional projection of the cloud of points yielding maximum scattering (Hill, 1974; Lebart, Morineau, and Warwick, 1984). On such projection, genes that have a similar codon usage will appear as neighbor (but the converse is not necessarily true).

To analyze this graphical representation in terms of individual classes, it is necessary to use a second method that automatically clusters the objects (here, the CDSs) that are close to one another. For example, in a first step, one splits the collection of objects into k groups by a dynamic clustering method; then, in a second step, objects that are always clustered together in the course of the different partition processes are selected (Delorme and Hénaut, 1988). Analysis of *E. coli* CDSs revealed that the hypothesis requiring the minimum assumptions is a clustering into three well-separated classes (Médigue, Rouxel, Vigier, Hénaut, and Danchin, 1991a). The codon bias is very strong in class II, intermediate in class I, and weak in class III. For example, in class II, CTA triplet is used in fewer than 1 percent of all leucine codons, whereas CTG is used in 76 percent of all cases (this corresponds to a major leucine tRNA, $tRNA_i^{leu}$).

The proof that this partition is significant comes from the observation that these three classes of *E. coli* genes can also be clearly distinguished by their biological properties. Class I contains almost all genes of intermediary metabolism, with the noticeable exception of genes involved in the core of carbon assimilation (glycolysis, tricarboxylic acid cycle [TCA] cycle, and fatty acids synthesis). It also contains genes specifying gene regulation (activators and repressors) and genes responsible for DNA metabolism. Thus, class I comprises those genes that maintain a low or intermediary level of expression (Bennetzen and Hall, 1982; Gribskov, Devereux, and Burgess, 1984; McLachlan, Staden, and Boswell, 1984; Sharp and Li, 1986, 1987) but can be

potentially expressed at a very high level (e.g., the lactose operon). In contrast, class II contains genes that are constitutively expressed at a high level during exponential growth. Most of these genes are involved in the translation and transcription machinery, as well as in the folding of proteins. Up to this point, the clustering of *E. coli* genes fits the codon usage clusters described by Gouy and Gautier (1982). It also fits the general gene clustering described by Sharp and coworkers (Sharp and Li, 1986, 1987). In contrast, in class III, the "frequent" codon CTG is used in only 30 percent of all leucine codons. In the same way, the second and first classes show significant bias against a few codons (mainly ATA, AGA, and AGG) that are not discriminated against in class III. In general, the codon usage characterizing the third class is different from that of classes I and II, for the distribution of codons is quite even. It therefore is significant that most of the corresponding genes have been shown to be involved in functions required for horizontal transfer of genes (Médigue, Rouxel, Vigier, Hénaut, and Danchin, 1991a).

With these examples, we observe that investigation at a first level of the information content of a genome is rewarding. In particular, it tells us that the actual selection pressure on the fine evolution of sequences inside genes is actually visible when one possesses a large collection of genes for analysis. Because this corresponds to a fine variation, such as a general bias in nucleotide distribution in genes, and comprises many genes at the same time, it becomes difficult to propose experiments to test whether the interpretation of the observation adequately reflects reality. Indeed, it must be difficult to submit a modified organism to a "normal" life cycle (i.e., a life cycle that differs strongly from laboratory conditions, including chemostats) and to test whether its DNA has evolved according to the predicted consequences of a model. In some cases, however, experiments have been performed that can answer some of the predictions: Cox and Yanofsky (1967), for example, have submitted a culture of *E. coli* lacking a functional *mutT* gene (the product of which prevents AT to GC transitions) for evolution during a large number of generations, and they have observed that, after some time, the overall content of the DNA was significantly more GC-rich than was the parental strain.

In the examples just given, the information content of a sequence was considered mostly as a function of its *local* value. A more appropriate way to consider the same information would be to evaluate it as a function of its context, as we shall now do.

