Orig Life Evol Biosph (2015) 45:249–255
DOI 10.1007/s11084-015-9434-5

ORIGINS 2014

# Emergence of Coding and its Specificity as a Physico-Informatic Problem

**Peter R. Wills**[1] · **Kay Nieselt**[2] · **John S. McCaskill**[3]

Published online: 28 March 2015


**Abstract** We explore the origin-of-life consequences of the view that biological systems are demarcated from inanimate matter by their possession of referential information, which is processed computationally to control choices of specific physico-chemical events. Cells are cybernetic: they use genetic information in processes of communication and control, subjecting physical events to a system of integrated governance. The genetic code is the most obvious example of how cells use information computationally, but the historical origin of the usefulness of molecular information is not well understood. Genetic coding made information useful because it imposed a modular metric on the evolutionary search and thereby offered a general solution to the problem of finding catalysts of any specificity. We use the term "quasispecies symmetry breaking" to describe the iterated process of self-organisation whereby the alphabets of distinguishable codons and amino acids increased, step by step.

**Keywords** Referential information · Computational processes · Genetic coding · General solutions · Quasispecies symmetry breaking

When we talk about a genetic code, we are no longer using the language of physics and chemistry to describe the processes occurring in biological systems – we are looking at the molecular operation of cells in *computational* or *informatic* terms. Do cells really process information and transmit it through the mechanisms of heredity as Schrödinger (1944) suggested, or have we invented computational descriptions of molecular biological events as

Paper presented at ORIGINS 2014, Nara Japan, July 6-11 2014.

✉ Peter R. Wills
p.wills@auckland.ac.nz

[1] Department of Physics, University of Auckland, PB 92019, Auckland 1142, New Zealand

[2] Integrative Transcriptomics, Center for Bioinformatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany

[3] Microsystems Chemistry & BioIT: BioMIP, Faculty of Chemistry & Biochemistry, Ruhr Universität Bochum, Universitätsstr. 150, Bochum 44801, Germany

fictional devices to help us better understand them by analogy? In this contribution we explore the origin-of-life consequences of the view that biological systems *do* process information in a referential fashion, such that genetic information is *about* something, rather than being an epiphenomenon of no causal relevance to events in the world of molecules in which it is instantiated.

To say that life is a chemical reaction is an empty truism. So is weather a chemical reaction at the same level of description. Defining life as a complex, self-sustaining chemical system capable of Darwinian evolution (Joyce 1994) gets us further, but it does not include the most important empirical findings of modern molecular biology, anticipated by Schrödinger (1944) in his idea of a genetic "codescript": that the mechanisms of heredity and genetic expression are essentially informatic; that genes can be regarded as mathematically describable patterns, abstract *sequences* embodied in nucleic acid form; and that variations in and alterations to genetic sequences comprise one of the main means whereby organisms evolve naturally and can be manipulated artificially. Evolutionary selection is displayed by replicating polymers, metabolism is displayed by dissipative structures and compartmentalization is displayed by all manner of vesicles, but informatic processes are only evident in biological systems or artifacts produced by them.

This begs of us the question: how did polymer sequence information first acquire some sort of referential meaning in respect of the systems in which it occurred? (Wills 1994). It is not necessary to have either replication or genetic coding in order to maintain information above the error catastrophe threshold of Eigen (1971) in an autocatalytic system of heteropolymers (Wills and Henderson 2000). An RNA World devoid of the complicated machinery needed to reproduce sequence information directly or transfer it accurately into another domain (e.g., catalytic proteins) is capable of displaying all of the main features of biochemical evolution – especially transitions to dynamic states of increasing complexity and higher information content (Takeuchi and Hogeweg 2012). However, the evolution of complex systems of catalytic polymers seems to require the natural operation of *bioinformatic* processes, according to the original conception of that term (Hogeweg 2011).

There is a view that genetic information is unnecessary for life, that genes are an afterthought, a kind of syntax that enables, but does not cause, wider evolution. The idea is that genes provide just one way of exploring spaces of proteins and RNA sequences that may have direct or indirect functional consequences for the coherent, integrated systems in which they occur. This view is supported by studies of "reflexively autocatalytic and food-generated (RAF)" sets of chemical species (Hordijk et al. 2012, 2014). So far, RAF theory has not taken comprehensive account of the exigencies of non-equilibrium thermodynamics in complex chemical systems and has been applied only to homogeneous systems in which they cannot generally be maintained. However, it is possible that the coupling of autocatalytic functional closure with some sort of spatial localization or encapsulation, as mooted by Gánti (2003) can account for the modes of integration characteristic of systems that propagate in biological evolution, irrespective of whether they store information in genes.

We hold to a contrary position: that the essential feature of biological systems that demarcates their constitution from "inanimate matter" is their possession of referential information, which is processed computationally to control choices of specific physico-chemical events, resulting in the rejection of a myriad of other possibilities. The information is maintained as a particular state of an otherwise highly degenerate molecular subsystem: a specific, fixed, heteropolymeric sequence selected from an astronomical number of alternatives. When we say that the information is *processed computationally*, we mean that there exist in the system operational mechanisms that could interact with practically any sequence of the informational heteropolymer, but with different consequences for the system's operation,

stability and survival. We wish to draw attention to the importance for the origin of life of the evolutionary emergence of general mechanisms that support computational operations.

