DNA as information

Peter R Wills

Department of Physics, University of Auckland, PB 92019 Auckland 1142, New Zealand

Institut für Biochemie und Molekularbiologie, Universität Hamburg, Martin-Luther-King-Platz 6, 20146 Hamburg, Germany

Department of Biochemistry and Biophysics, University of North Carolina, 120 Mason Farm Rd, Chapel Hill, NC 27599-7260 USA

Keywords: DNA; information; nominable entities; aperiodic crystal; computation; meaning; genetic coding; origin of life

1. Introduction

The biological significance of DNA lies in the role it plays as a carrier of information, especially across generations of reproducing organisms, and within cells as a coded repository of system specification and stability. These roles of DNA do not find any chemical explanation in terms of the average material properties of DNA as a irregular heteropolymer. To understand DNA's biological action one must go to the detailed molecular level. And then one also fails to find any simple answer in the DNA itself, because single molecules can have vastly different biological effects, covering the entire range of possibilities depending on the molecular biological context, even though they are identical except for the exchange of a particular one of the 10^9 nucleotide moieties of a genome, most such exchanges having very little effect. This drives us immediately to the conclusion that the DNA in organisms functions as *information* and that the internal DNA-dependent dynamics of cells embody functional information processing, that is, *computation*. DNA-based molecular biological computation can be said to control, perhaps even "direct", the entire panoply of biochemical events occurring in cells.

The obvious way in which information is stored in DNA, as sequences of letters drawn predominantly from the standard 4-letter {A, C, G, T} nucleotide alphabet, has been understood since the discovery of the substance's dual-linear base-paired molecular structure and its mode of complementary chain copying (Watson and Crick, 1953). However, as soon DNA is represented in these abstract terms, as information comprised of a sequence of arbitrary symbols, there is a new theoretical problem. Changes in the DNA sequence of an organism's genome translate in a regular causative way into biological changes in the concrete physical world: such regularities of causation are exploited by genetic engineers when they make calculated informational alterations to an organism's DNA. What is the nature of the causative connection between DNA sequence information, which is an arbitrary abstraction of a material property, and the reality of events in the physical world of molecules? This information/matter dichotomy is just one of a set of problems that has beset Western philosophy for centuries. To emphasise this point we refer to the theological ruminations of Thomas Aquinas concerning the Christian doctrine of the "incarnation" (Cross, 2002). If God became a man, was the man he became (Christ) still truly a man? In similar vein, how can we accept an arbitrarily defined class of abstract sequences as causes of biological events and preserve

our analytical view of organisms as causally closed physical systems? While scientists are unlikely to be empowered to perpetrate anything like the terror of the Inquisition in propagating their views as orthodox truth, we should not underestimate the magnitude and intensity of the internecine disputes that underlie discussion of the role of information in biology, such as have led to the separation of International Society for Biosemiotics Studies and the International Society of Code Biology (Barbieri, 2014). Barbieri (2016) solves the problem with his clear and unambiguous identification of *nominable entities* that embody an irreducible aspect of the natural world, manifest only in biology but as fundamental to reality as physical aspects of nature, such as mass, space and time.

The central problem of understanding the role of information in biology arises when attempts are made to go beyond counting the number of the bits of information in a genome and to link genotype with phenotype. DNA information accumulates predominantly through natural selection, a process which is well understood at the molecular level (Eigen, 1971). But knowing that what survives is the fittest does not enlighten one as to the character of the fitness landscape, that is, why, in terms of its internal structure, is one organism is fitter than another? Knowledge of what survives gives no insight into how genotypic information is mapped onto the phenotypic characteristics that define the internal factors, as opposed to environmental factors, contributing to an individual's fitness. In other words, how are we to understand the mapping from points in an extremely high dimensional (DNA) sequence space, entities as abstract as natural numbers, onto the real-world characteristics of organisms whose lives are at stake in the game of evolution? What is fundamental to the dynamic structure of systems in which such a mapping is maintained, in which the *meaning* of the information they store (in DNA) is exquisitely constrained by the system's detailed internal molecular structure? And what features of the material world provide for the spontaneous emergence of structures so remarkably ordered in comparison with what appears to be the bulk of abotic matter in which no integrated functionality is visible?

