

Cambridge University Press

0521802938 - Information Theory, Evolution, and the Origin of Life

Hubert P. Yockey

Excerpt

[More information](#)

# 1

## The genetic information system

*Socrates: Every sort of confusion like these is to be found in our minds; and it is this weakness in our nature that is exploited, with a quite magical effect, by many tricks of illusion, like scene-painting and conjuring.*

Glaucon: *True.*

*Socrates: But satisfactory means have been found for dispelling these illusions by measuring, counting and weighing. We are no longer at the mercy of apparent differences of size and quantity and weight; the faculty which has done the counting, measuring or weighing takes control instead. And this can only be the work of the calculating or reasoning element in the soul.*

*The Republic*, Book X, Plato (428–348 B.C.), translated by Francis M. Cornford, Oxford University Press.

### 1.1 Expressing knowledge in numbers

Socrates (*The Republic*, Book VI, p.745) had noted, in an earlier conversation with Glaucon, that students of geometry and reckoning first set up postulates appropriate to each branch of science, treating them as known absolute assumptions, taking it for granted that they are obvious to everybody. Thus, as Socrates taught us, the essence of science is measuring, counting, and weighing together with reasoning from postulates or axioms. This breaks molecular biology out of sophisticated *Just So Stories* (Kipling, 1902) into the quantitative mode, used by natural scientists (Gamow, 1954; Wolynes, 1998).

Hermann Rorschach (1884–1922), a Swiss psychiatrist, analyzed the interpretations, by his subjects, of ten standard inkblots to probe their thoughts. There is a danger that we may be looking at the shapes of Rorschach inkblots, so to speak, and seeing what we want to see when science attempts to proceed from qualitative arguments. The more discussions can be made quantitative and avoid *ad hoc* explanations, the better our understanding of biology is served. There are no “other ways of knowing” in science. *The absence of evidence is evidence of absence.*

The laws of physics and chemistry are much like the rules of a game such as football. The referees see to it that these laws are obeyed but that does not predict the winner of the Super Bowl. There is not enough information

in the rules of the game to make that prediction. That is why we play the game. Chaitin (1985, 1987a) has examined the information content of the laws of physics by actually programming them. He finds the information content amazingly small.

The reason that there are principles of biology that cannot be derived from the laws of physics and chemistry lies simply in the fact that the genetic information content of the genome for constructing even the simplest organisms is much larger than the information content of these laws (Yockey, 1992).

### *1.1.1 The definition of life and Louis Pasteur*

Mr. Justice Potter Stewart (1915–85, U.S. Supreme Court) said he couldn't define pornography, but he knew it when he saw it. It is often said that no broadly accepted definition of life exists. Like Edgar Allan Poe's (1809–48) *The Purloined Letter*, the definition of life has been in plain sight since 1848. One of Louis Pasteur's (1822–95) more important discoveries, relevant to the nature and origin of life, is that ammonium tartrate tetrahydrate when made from grapes has only the left-handed molecules, Pasteur (1848, 1922). When examined in a polarimeter, they are found to rotate the plane of polarization of light to the left. Ammonium tartrate tetrahydrate made synthetically is racemic, that is, composed of equal numbers of right-handed and left-handed molecules. The human hand is chiral. Each hand is the mirror image of the other. Neither can be superimposed on the other.

Pasteur carefully selected the two kinds of crystals, called optical isomers, and found that each rotated the plane of polarization in opposite directions, one left and the other right. He prepared a synthetic ammonium tartrate tetrahydrate solution and contaminated it with a mold. The solution became more optically active with time. It followed that the mold was using only the left-handed ammonium tartrate molecules. What a delicate appetite that mold had! This achievement of Pasteur is the first demonstration of chiral molecules as an essential and unique element in biology. It can serve as a definition of life, as any substance composed of only one optical isomer must have come from life (Section 8.1.3).

An additional criterion for this book is:

The existence of a genome and the genetic code divides living organisms from nonliving matter. There is nothing in the physico-chemical world that remotely resembles reactions being determined by a sequence and codes between sequences.

