

What is Information?

Marcello Barbieri

Received: 11 February 2011 / Accepted: 15 June 2011 / Published online: 26 January 2012
© Springer Science+Business Media B.V. 2012

The discovery of the double helix suggested in no uncertain terms that the sequence of nucleotides is the *information* carried by a gene (Watson and Crick 1953). A few years later, the study of protein synthesis revealed that the sequence of nucleotides in genes determines the sequence of amino acids in proteins, with a process that amounts to a transfer of linear information from genes to proteins. In both types of molecules, therefore, *biological information* was identified with, and defined by, the specific sequence of their subunits.

This idea was immediately accepted into the Modern Synthesis because it provided the molecular basis of both heredity and natural selection. Heredity became the transmission of information from one generation to the next, the short-term result of molecular copying. The long-term repetition of copying, on the other hand, is inevitably accompanied by errors, and in a world of limited resources not all copies can survive and a selection is bound to take place. That is how natural selection came into being. It is the long-term result of molecular copying, and can exist only in a world of informational molecules.

Information has become in this way the key concept of modern biology, but also the object of a fierce controversy. The reason is that information, heredity and natural selection simply do not exist in the world of chemistry, and this creates a contrast with the idea that “*life is chemistry*”. This view was proposed for the first time by Jan Baptist van Helmont (1648), and has been re-proposed countless times ever since. One of the most recent formulations has been given by Günther Wächtershäuser (1997) in these terms “If we could ever trace the historic process backwards far enough in time, we would wind up with an origin of life in purely chemical processes”.

Wächtershäuser claimed that “information is a *teleological* concept”, and that science does not really need it: “On the level of nucleic acid sequences it is quite convenient to use the information metaphor... and apply teleological notions such as ‘function’ or ‘information’... but in the course of the process of retrodiction the

M. Barbieri (✉)

Dipartimento di Morfologia ed Embriologia, Via Fossato di Mortara 64a, 44121 Ferrara, Italy
e-mail: brr@unife.it

teleological notions, whence we started, fade away. And what remains is purely chemical mechanism". This amounts to saying that biological information, the most basic concept of modern biology, is nothing more than a verbal expression.

The same charge has been made by the supporters of *physicalism*, the view that all natural processes are completely described, in principle, by physical quantities. The specific sequence of genes and proteins and the rules of the genetic code cannot be expressed by physical quantities, so what are they? According to physicalism, they are mere *metaphors*. They are linguistic terms that we use as shortcuts in order to avoid long periphrases. They are compared to those computer programs that allow us to write our instructions in English, thus saving us the trouble to write them in the binary digits of the machine language. Ultimately, however, there are only binary digits in the machine language of the computer, and in the same way, it is argued, there are only physical quantities at the most fundamental level of Nature.

This conclusion, known as *the physicalist thesis*, has been proposed by various scientists and philosophers (Chargaff 1963; Sarkar 2000; Mahner and Bunge 1997; Griffith 2001; Boniolo 2003) and is one of the most deeply dividing issues of modern science. Most biologists are convinced that information and the genetic code are real and fundamental components of life, but physicalists insist that they are real only in a very superficial sense and that there is nothing fundamental about them. Clearly, we must face this charge, and we must discuss it on its own grounds, i.e., in terms of physical theory.

The New Observables

The physicalist thesis would be absolutely correct if genes and proteins were formed spontaneously because all spontaneous reactions are completely accounted for by physical quantities. This, however, is precisely the point that molecular biology has proved wrong. Genes and proteins are *never* produced by spontaneous processes in living systems. They are produced by molecular machines that physically stick their subunits together according to sequences and codes. They are *manufactured molecules*, i.e., *molecular artifacts*.

The crucial point is that the manufacture of genes and proteins requires sequences and a genetic code, and we need therefore to find out what these entities actually are. By tradition, we call 'observables' the entities that allow us to describe the world, and since we cannot describe living systems without sequences and coding rules we must conclude that these entities are precisely that—*new observables* (Barbieri 2003, 2008). But what are, exactly, these new observables?