The answers to the closely associated questions "What is information?" and "What is the origin of life?" provided in this volume are quite disparate and although they do not span the gamut of what has been proposed since the discovery of DNA as information, their diversity, similarities and differences are worth exploring. It would be inappropriate for me as one of the authors involved to assume the privilege of setting up standards and evaluating colleagues' deliberations. However, without doing that it is possible, within a context of explicitly states premises, to describe the relationships between different approaches and to offer some observations concerning the range and relationships of the views which have been expressed. It is the purpose of this short review to compare and contrast the underlying concepts of information and biological processes expressed in a selection of the contributions to this volume (Adami, 2016; Barbieri, 2016; Koonin, 2016; Roederer, 2016; Varn and Crutchfield, 2016; Walker *et al*., 2016 and Wills, 2016).

2. Measuring the information in DNA

How the information content of DNA is to be defined and measured has been the subject of considerable disputation. Without doubt, the application of Shannon's formula is relevant, but views concerning the correct way of interpreting and applying the formula differ widely (Yockey, 1992; Muller, 2007; Schneider 2010; Battail, 2014). In the tutorial of Adami (2016) the ideas of information, uncertainty and entropy are used to relate the material world of physics to the world of human knowledge but we are not told how any of this relates to the information content of DNA. Varn and Crutchfield (2016) provide an enlightening treatment of the problem, absolutely rigorous in terms of the foundations of both physics and computation, locating their discussion within the context the biological necessity for an *aperiodic* crystalline structure in which to store information in nanoscopic matter (Schrödinger, 1944). The analysis likens DNA to the novel class of *chaotic* crystalline structures, establishing a new equivalence between molecular events in living and non-living systems, especially in respect of the possibility and measure of molecular information processing. Their results are applicable to both biological and artificially constructed systems. Although there is mention of the origin of life, the work circumvents any attempt to define the boundary of the "living", but the territory left for those who undertake such a task is considerably narrowed. The information theoretic version of the second law of thermodynamics presented by Varn and Crutchfield (2016) is of significance to biology yet to be determined by its application, but the insight that "[t]he existence of natural [Maxwell's] Demons with memory (internal states) is a sign that they have been adapted to leverage temporally correlated fluctuations in their environment" is seminal, nucleating in the description of a single system concepts from the theories of molecular evolution, computation and nonlinear irreversible thermodynamics.

Elsewhere, Adami (2015) has described how it is that the probability of emergence of replicating biotic systems is increased if monomers are supplied at individual relative rates that are in proportion to their relative abundances in the biotic polymers. This is the same as saying that random typing is more likely to produce meaningful sequences of letters if the probability of typing letters matches the frequency of occurrence of the letters in the language of choice. However, Adami (2015) links the biological notion of selective adaptation to his description of an integrated system of functional polymers with an overall monomer composition matching that of the environment. Restricting consideration to the case of natural selection among molecular self-replicators constrained by the supply of monomers should provide opportunity for a clear demonstration of the results presented by Varn and Crutchfield, perhaps with the explicit connection between thermodynamics and the theory of evolution left incomplete by Eigen (1971) and recently considered by England (2014).

Koonin (2016) takes the view that biological information is "effectively orthogonal" to Shannon information, because it has to do with meaning, not the statistical distribution of symbols in a sequence – the biological meaning of sequences can be found only through the alignment of homologous sequences, not by examining individual sequences. No specification of how homology can be rigorously determined is provided – it can only be assumed that bioinformatic

techniques for aligning sequences are adequate – but meaning can be measured through the "vertical" comparison of aligned sequences from different sources, rather than the "horizontal" comparison of sites along a single sequence. This meaning is seen to be transferred by DNA exchange between genomes during evolution. Koonin identifies a quantity he calls the "information density" as a measure of the meaning of a DNA sequence. It corresponds to the average, across an alignment, of the single-sequence-position deviation from randomness. However, in the end Koonin completely relativizes his information theoretic measure of biological meaning by pointing out that the question "meaning for whom?" is answered through the range of orthologous sequences chosen for the alignment in relation to which information density is calculated.