### *1.1.2 The work of Gregor Mendel (1822–84) leading to molecular biology and genetics*

In the nineteenth century, intuition led many to believe that the inheritance of characteristics, such as tall and short, would yield a blending of these traits and produce plants of medium height (Jenkin, 1867). The theory of blending inheritance predicts either the disappearance of favored traits or that mutations must be several thousand times as frequent as they are known to be (Fisher, 1930). Therefore, Jenkin concluded that this was evidence that Darwinian evolution would never occur.

However, the Gregor Mendel's experiments with strains of pea (Mendel, 1865) proved that inheritance is *segregated* and does not blend. This was the first step to the molecular biology and genetics we have today. The structure of DNA found by Watson (1928–) and Crick (1916–2004) could have been just that of another large molecule, such as hemoglobin, if it had not been that DNA carries the genetic message that is transferred to the proteome by the genetic code. Their work completed the modern view that the message in the genetic information system is *segregated, linear, and digital*. Watson and Crick's finding is just as much a new axiom in science, as Max Planck's discovery that Newton's particles of light are electromagnetic wave packets and the frequency,  $\nu$ , is related to the energy,  $E$ , by  $E = h\nu$  where  $h$  is Planck's constant.

The genetical information system, because it is *segregated, linear, and digital*, resembles the algorithmic language by which a computer completes its logical operation (C. H. Bennett, 1973; Chaitin, 1979). Computer users are well aware that the amount of information in a sequence or a message can be measured without regard for its meaning. A computer user buying a floppy disk or a hard drive does not expect it to hold either more or less information depending on whether it will be used to store children's drawings or translations of the plays of Sophocles. Information theory and coding theory and their tools of measuring the information in the sequences of the genome and the proteome are essential to understanding the crucial questions of the nature and the origin of life.

### *1.1.3 Evolution and the sequencing of DNA*

As all the living forms of life are the lineal descendants of those which lived long before the Cambrian epoch, we may feel certain that the ordinary progression by generation has never once been broken and no cataclysm has devastated the world. "... from so

simple a beginning endless forms most beautiful and most wonderful have been, and are being evolved. (Darwin, 1872, Ch. XV)

The recent accomplishments in the sequencing of the DNA of the human genome as well as those of a number of other organisms establishes the remarks of Darwin beyond question. Darwin was concerned with “missing links” and based much of his “one long argument” (*Origin of Species*, 1872 edition, Ch. XV) on comparative morphology. The old arguments based on “missing links” proposed in opposition of Darwin’s theory are no longer relevant. For that reason, foolish discussions on either side of the debate on Darwinism about how the giraffe got his long neck are no longer pertinent. Although the details may be unknowable, there is indeed a phylogenetic evolutionary message or signal from which all organisms have branched (Woese, 1998, 2000, 2002) (see Section 11.2.3). The e-mail one sends to colleagues traces its way through the Internet from source to destination. By the same token, the “code-script,” as Schrödinger (1992) called it, unites all living things on Earth.

The transmission of genetic messages for more than 3.85 billion years since the origin of life (Mojzsis et al. 1999; Woese, 2000), with modification and diversification by evolution, could have been done *only* because the message in the genome is *segregated, linear, and digital* (Chapter 12). It is impossible to remove the effect of noise in analog signals. Early analog records of the glorious voice of Enrico Caruso (1873–1921) do not compare with the modern digital recordings of the Three Tenors: Plácido Domingo, José Carreras, and Luciano Pavarotti. Shannon’s Channel Capacity Theorem (Shannon, 1948) showed how to eliminate the effect of noise as much as we wish by digitizing the signal. The digital revolution has now provided digital television eliminating noise almost to the theoretical limit. Even cameras are now digital.

Evolution would be quite impossible if inheritance were by analog means. Nevertheless, distinguished biologists Szathmáry and Maynard Smith (1997) wrote:

To explain the origin of life, we need to explain the origin of heredity in terms of chemistry.

Morowitz et al. (2000) wrote:

A small number of selection rules generates a very constrained subset, suggesting that this is the type of reaction model that will prove useful in the study of biogenesis. The model indicates that the metabolism shown in the universal chart of pathways may be

Cambridge University Press

0521802938 - Information Theory, Evolution, and the Origin of Life

Hubert P. Yockey

Excerpt

[More information](#)*The contributions of Niels Bohr*

5

central to the origin of life, is emergent from organic chemistry, and may be unique.