Eukaryotes and Prokaryotes: Algorithmic Complexity

A genome sequence is certainly far from being random. This is due not only to the fact that it actually contains functional information (i.e., information used for the definition of molecules that operate in metabolism and cell construction), but also because the starting DNA molecules were certainly not random. For this reason, given a sequence, it is very difficult to estimate

its actual probability: One has to compare it to an unknown baseline. When studying the frequency of words (such as *TA, CCTTGG,* for instance), one has to estimate the probability of occurrence for them if the sequence were random. However, this does not mean that the real sequence can compare to a purely stochastic sequence of letters of same overall probability of occurrence: A way to overcome this difficulty is to choose regions of the genome, such as coding sequences, where it is known that the major selection pressure has been operating only indirectly on DNA, using this as a means to calculate a baseline. For example, as proposed by Hanai and Wada (1989) or Hénaut, Limaiem, and Vigier (1985), one can study the frequency of words overlapping codons (four or more letters) by comparing the level of occurrence of a given word with its level of occurrence if all the codons used in every CDS were taken into account but their actual succession were random. In this case, most of the bias comes from the style of the genome (which can, in part, be reflected by a nonrandom frequency of dipeptides in proteins [Petrilli, 1993]), resulting in some contribution to the bias in the usage of words). To illustrate, it can be observed that among four-letter palindromic sequences, CTAG is counterselected in all organisms (except those in which *C* is modified, such as bacteriophage T_4 where *C* is replaced by hydroxymethyl cytosine) (Médigue et al., 1991a). This seems to be due to a special conformation of the double-stranded helix when it carries such sequences of motifs, resulting in some weakness (perhaps inducing spontaneous breaks) (Hunter, Langlois d'Estainot, and Kennard, 1989). In all other cases, the word's frequency reflects the specific style of each genome.

Is it possible to go farther and extract significant global properties of genomes? Relationships between objects, such as those that are of fundamental importance in biology, should be taken into consideration. In particular, the very nature of the information carried from generation to generation in the DNA molecule, which constitute genes, is not at all present in the analogy between entropy and Shannon's information. For example "context," "meaning," or "semantics" are, by construction, absent from this extremely primitive model. Biology provides us with a metaphor that displaces the idea of information toward a new field, that of programming and informatics. Is there more insight in these new fields than in the "natural" way of considering information ? At least since 1965, Kolmogorov and Chaitin, following Solomonoff in the USA and the Russian school of electronic engineers, have formulated the problem in details in the following way (for a general description, see Yockey, 1992). Let us consider the simple case of a chain of characters such as those found in computer sciences. What can be said about that chain's information content? A way to consider the chain is to try to reduce its length so that it can be accommodated in a memory—for example, using the minimum space—without altering its performance when used in a program, for example (at least in terms of accuracy, if not in terms of time)—in short, without losing its information content. This is called *compressing* the data. This problem is of very broad and general interest. Because a chain of

characters can be identified with an integer, the universal Turing's computation rules apply. It is particularly possible, given a chain, to define the shortest formal program (in terms of Turing's universal computation algorithms—that is, algorithms that can be implemented on any machine operating on integers) that can compress an original chain, or restore it given its compression state. The information value of a chain S is therefore defined in this model as the minimal length of the universal program that can represent S in a compressed form.

With this definition, it appears that a completely random chain S cannot be compressed, implying that the minimal program required to compress S is identical to S. It should be noticed here that an *apparently* random chain can be compressible: This is the case for instance, with the chain formed by the digits of the decimal writing of the number π, which can be generated by a short program. In contrast, a chain made of repetitive elements can be summarized as a very short program: the sequence of the element and the instruction "repeat." In this context, the information of a sequence is defined as the measure of its compressibility (one often uses the concept of complexity rather than information in this context). It is, however, apparent here that this is not all of what can be said, should we wish to define information in a finer sense. Indeed, the program defining π is highly compressible, as is the program generating the sequence 01010101..., but we would like to say that the information embedded in π is richer than the one embedded in 01010101....

Represented as sequences of letters, genes and chromosomes have an intermediate information (complexity) content: One finds local repetitions or, in contrast, sequences that are impossible to predict locally (Lebbe and Vignes, 1993). Their complexity is intermediary between randomness and repetition. The apparent complexity of sequences originating from higher eukaryotes or prokaryotes is very different, and this links genomic sequences to both sides of the seemingly uninteresting fraction of information (because it corresponds on the eukaryotic side to repetition and on the prokaryotic side to randomness). The complexity of the former is more repetitive and usually looks much lower than that of the latter, which appears more random (Graniero-Porati and Porati, 1988). This is quite understandable if one remembers that bacterial or viral genomes are submitted to stringent economy constraints, implying that they must remain very compact. In contrast, genomes of higher eukaryotes are extremely loosely constrained, and they contain, for instance, many repetitive sequences.