3. Biological information processing

The essence of computation is information processing, and the essence of biological information processing is control of the molecular events inside a system. Thus, Walker *et al*. (2016), taking fission of a complete single-celled yeast cell as an example, locate the special character of biological systems – the connection between information and causation – in the "informational architecture" of the cell, the spatio-temporal structure of the transformation of information during biological processes. The information theoretic analysis they present is rigorous and the conclusion reached is that "biology is distinguished from physics … in how the flow of information directs the execution of function". They describe this in terms of the informational architecture interacting with the system's causal structure, which is construed to be physico-chemical. While the analysis is intended to apply to any level of biological organisation, including DNA, there is no exposition of how the connection works at the level of nucleotide sequences and chemical reactions. However, the implicit genetic control of the individual processes of the yeast fission system makes the study relevant to an understanding of exactly how it is that quantities of functional information, originating in a DNA repository and corresponding to dynamically constrained distributions of alternative states, can operate to govern the whole-system behaviour.

Roederer (2016) also characterises biological systems in terms of a causal connection between informational patterns and events in the material world. Biological information is *pragmatic*, its importance lying not in the quantity of it but in what it effects in any system. For example, there is an essentially univocal correspondence between the linear pattern of DNA bases presented to cells of a certain species and the ensuing complex of molecular biological dynamic changes that take place. However, pragmatic information cannot be measured, because it represents a correspondence between a pattern and a change; and thus it is highly context-dependent. This idea of pragmatic information is very close to what others like Koonin (2016) and Wills (2016) would refer to as the *meaning* of information like the pattern of nucleotide bases in a genome. As Roederer (2016) states "a pattern all by itself has no meaning", which is echoed in the ansatz of Wills (2014): "Any body of information can be given any meaning whatsoever, by creating a device that functions as an interpreter to deliver the

specified meaning upon reception of that information". Roederer (2016) and Wills (2014) both imply the possibility of epigenetically defined phenotypes, such as the distinguishable strains of yeast that breed true in the same environment even though they have identical genomes (Wickner *et al*., 2015).

Like Varn and Crutchfield (2016), Wills (2016) focuses on the linear, aperiodic crystalline structure of DNA as the medium for the static physical instantiation of biological information, a body of which can constitute the complete, heritable "specification for the construction and maintenance of an entire organism" (Schrödinger, 1944). His enquiry delves into how molecular systems can self-organise to fuse a union between the DNA sequence information they contain and the internal molecular componentry that cooperatively generates meaning from the information. The investigation considers the machinery of translation, the system for executing the rules of the genetic code, as an example of a molecular biological interpreter, representative of a system of functional computation that is fundamental to all biological systems. It is argued that the principle of natural selection does not alone account for the evolutionary accumulation of information in DNA (Eigen 1971). Rather, equal emphasis should be given to processes of *epigenesis*, whereby selective advantage is conferred on genetic sequences by virtue of their coincidental occurrence with new interpretations of them, interpretations that simultaneously emerge, together with the selected information, as a result of functional self-organisation within the system. Neither information nor function is given causative precedence.

4. Origins

Both Barbieri (2016) and Wills (2016) seek a deeper understanding of the nature of biological information in the historical origin of genetic coding. Both authors take the emergence of coding as an essential element of the transition from abiotic and biotic chemistry. For Barbieri, the transition across the boundary entails the appearance of *nominable entities*, which can be described only by naming the order of their components, that is, by specifying a pattern of information. This property qualifies them to be designated as manufactured *artifacts*, in stark contrast to all the other molecules in the universe, which are "spontaneous". The first genes and proteins were spontaneous, but some molecules somehow took on the status of *machines*, functioning as "bondmakers" and "copymakers", in turn conferring on some genes the status of being pattern-preserving, information-bearing *templates*. This enabled the appearance of the first artifacts, molecules whose components were ordered by pre-existing information. However, before the emergence of the genetic code none of the template-maintained information had *meaning* and it is on this basis that Barbieri (2016) proposes his *code paradigm* of life as "chemistry+information+codes". Elsewhere (Barbieri, 2015) attributes the emergence of accurate, unambiguous coding to the evolution of ribosomal proteins, but comprehensive bioinformatic analyses (Harish and Caetano-Anollés, 2012; Caetano-Anollés *et al*., 2013; Petrov *et al*., 2015) point to the