## 2.1 The contributions of Niels Bohr

Niels Bohr (1885–1962) proposed that life is *consistent* with but *undecidable* or *unknowable* by human reasoning from physics and chemistry. Bohr (1933) made this point in his famous “Light and Life” lecture:

The recognition of the essential importance of fundamentally atomistic features in the function of living organisms is by no means sufficient, however, for a comprehensive explanation of biological phenomena, before we can reach an understanding of life on the basis of physical experience. Thus, we should doubtless kill an animal if we tried to carry the investigation of its organs so far that we could describe the role played by single atoms in vital functions. In every experiment on living organisms, there must remain an uncertainty as regards the physical conditions to which they are subjected, and the idea suggests itself that the minimal freedom we must allow the organism in this respect is just large enough to permit it, so to say, to hide its ultimate secrets from us.

It may seem strange that the numerous biological compounds in all living things, from amoeba to man, are constructed from the same twenty (or twenty-two) amino acids. The twenty-six letters of the English alphabet are enough to form all the plays of Shakespeare. The eighty-eight keys of the piano are enough for the piano concertos of Beethoven. The segregated, linear, and digital character of the genetic message is an elementary fact. Therefore, it answers the question: “What is Life” (Yockey, 1977b, 1992, 2000, 2002). There is an abyss between living organisms and inanimate matter. As Ernst Mayr (1982) put it:

One of the properties of the genetic program is that it can supervise its own precise replication and that of other living systems such as organelles, cells and whole organisms. There is nothing exactly equivalent in inorganic nature.

The belief of mechanist-reductionists that the chemical processes in living matter do not differ in principle from those in dead matter is incorrect. There is no trace of messages determining the results of chemical reactions in inanimate matter. If genetical processes were just complicated biochemistry, the laws of mass action and thermodynamics would govern the placement of amino acids in the protein sequences.

## 2.2 Information as the central concept in molecular biology

Information, transcription, translation, code, redundancy, synonymous, messenger, editing, and proofreading are all appropriate terms in biology. They take their meaning from information theory (Shannon, 1948) and are not synonyms, metaphors, or analogies.

The genome is sometimes called a “blueprint” by people who have never seen a blueprint. Blueprints, no longer used, were two-dimensional, a poor metaphor indeed, for the linear and digital sequence of nucleotides in the genome. The linear structure of DNA and mRNA is often referred to as a *template*. A template is two-dimensional, it is not subject to mutations, nor can it reproduce itself. This is a poor metaphor as anyone who has used a jigsaw will be aware. One must be careful not to make a play on words.

### 1.2.1 Information, knowledge, and meaning

The messages conveyed by sequences of symbols sent through a communication system generally have meaning (otherwise, why are we sending them?). It often is overlooked that the meaning of a sequence of letters, if any, is arbitrary. It is determined by the natural language and is not a property of the letters or their arrangement. For example, the English word “hell” means “bright” in German, “fern” means “far,” “gift” means “poison,” “bald” means “soon,” “Boot” means “boat,” and “singe” means “sing.” In French “pain” means “bread,” “ballot” means a “bundle,” “coin” means a “corner or a wedge,” “chair” means “flesh,” “cent” means “hundred,” “son” means “his,” “tire” means a “pull,” and “ton” means “your.” In French, the English word “main” means “hand,” “sale” means “dirty.” French-speaking visitors to English-speaking countries will be astonished at department stores having a “Sale” and especially if it is the “Main Sale.” This confusion of meaning goes as far as sentences. For example, “O singe fort” has no meaning in English, although each is an English word, yet in German it means “O sing on,” and in French it means “O strong monkey.”

*Meaning according to Humpty Dumpty.* When I use a word, Humpty Dumpty said in a rather scornful tone, it means just what I choose it to mean – neither more or less. “The question is,” said Alice, “whether you can make word mean so many different things” “The question is,” said Humpty Dumpty, “which is to be master – that’s all.”