According to a long tradition, natural entities are divided into *quantities* and *qualities*. Quantities can be measured and are objective, whereas qualities are subjective and cannot be measured. In the case of information and meaning, however, this scheme breaks down. Biological (or organic) information, for example, is not a quantity because a specific sequence cannot be measured. But it is not a quality either, because linear specificity is an objective feature of the molecules, not a subjective one. The same is true for the meaning of the coding rules. This too cannot be measured, so it is not a quantity, but it is not a quality either because the rules of the genetic code are the same for all observers in all living systems.

A scheme based on quantities and qualities alone, in short, is not enough to describe the world. In addition to quantities (*objective and measurable*) and qualities (*subjective and not-measurable*) we must recognize the existence in Nature of a third type of entities (*objective but not-measurable*).

Organic information and organic meaning belong precisely to this new type of entities, and we can also give them a suitable name. Since they can be described only by *naming* their components in their natural order, we can say that they are *nominable* entities, or that they belongs to the class of the nominable entities of Nature (Barbieri 2004, 2006, 2008).

We realize in this way that organic information is an observable just as the physical quantities are, and this means that it *has the same scientific 'status' of the physical quantities*. This conclusion, in turn, raises immediately a new problem because there are two distinct groups of physical entities: a small group of *fundamental* entities (space, time, mass, charge and temperature) and a much larger group of *derived* entities. We need therefore to find out whether information belongs to the first or to the second group.

Luckily, this problem has a straightforward solution because the sequences of genes and proteins have two very special characteristics. One is that *a change in a single component of a biological sequence may produce a sequence which has entirely new properties*. This means that although a biological sequence can be said to have 'components', it is at the same time a single indivisible whole. The second outstanding feature is that *from the knowledge of n elements of a biological sequence we cannot predict the element $(n+1)$* . This is equivalent to saying that *a specific sequence cannot be described by anything simpler than itself*, so it cannot be a derived entity. Which means that organic information is not only a new observable, but a new *fundamental* observable.

Names and 'Nominable' Entities

Science is always expressed in words and we need therefore to give *names* to whatever we observe in Nature. Names (including those that we call 'numbers') are a necessary component of physical theory but have a peculiar feature. Laws, constraints and observables remain the same in all languages, whereas names are totally language-dependent. This is because names (or *nominal entities*, to use a classical term) in general have nothing to do with the intrinsic features of the named objects, and are therefore mere labels that we attach to them.

The deep divide that exists between 'names' and 'objects' has been at the centre of many controversies in the past, in particular of the celebrated medieval dispute over 'nominal entities' and 'real entities'. It has also had a long history in the philosophy of mathematics, where some have argued that numbers are 'invented' by the human mind, and others that they are 'discovered', a conclusion which implies that they have an existence of their own in some abstract Platonic world.

The relationship between names and objects is also a crucial issue in science, but here it has taken on a new form. Let us underline that all names are sequences of characters (alphabetic, numerical or alpha-numerical) and that each sequence is unique. Names, in other words, have *specificity*. In general, the specificity of a name

has nothing to do with the characteristics of the named object, and in these cases we can truly say that names are mere labels. Science, however, has invented a new type of names where the sequence of characters does represent an order that is objectively present in the named objects.

The chemical formula of a molecule, for example, describes an objective sequence of atoms, and any atom can be described by the objective sequence of its quantum numbers. In these cases, the names are no longer arbitrary labels but true '*observables*' because they describe characteristics that we observe in Nature. This shows that there are two distinct types of names in science: labels and observables.

In the case of the observables, furthermore, there is another distinction that must be made. When a molecule is formed spontaneously, its final sequence is due to the interactions between its own components, and in most cases it is completely determined by them. In the case of a protein, however, all its different amino acids interact by the same peptide bonds and a spontaneous assembly would produce a completely random order (which is incompatible with life). In this case, a specific sequence can be obtained only if the amino acids are put together by a molecular machine according to the order provided by a template that is *external* to the protein itself. We need therefore to distinguish between two different types of observables.

The sequence of quantum numbers in an atom, or the sequence of atoms in inorganic molecules, is determined *from within*, by internal factors, whereas the sequence of amino acids in a protein is determined *from without*, by external templates. In the first case the sequence is a *physically computable* entity, in the sense that it can be predicted by a formula, whereas in the second case it can only be described by 'naming' its components, and is therefore a *nominable* entity (this term should not be confused with the classical concept of *nominal* entity, which applies to all names). A *nominable* entity is not a label but an observable, and more precisely a *non-computable* observable.