precedence of specifically functional aminoacyl-tRNA synthetase (aaRS) proteins.

The analysis of the origin of coding presented by Wills (2016) has a completely different theoretical foundation. It starts with considerations of the dynamics of autocatalytic networks of polymers and the chicken-egg problem of the coding enzymes (aaRSs) being needed for their own synthesis. Instabilities connecting alternative solutions to the dynamic equations describing the elementary chemical processes of gene replication and translation provide the context in which coding can arise from random peptide synthesis as a result of relatively simple systems self-organising thermodynamically. The need to preserve, through the process of natural selection, the genetic information coding the aaRSs demands that the products of genetic replication and translation be colocalised. In the absence of compartmentalized proto-cells that reproduce their entire contents in the right proportions as living cells do, Turing reaction-diffusion coupling provides an elementary mechanism for polymeric genotypes and phenotypes to maintain colocalisation and stave off the caustic effects of computational errors. As a result of dynamic transitions described as "quasi-species bifurcations" in such systems, coding ambiguity is reduced in parallel with the progressive accumulation of genetic information sufficient for progressively more complex populations of aaRS "statistical proteins" to specify themselves with expanded codes of increasing precision (reduced error). Thus, a solution to the problem raised by Barbieri (2015) has existed for a decade and a half (Wills, 2001) and has since been carefully elaborated, most recently spawning a novel bioinformatic analysis of the deep co-phylogenies of functional aaRS structures (Wills *et al*., 2015).

5. Conclusion

The topic "DNA as information" focuses attention on biological information as it can be stored statically in molecular structures. But the essence of information in biology is its dynamic transfer into different forms and the effects of such transfers. Nowhere is that more evident than in the contribution of Walker *et al*. (2016), who analyse the consequences of information flows and processing in molecular biological control systems. Their methodology is more reminiscent of the thermodynamic approach of England (2014) than others whose work has a direct connection the Central Dogma (Crick, 1958). However we have come a long way from the simple maxim "DNA makes RNA makes protein". Every new discovery of a biological macromolecule or function modifies the estimate of the information content of the DNA of the species concerned and other members of its clade. Koonin (2016) has proposed a way of quantifying the information in DNA relative to the breadth of the selected clade, giving us a view of how the meaning of information in DNA continually evolves as a result of mutation and gene transfer. Roederer (2016) considers the general consequences of some pre-existing molecular pattern influencing the path of physical processes, emphasising the need for what amounts to "recognition" of information for it to have meaning in biological contexts. The origin of such recognition processes and their consequences is the focus of the contributions from Barbieri (2016)

and Wills (2016), both of whom take the recognition and matching of codons and amino acids, genetic coding, to be a defining feature of biology and the origin of life. However, as the work of Varn and Crutchfield (2016) shows, there is a much broader class of structures that could potentially mimic the function of "DNA as information" in molecular biological-like systems. Might it be possible, in completely different environments, for systems of nanoscopic processes to bootstrap themselves into existence and evolve as a result of their association with a colocalised repository of non-DNA information, which they manage to interpret as a programme for their construction through a network of processes not involving a simple translation step? The possibility of creating artificial systems of that sort is certainly being explored (McCaskill *et al*., 2012).

References

Adami, C. 2016 What is information? Phil. Trans. R. Soc. A (THIS VOLUME).