Nonetheless, would it not be more natural to say that eukaryotes are more complex than prokaryotes? We can propose here that this difference in the form of algorithmic complexity is perhaps the actual *cause* of the main morphological differences between prokaryotes and differentiated organisms. The former, following their strategy for occupying biotopes owing to a rapid and versatile adaptive power, are forced to keep up with a small genome.

This implies that several levels of meaning must often be superimposed in a given sequence. The apparent result of this constraint is that their actual algorithmic complexity looks more like that of a random sequence. However, this constraint has an important physical consequence (because, after all, DNA is not only an abstract sequence of symbols but also a molecule that must occupy space): *Superposing signals on the same DNA stretch generally precludes combining synchronous recognition processes by specific proteins, for it is not possible to place two different objects at the same time at the same location.* To exclude such combinations (which alone can fulfil the logical principle of third-party exclusion), however, prevents investigating the refined principles of a delicate modulation of gene expression. The exactly opposite situation is found in differentiated organisms such as animals (and perhaps in the regions of prokaryotic genomes displaying a low complexity). *The lack of limitation in the length of the DNA text permits juxtaposition of recognition signals by regulatory proteins* and allows exploration of the properties of their association according to the rules of combinatorial logic. This is probably what permits differentiated expression as observed in the various cell types. Indeed, this is already well understood for the regions controlling transcription (promoters, enhancers, etc.), but this is probably also the reason for the multiplication of introns in higher eukaryotes, as they behave as inserts without apparent signification, in the very coding regions of genes. Thus, analysis of information in terms of algorithmic complexity permits us to propose explanations to account for the differences among major organism types, whereas Shannon's information gives a meaning only to conservation of a DNA sequence, with a low level of errors, from generation to generation, through the replication machinery.

Many analyses of genomes indirectly pertain to analysis of their algorithmic complexity, thus providing an interesting view of the general organization of base sequences in chromosomes, but such analyses consider genomes as fixed, synchronous entities. In general terms, higher eukaryotes seem to be much more complex than bacteria, be it only because they are able to exhibit time-dependent differentiation patterns. It appears, therefore, that something more should be said about their information content: Information cannot be identified simply by algorithmic complexity, so that Chaitin and Kolmogorov's information (see Yockey, 1992) represents only part of what interests us. In the examples just provided, the concept of information did not imply the notion of *value* (depending both on the object considered and on the nature of the problem that has to involve the object). In fact, if one considers a particular piece of information, its actual availability must have some *value*. Information that is within everyone's reach must differ from that which is difficult or slow to obtain. In the case of living organisms, knowledge of the paths that have been followed through evolution and have led to organisms as we know them today is very valuable and difficult to obtain. Indeed, having access to this type of information may require the introduction of *time*, as we shall now consider.

Evolution: Bennett's Logical Depth

We have seen that we would like to have an idea not only of the complexity of a sequence (as evaluated using the previously described approaches) but also of the availability of the information contained in the sequence. This is particularly relevant in the case of DNA sequences, because all that can be inferred from the knowledge of a sequence derives from the accessibility of its actual information content. To see more repetitions in eukaryotes and more apparent randomness in prokaryotes does not account for the paths that have led to such a large overall difference in genomes. We wish to know the *usefulness* of the information in the sequence. A very short recursive program can be written to express the digits of π and it is possible to write a program of the same length but generating a repetitive sequence. However, the availability of the information content of the corresponding programs' outcome is very different. Although it is easy to predict, after a very short computation, the value of the n^{th} digit of the repetitive sequence, the same is not true for π. In the latter case, if n is large enough, a long computation time is required until the digit is generated, using the program permitting calculation of π.