All names, in conclusion, are specific sequences of characters, and in science they can be divided into two great classes: labels and observables. The observables, in turn, can be divided into *computable* entities and *nominable* entities. The important point is that physics and chemistry deal exclusively with computable entities (physical quantities), whereas nominable entities (information and coding rules) exist only in living systems.

The Discovery of New Worlds

The history of physics tells us that there are three logical steps in the scientific procedure. First we look at the world and choose a number of entities to describe it, entities that are called *observables* (space, time, mass, etc.) precisely because they represent what we observe. Then we look for relationships between observables and obtain models of the observed phenomena (equations, laws, regularities, etc.). Finally we use our models to make predictions that test them (we predict, for example, the eclipses of the moon etc.).

The choice of the observables is the first step in this procedure, and the most crucial one. The movements of planets and stars, for example, can be described with

only two observables - space and time - and in that case we get either a Ptolemaic system or a Copernican system. By introducing a third observable - mass - we obtain the laws of motion, universal gravitation and the Newton model of the world.

The three fundamental observables of classical physics can be changed (space and time, for example, can be replaced by velocity and time, and in that case space becomes a derived entity), but the number of fundamental observables does not change. That number defines a whole world of phenomena, and the discovery of new worlds is always associated with the discovery of new fundamental observables.

Mechanics, thermodynamics, electromagnetism and nuclear physics, have all been built on the discovery of new fundamental observables, and now we realize that this is true also in biology. Life began when the first molecular machines appeared on the primitive Earth and started manufacturing genes and proteins by copying and coding, i.e., by processes that necessarily require two new observables.

The nature of these observables has so far remained elusive because biologists have not yet come to grip with the idea that life is artifact-making. As soon as we realize that genes and proteins are manufactured molecules, we immediately see that sequences and coding rules are real observables. This conclusion, in turn, has three major consequences.

- (1) The first is that the physicalist thesis is *not* valid because an observable is an objective tool of science, not a metaphorical entity.
- (2) The second is that information is *not* a teleological concept, but a *descriptive* one. There is no more teleology in information and in the genetic code than there is in the quantities of physics and chemistry, because all of them are *observables*.
- (3) The third consequence is a new understanding of information. Biological information is indeed the sequence of genes and proteins, but the *nature* of these sequences has so far eluded us. Now we realize that they are objective and reproducible but non-computable observables. They are *nominable* entities, a new type of fundamental observables without which we simply cannot describe the world of life.

References

- Barbieri, M. (2003). *The organic codes. An introduction to semantic biology*. Cambridge, UK: Cambridge University Press.
- Barbieri, M. (2004). The definitions of information and meaning. Two possible boundaries between physics and biology. *Rivista di Biologia-Biology Forum*, 97, 91–110.
- Barbieri, M. (2006). Life and semiosis: the real nature of information and meaning. *Semiotica*, 158(1/4), 233–254.
- Barbieri, M. (2008). Biosemiotics: a new understanding of life. *Naturwissenschaften*, 95, 577–599.
- Boniolo, G. (2003). Biology without information. *History and Philosophy of the Life Sciences*, 25, 255–273.
- Chargaff, E. (1963). *Essays on nucleic acids*. Amsterdam: Elsevier.
- Griffith, P. E. (2001). Genetic information: a metaphor in search of a theory. *Philosophy of Science*, 68, 394–412.
- Mahner, M., & Bunge, M. (1997). *Foundations of biophilosophy*. Berlin: Springer Verlag.

- Sarkar, S. (2000). Information in genetics and developmental biology. *Philosophy of Science*, 67, 208–213.
- van Helmont, J.B. (1648). *Ortus Medicinae*. Amsterdam
- Wächtershäuser, G. (1997). The origin of life and its methodological challenge. *Journal of Theoretical Biology*, 187, 483–494.
- Watson, J. D., & Crick, F. H. C. (1953). Genetical implications of the structure of deoxyribose nucleic acid. *Nature*, 171, 964–996.