Cambridge University Press

0521802938 - Information Theory, Evolution, and the Origin of Life

Hubert P. Yockey

Excerpt

[More information](#)*Information as the central concept in molecular biology*

7

Alice was too much puzzled to say anything, so after a minute Humpty Dumpty began again. “They’ve a temper, some of them – particularly verbs, they’re the proudest – adjectives you can do anything with, but not verbs – however, I can manage the whole of them! Impenetrability! That’s what I say!

Would you tell me, please said Alice “what that means”?

“Now you talk like a reasonable child,” said Humpty Dumpty, looking very much pleased. “I meant by impenetrability that we have had enough of that subject, and it would be just as well if you’d mention what you mean to do next, as I suppose you don’t mean to stop here all the rest of your life.”

“That’s a great deal to make one word mean,” said Alice in a thoughtful tone.

“When I make a word do a lot of work like that,” said Humpty Dumpty, “I always pay it extra.”

“Oh!” said Alice. She was too much puzzled to make any other remark.

From *Through the Looking Glass*, by Lewis Carroll (1832–98), aka Reverend Charles Lutwidge Dodgson.

\*\*\*\*\*

Similarly, the sequences of nucleotides or amino acids that carry a genetic message have explicit specificity. (Otherwise how does the organism live?) Now, in this book, the term *information* does not mean *knowledge*, although a message composed of a sequence of symbols may transfer knowledge to the receiver of the message.

The genetic information system operates without regard for the significance or meaning of the message, because it must be capable of handling *all* genetic messages of *all* organisms, extinct and living, as well as those not yet evolved. It does not have to be “about something.”

The genetic information system is the software of life and, like the symbols in a computer, it is purely symbolic and independent of its environment. Of course, the genetic message, when expressed as a sequence of symbols, is nonmaterial but must be recorded in matter or energy. We could, in principle, send the genome of a mosquito to our little green friends on an Earth-like planet somewhere in the Milky Way Galaxy.

## 2

## James Watson, Francis Crick, George Gamow, and the genetic code

*The evidence presented supports the belief that a nucleic acid of the desoxyribose type is the fundamental unit of the transforming principle of Pneumococcus Type III.*

Avery et al. *Journal Experimental Medicine* 79, 137–159 (1943)

*The phosphate-sugar backbone of our model is completely regular, but any sequence of the pairs of bases may fit into the structure. It follows that in a long molecule many different permutations are possible, and it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information. If the actual order of the bases on one of the pair of chains were given, one could write down the exact order on the other one. Thus one chain is, as it were, the complement of the other, and it is this feature which suggests how the deoxyribonucleic acid might duplicate itself.*

Watson and Crick (1953b, pp. 964–5)

*In a communication in Nature of May 30, p 964, J. D. Watson and F. H. C. Crick showed that the molecule of deoxyribonucleic acid, which can be considered as a chromosome fibre, consists of two parallel chains formed by only four different kinds of nucleotides. These are either (1) adenine, or (2) thymine, or (3) guanine, or (4) cytosine with sugar and phosphate molecules attached to them. Thus the hereditary properties of any given organism could be characterized by a long number written in a four-digital system. On the other hand, the enzymes (proteins), the composition of which must be completely determined by the deoxyribonucleic acid molecule, are long peptide chains formed by about twenty different kinds of amino-acids, and can be considered as long 'words' based on a 20-letter alphabet. Thus the question arises about the way in which four-digital numbers can be translated into such 'words'.*

G. Gamow, *Nature*, 173, 318 (1954a)

### 2.1 Watson and Crick's proposal of the role of the sequences of DNA in genetics

Those readers of this book who are computer-oriented will easily understand that the chemistry of life is controlled by digital sequences recorded in DNA, as Gamow (1954a) was the first to realize. Life is guided by information and inorganic processes are not. The publicity after fifty years still dwells on the “double helix” and biochemistry, whereas the important discovery is that the life message is digital, linear, and segregated.

The discovery by Avery and his laboratory that a nucleic acid is the fundamental carrier of the hereditary properties of life set the stage for finding the structure of DNA.