This observation is essential when one considers sequences generated by recursive programs. In many cases, the writing of the program can be short because the procedure that is reiterated is nested inside the program. Therefore, the information content of the sequence, in terms of algorithmic complexity, is small, because it can be summarized as a short program. Nonetheless, the information on a given digit could be extremely tedious to obtain and might be obtained only after a very long computation time. In this respect, the knowledge of the precise value of this digit—the *information* on it—would seem to be very high, had we another quick means to obtain it. Bennett (1988) has formalized this aspect of sequences in terms of logical depth. The *logical depth* of a sequence S is the minimum time required for generating S using a universal Turing machine (Bennett, 1988). Obviously, knowing a sequence, it is not possible to know its algorithmic complexity. It is even less obvious to have an idea of its logical depth. For this reason, the conceptual analysis made earlier might appear to be merely philosophical (and philosophical is a bad word in English, though not in Latin or Greek civilizations!). We can propose efficient means to analyze both the algorithmic complexity and the logical depth, by proposing algorithms meant to generate the sequence being analyzed. For this we must propose educated guesses, and we contend that, starting from an analysis of the underlying biological processes, we can indeed propose constructive hypotheses. For example, trying to represent how RNA polymerase recognizes its promoter will permit us to generate an algorithm that, in turn, allows us to identify promoters from sequences. Cases in point are the algorithms proposed by d'Aubenton, Carafa, Brody, and Thermes (1990) for the recognition of transcription terminators or by Lisacek, Dinz, and Michel (1994) for the recognition of group I introns. The same is true in the case of logical depth: If this aspect of informa-

tion carried by DNA is relevant, then the algorithms needed for analysis should require significant computing time to produce interesting outputs. This is particularly true in the case of phylogenetic analyses and indicates that algorithms that perform too fast are necessarily flawed with respect to the biological relevance of their output. A positive side of this very negative observation is that time adds a new dimension to sequence analyses: It could be interesting to add to the features permitting comparisons between sequences the time required for the computation algorithms performed on them, because similar sequences ought to perform with similar computation times under a given hardware and software environment.

Logical depth relates to the Aristotelian distinction between potentiality and reality. It rightly suggests that one should not be allowed to speak in terms of potentiality as if in terms of reality. *There are cases when to speak of potentiality is nonsense, because the actualization of the potential, in true reality, is impossible in a reasonable period of time.* Only those facts that are mechanistically generated could permit such a gross identification between potentiality and future reality. It must be stressed here that reflection on this aspect of information does not completely exhaust the nonclarified (natural) concept of information. In particular, it rests on the existence of an abstract, but perfect, universal Turing machine. In the case of living organisms—provided that one can identify their behavior as that of Turing machines (see Danchin, in press, for justification of this hypothesis)—the building up of the machine is the result of the actualization of a finite program present in the machine. Here one should take time into account once again and consider only those algorithms the length of which is smaller than the lifetime of the machine as a given, unaltered structure. Indeed, a time limit defining the algorithms permitting construction of the machine must also be taken into account: This limit defines a *critical depth*, likely to be of major importance in the case of life. We shall limit ourselves to these descriptions, however, to consider now the specific case of living organisms.

GENOMES *IN SILICO*: INFORMATION AND HEREDITY

The argument that has just been developed is meant to be a substratum for reflection on the metaphor of *program* that currently is used in molecular genetics. It is known that a giant molecule, DNA, made of the sequential arrangement of four—and only four—types of related molecular species, specifies entirely the structure and dynamics of any living organism (adding its own constraints to a world where physics and chemistry operate, permitting in particular the generation of compartmentalization and metabolism). We shall not consider here the problem of origins of such organisms (Danchin 1990), but shall try only to investigate the nature of the relationship between the DNA sequence and the building up of an organism.

As in the case of the reflection of Turing in number theory (Hofestadter, 1979), one can surmise that there must exist a machine to interpret the

content of a DNA sequence. This machine is a preexisting living cell, made of a large but finite number of individual components organized in a highly ordered way. Given such a machine, a fragment of DNA, provided it contains a well-formed sequence, is necessary and sufficient to produce a specific behavior representing part or all of the structure and dynamics of an organism. The appropriateness of the alphabetical metaphor used to describe the DNA sequence is further demonstrated by the fact that it is possible to manipulate on paper four symbols—*A, T, G, C*—and to organize them in such a way that, when interpreted into chemical terms (i.e., linked into the proper sequence arrangement into an artificial DNA molecule), they produce a behavior having at least some of the properties of real natural biological entities. This is the basis of survival and multiplication of viruses as well as of a success story in biotechnology: It is possible to make bacteria synthesize a human growth hormone that has been demonstrated on children to be as active as the natural product. That the machine is, in a way, independent of DNA has also been further indicated by the fact that it is possible in some organisms (amphibians) to replace the nucleus of a cell by a new nucleus, resulting in the construction of a new organism (i.e., both a new DNA molecule and a new machine).