Adami, C. 2015 Information-theoretic considerations concerning the origin of life. Orig. Life Evol. Biosph. 45, 309-17 (doi: 10.1007/s11084-015-9439-0).

Barbieri, M. 2014 From biosemiotics to code biology. Biol. Theory 9, 239-249.

Barbieri, M. 2015 Evolution of the genetic code: The ribosome-oriented model. Biol Theory 10, 301–310 (DOI 10.1007/s13752-015-0225-z).

Barbieri, M. 2016 What is information? Phil. Trans. R. Soc. A (THIS VOLUME).

Battail, G. 2014 Information and Life. Berlin: Springer.

Caetano-Anollés, G, Wang, M, Caetano-Anollés, D. 2013 Structural phylogenomics retrodicts the origin of the genetic code and uncovers the evolutionary impact of protein flexibility. PLOS ONE 8(8): e72225.

Crick, F. H. C. 1958 On protein synthesis. Symposia of the Society for Experimental Biology 12, 138–163.

Cross, R. 2002 The metaphysics of the incarnation, Thomas Aquinas to Duns Scotus. Oxford: Oxford University Press.

Eigen, M. 1971. Self-organisation of matter and the evolution of biological macromolecules. Naturwissenschaften 58, 465–532.

England, J. L. 2014. Statistical physics of self-replication. J. Chem. Phys. 139, 121923. (DOI 10.1063/1.4818538).

Harish, A, Caetano-Anollés, G. 2012. Ribosomal history reveals origins of modern protein synthesis. PLoS ONE 7(3): e32776.

Koonin, E. V. 2016 The meaning of biological information. Phil. Trans. R.

Soc. A (THIS VOLUME).

McCaskill, J S, v. Kiedrowski, G, Oehm, J, et al. 2012. Microscale Chemically Reactive Electronic Agents. Int. J. Unconv. Comp. 8, 289–299.

Muller, S. 2007. Asymmetry. Berlin: Springer

Petrov AS, Gulen B, Norris AM *et al*. 2015. History of the ribosome and the origin of translation. PNAS Early Edition, www.pnas.org/cgi/doi/10.1073/pnas.1509761112

Roederer, J. G. 2016 Pragmatic information in biology and physics. Phil. Trans. R. Soc. A (THIS VOLUME).

Schneider, T. D. 2010 A brief review of molecular information theory. Nano. Comm. Networks 1, 173–180 (doi:10.1016/j.nancom.2010.09.002).

Schrödinger, E. 1944 What is life? Cambridge: Cambridge University Press.

Varn, D. P., Crutchfield, J. P. 2016. What did Erwin mean? The physics of information from the materials genomics of aperiodic crystals and water to molecular information catalysts and life. Phil. Trans. R. Soc. A (THIS VOLUME).

Walker, S. I., Kim, H, Davies, P. C. W. 2016 The informational architecture of the cell. Phil. Trans. R. Soc. A (THIS VOLUME).

Watson, J. D., Crick, F. H. C. 1953 Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. Nature 171, 737–738; Genetical Implications of the Structure of Deoxyribose Nucleic Acid. Nature 171, 964-67.

Wickner, R. B., Shewmaker, F.P., Bateman, D.A. *et al.* 2015 Yeast prions: structure, biology, and prion-handling systems. Microbiol. Mol. Biol. Rev. 79, 1-17 (doi: 10.1128/MMBR.00041-14).

Wills, P. R. 2001 Autocatalysis, Information and Coding. BioSystems 60, 49-57.

Wills, P. R. 2014 Genetic information, physical interpreters and thermodynamics; the material-informatic basis of biosemiosis. Biosemiotics 7, 141-165 (DOI 10.1007/s12304-013-9196-2).

Wills, P. R. 2016 The generation of meaningful information in molecular systems. Phil. Trans. R. Soc. A (THIS VOLUME).

Wills, P. R., Nieselt, K., McCaskill, J. S. 2015 Emergence of coding and its specificity as a physico-informatic problem. Orig. Life Evol. Biosph. 45,

249–255 (DOI 10.1007/s11084-015-9434-5).

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