Cambridge University Press

0521802938 - Information Theory, Evolution, and the Origin of Life

Hubert P. Yockey

Excerpt

[More information](#)*Watson and Crick's proposal of DNA in genetics*

9

Watson and Crick (1953a) began their paper stating that the DNA structure published by Linus Pauling (Pauling and Corey, 1953) was wrong and proposed their own. This was incredible *chutzpah* for these two young scientists who were unknown at the time. Linus C. Pauling (1901–94) is the only person to have been awarded two unshared Nobel prizes: Chemistry (1954) and Peace (1961).

The importance of deoxyribosenucleic acid (DNA) within living cells is undisputed. It is found in all dividing cells, largely if not entirely in the nucleus, where it is an essential constituent of the chromosomes. Many lines of evidence indicate that it is the carrier of a part of (if not all) the genetic specificity of the chromosomes and thus of the gene itself.

The phosphate-sugar backbone of our model is completely regular, but any sequence of the pairs of bases can fit into the structure. It follows that in a long molecule many different permutations are, possible, and it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information. (Watson and Crick, 1953b).

*Rosalind Franklin, the dark lady of DNA.* The reader may have assumed that science is carried out by selfless academics eager to give credit to all their predecessors and colleagues. That is hardly the case. Watson and Crick (1953a) cited Dr. Rosalind Franklin in a footnote of their seminal paper on the structure of DNA.

We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their coworkers at King's College, London.

When Watson (1968, 2001) referred to Dr. Franklin as “Rosy,” “as we all called her from a distance,” it was hardly a term of respect. Dr. M. H. F. Wilkins, who would share the Nobel Prize in 1962, mentioned Dr. Franklin in a footnote along with several others (Wilkins, Stokes, and Wilson, 1953). Much later, and after he had been awarded the Nobel Prize in chemistry, Watson (1968) confessed that: “Rosy, of course, did not directly give us her data. For that matter, no one at King's realized they were in our hands.” But, in their first paper (Watson and Crick, 1953), they mentioned that Pauling and Corey (1953a, 1953b) had “. . . kindly made their manuscript available

Cambridge University Press

0521802938 - Information Theory, Evolution, and the Origin of Life

Hubert P. Yockey

Excerpt

[More information](#)

*James Watson, Francis Crick, George Gamow, and genetic code* 10

to us in advance of publication.” They did not consider it necessary to return this courtesy or to acknowledge also that they had Dr. Franklin’s data before publication.

The success of their seminal paper (Watson and Crick 1953) depended critically on their possession of the Pauling and Corey paper and on Dr. Franklin’s data. Possession of both provided an opportunity to put their paper in final form and remove any mistakes. Dr. Franklin should have been a third author although she would not have been awarded the Nobel Prize in 1962. The Dark Lady of DNA died in April 1958 of ovarian cancer (Elkin, 2003; Maddox, 2002). The Nobel Prize is not awarded posthumously.

Gamow’s suggestion is that life is more than complicated chemistry and that the digital information in DNA sequences is sent to the digital information in the proteome by means of a code. It was not until 2001 that Watson acknowledged the essential contribution made by George Gamow (1904–68).

### 2.1.1 *George Gamow and his proposal of the genetic code*

Gamow, promptly upon reading the paper of Watson and Crick (1953b), wrote to them proposing that the sequences of nucleotides of DNA were mapped onto the sequences of the amino acids in protein by a code (Gamow, 1954a, 1954b, 1961). Gamow’s handwritten letter to Watson and Crick dated July 8, 1953, has now come to light after fifty years (Watson, 2001). Gamow wrote:

But I am very much excited by your article in *Nature* May 30th and think that this brings biology over into the group of “exact” sciences. . . . If your point of view is correct, and I am sure it is at least in its essentials, each organism will be characterized by a long number written in quadrucal (?) system with figures 1, 2, 3, 4, standing for the four bases (or by several such numbers, one for each chromosome). It seems to me more logical to assume that different properties (single genes?) of any particular organism are not “located” in definite spots of chromosome, but are rather determined by different mathematical characters of the entire number.

\* As assumed in classical genetics

(Later at the Gatlinburg Symposium on Information Theory in Biology, I heard Gamow refer to this as “the number of the beast” [*Revelations* 13:18]: *This calls for wisdom: let him who has understanding reckon the number of the beast, for it is a human number, its number is 666.*)