However, DNA alone is absolutely unable to develop into a living organism: This is why the question of origin must be considered if one wishes to resolve this chicken-and-egg paradox, through the coevolution of the machine and of the program. It must be stressed here that to apply the consequences of Turing's arguments, it is, in principle, necessary that the program and data be well-identified and separate entities. As stated earlier, biotechnology uses this fact to permit synthesis of human proteins in bacteria, but this corresponds to small genome fragments. Viruses, which infect cells, use bacteria in a similar way to reproduce. Finally, a recent discovery seems to indicate that, at least as a first approximation, bacteria behave as if separating data (the cell machinery, excluding DNA) and program (the chromosome): We have found that one-fifth of the *E. coli* chromosome is made of DNA on which selection operates in such a way that it favors and sustains horizontal exchanges not only between bacteria of the same species but also between distantly related organisms (class III genes) (Médigue et al., 1991a).

Now, if one considers the role of a DNA molecule in a cell, it is possible to describe the various functions derived from its presence as calculable in an algorithmic way. Many routines in the process are recursive (i.e., call themselves but must terminate). It is a generative process that permits generation of both a DNA molecule and of the machine and data necessary to interpret it. Although this has often been called *self-reproduction,* it is production of an organism that is similar if not identical to the starting organism. One should speak, therefore, of *self-reference.* A living organism is, in this context, the algorithmic actualization of a program, the initiation of which has been obtained through a selective process, keeping only actualizations that are stable in time. It should be noticed here that this means that selection does

not operate on the phenotype (data) nor on the genotype (program) but on the unfolding of the algorithm (i.e., on the organism itself), thus placing in a new light the usual paradox facing all selective theories and standard Darwinism in particular.

In this respect, it is possible to discuss the information content of a DNA sequence or its complexity using the various definitions previously given. At first sight, most, if not all, of the sequence must be specified to generate the daughter sequence. Things would be simple, and the corresponding information—that which is required for reproduction—would even be rather poor, if there were an absolute fidelity in replicating the original. However, the rule of replication (A → T, T → A, G → C, C → G) is not absolutely faithful, even if it can be considered to result from an algorithmic process. Variations are embedded in the replication process as a function of the general environment of the organism as well as a function of the local sequence of the DNA segment that is replicated. This corresponds to information that should be taken into account, and that is very important indeed, as we shall see later.

However, it must be emphasized at this point that the existence of error management results in specific features of the DNA sequence. As already stressed, CTAG sequences are generally rare, and in *E. coli* this is compensated by an excess in CTGG words (Médigue et al., 1991a). Spontaneous deamination of cytosines is likely to be compensated by a bias in the fidelity of replication, tending to create more GC pairs. In turn, this mechanism can result in an overshoot, as seen in *E. coli* strains that lack the *mutT* gene. Many other processes are certainly modeling the overall GC content of genomes, resulting in a genome style (in this case G + C content) that has been taken into account by taxonomists for preliminary classification of organisms (see Bergey, 1984, for bacteria). Still, such evolution can be much subtler. Chi sites are oriented along the replication fork direction in *E. coli* (Médigue, Viari, Hénaut, and Danchin, 1993), and it is likely that, because replication is semiconservative, there exists some bias in regions where Okasaki fragments are initiated (this should result in biases having a periodicity of the order of a few kilobases). Another bias should be introduced by the restriction modification systems present in bacteria, as well as by the existence of methylated sites (as have been observed in a few cases). Finally, there should exist biases in the DNA composition along genes corresponding to the fine tuning of gene expression as a function of environmental conditions or of location of the transcribed or translated regions in the genes. Limaiem and Hénaut (1984) have observed such biases as the enrichment in C + G of the third codon letter as a function of the position of the codon in the genes' sequence. Despite their evidence, these observations are difficult to submit to experimental falsification because this would require, for example, construction of artificial genes and study of their evolution for a very long period of time in conditions that may be very different from natural conditions (chemostats). Therefore, it becomes important to make predictive tests using computers, but to what extent can we make predictions?

Gödel has demonstrated that recursive processes have the very interesting property of being open, in such a way that there exist propositions in natural number theory that can be generated from a corpus of axioms and definitions and that are not decidable in the original corpus (Hofestadter, 1979). Gödel's demonstration is ingenious; it uses a coding process that is certainly reminiscent of the coding process known as the *genetic code*. However, this demonstration has a specific feature: The aim of the demonstration, permitting Gödel to generate his famous "sentences," is known beforehand. The situation faced by living organisms is different: They cannot predict the future; they can have no aim except for (re)producing themselves. Our hypothesis here will be that *the variation provided in the replication and recombination machinery, permitting the generation of new DNA sequences, is the very process that permits living organisms to generate the equivalent of Gödel's sentences in terms of new, deterministic but utterly unpredictable organisms*. The process involves both a procedure for generating variants and a procedure that permits accumulation of "adapted" variants. The latter is so-called natural selection, which must be understood as a process screening stability (and *not* fitness) of the surviving organisms. According to this view, the Spencerian "survival of the fittest" is therefore but a surviving variant of the "instructive" Lamarckian thought.

If this is accepted, one is faced with the puzzling observation that to evaluate the actual information content of a DNA molecule, its history must be traced, which requires an extremely slow procedure: The information content of a DNA sequence is so deep (in Bennett's terms) that it would require the age of life on Earth to evaluate its true complexity. Bearing this in mind, we should be very modest when trying to predict the future of living organisms. In addition, one observes that the process of natural selection associated with generation of mutations provides a blind procedure for generating the equivalent of Gödel's sentences. (In fact, Gödel was cheating somewhat, as we intimated earlier, because he knew what he wanted to obtain, which living organisms cannot do.) This process also generates many more sentences of a type the existence of which we are unable to predict (but the existence of which we can conjecture, simply because at least one type—Gödel's type—does indeed exist). This corresponds to the truly *creative* phenomena that are derived from living organisms.

CONCLUSION

Genome analysis will yield major insights into the chemical definition of the nucleic acids and proteins involved in the construction of a living organism. Further insight comes from chemical definition of the small molecules that are the building blocks of organisms, through the generation of intermediary metabolism. Because life requires also in its definition the processes of metabolism and compartmentalization, it is important to relate intermediary metabolism to genome structure, function and evolution. This requires elabo-

ration of systems for constructing actual metabolic pathways (see chapter 11) and, when possible, dynamic modeling of metabolism (see chapter 6) and for correlating pathways to genes and gene expression (as discussed at the experimental level by Neidhardt [see chapter 9] and theoretically by Thomas [see chapter 8]). In fact, most of the corresponding work cannot be of much use because the data on which it rests have been collected from extremely heterogeneous sources and most often are obtained by in vitro studies.

The initial need is to collect, organize, and actualize the existing data. For the data to be effective and lasting, data collection should proceed through the creation of specialized databases (see Karp and Riley, 1994). To manage the flood of data issued from the programs aiming at sequencing whole genomes, specialized databases have been developed (Médigue et al., 1993; Rudd, 1993), that make it possible not only to bypass meticulous and time-consuming literature searches but also to organize data into consistent patterns through the use of appropriate procedures aimed at illustrating collective properties of genes or sequences. In addition to sequence databases, it has become important to create databases wherein the knowledge progressively acquired on intermediary metabolism could be symbolized, organized, and made available for interrogation according to multiple criteria (Rouxel, Danchin, and Hénaut, 1993).

Using organized data, it will become possible to make *in silico* assays of plausible pathways or regulation before making the actual test in vivo. Well-constructed databases will also permit investigation of properties of life that go back to life's origin, placing biology in a situation not unlike that of cosmology.

NOTES

1. This software, though still unpublished, is available through Internet; mail to viari@radium.jussieu.fr.

SUGGESTED READING

Benzécri, J. P. (Ed.). (1984). *L'analyse des données: I. La taxinomie; II. Les Correspondances* [in French; there is no equivalent in English]. Paris: Dunod.

Cover, T. M., and Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Hénaut, A., and Danchin, A. (in press). Analysis and predictions for *Escherichia coli* sequences, or *E. coli in silico*. In F. C. Neidhardt, J. L. Ingraham, K. Brooks Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (Eds.), Escherichia coli and Salmonella typhimurium (Vol. 2) (2nd ed.). Cellular and molecular biology: Washington, DC: American Society for Microbiology.

Yockey, H. P. (1992). *Information theory and molecular biology*. Cambridge: Cambridge University Press.