The control of any dynamic system requires the internal transfer of information, whether through analog "servo" linkages or digitally-directed local changes of state. However, even in the most highly digitally controlled systems, like large assembly lines or scientific instruments, sequences of bits are ultimately translated into physical effects in some secondary domain (the primary domain being the one in which the information is stored, manipulated and transmitted in the system). In conventional technology, the secondary domain is usually the world of macroscopic objects and components that are subject to the classical laws of Newtonian mechanics and Maxwellian electromagnetism, but in nanotechnology the scale of the secondary domain can be reduced to that of single molecules. This is the scale at which living cells maintain detailed control of their internal operation. Their molecular structure is chemically and spatially differentiated down to the finest detail. The integration and coherence of their internal dynamic control, which includes physical boundary maintenance, is said to imbue them with operational "autonomy" (Maturana and Varela 1980).

It is clear that a large proportion of the molecular processes and interactions that occur inside organisms cannot be regarded as anything more than "servo-regulated" in a statistical thermodynamic sense. Rates of enzyme catalysis appear to respond to changes in the concentrations of allosteric effectors, substrates and products in a quasi-analog manner and are subject to fluctuations and variations due to mass action-type macromolecular interactions. However, there are many others, perhaps those regarded as occurring at higher levels of the system control hierarchy, that rely on either the direct or indirect recognition of genetic sequence information, or copies of it. The ribosomal translation of nucleic acid sequence information into the secondary domain of protein amino acid sequences marks its irretrievable use, its transfer from "sign" to "signified": the Central Dogma of molecular biology (Crick 1958) notes that there is no direct "back translation" from protein to nucleic acid sequences.

To the extent that many molecular biological interactions involve the recognition of specific polymeric sequences, they can be looked at from a computational perspective. Then in a wider view, the internal operation of a cell can be regarded as cybernetic: information is used in processes of communication and control; physical events are subject to a system of integrated governance. What is most remarkable about this aspect of living systems is the specificity and precision of the narrow range of processes which they maintain. Without this high degree of selectivity, the level of biological complexity observed in nature could not exist, or so we presume. Such precision of control is necessary to ward off any tendency to err into the vast array of alternative, thermodynamically degenerate pathways that lead to death. In the most extreme case, loss of just one critical bit of information from the billions in a genome can yield an organism non-viable (Gibson et al. 2010). It is our contention that the stable control and constructive management of energy flows in complex molecular biological systems requires that many individual steps be regulated with precision equivalent to a significant quantity of information – many control bits. Thus, the emergence of a domain of chemical activity in which the processing of reaction control information was hierarchically separated from the rest of the system enabled life at its origin.

The transition to genetic coding produced a platform for the evolution of more and more differentiated and refined control over energy flows and thereby the definition of system functions of ever-increasing specificity, far beyond what could occur in a non-informatic autocatalytic system. The division of labor between two domains, information carrying template and functional catalysts, of itself gives an evolutionary advantage to replicating

systems (Takeuchi et al. 2011). Such advantage is amplified if the range and density of catalytic possibilities of the functional polymer (e.g., protein) in its sequence space is much richer than that of the information-carrying polymer (e.g., RNA) in its sequence space. Amino acids are much more compact than nucleotides and their coded utilization does not restrict their choice to pairs suitable for replication through complementarity. In a globular protein, the variation in local chemical structure can be rated at about 20 choices per 0.13 $nm^3$, on average, whereas the corresponding density of variation in nucleic acids is only 4 choices per 0.30 $nm^3$. Considering the combinatorial possibilities of adjoining neighbourhoods, it is no wonder that protein catalysts offer much greater opportunity than nucleic acid catalysts for the fine, specific control of chemical reactions.

The problem with proteins is that their sequences are not amenable to replication, so it is to be expected that RAF sets of proteins would (i) be very large (Kauffman 1986), (ii) involve a large number of reactions of different specificities and (iii) require a high energetic investment in system "housekeeping". On the other hand, the complementary base pairing of nucleotide sequences provides the possibility of copying and transferring information almost "for free". Thus, coding offers a solution to the problem of narrowing down the range of molecular species and reaction specificities needed to maintain a system of non-replicating polymers. Coding fixes an easily reproducible informatic *representation* of an entity of high functional specificity (protein) in an energetically inert information-holding subsystem (nucleic acid gene).

At this point we note that coding offers a *general* solution to the problem of finding catalysts of any specificity, because coding imposes a modular metric on the evolutionary search: the mapping from the information space (nucleic acid sequences) onto the functional domain (protein catalysis) defines the resolution (single steps in protein sequence space) at which evolution explores functional possibilities and at which adaptive change is significant. In fact, when we say that coding offers a general solution we mean that it defines the entire space of evolutionary possibilities as problems of a single category. That is why genetic algorithms are successful: they exploit the inherent advantage of representing an evolutionary search problem by encoding possible outcomes as arbitrary representations in a space that can be traversed in steps consisting of a relatively small number of very simple operations. Such encodings also create the potential for "open-ended evolution", through the emergence of simple subroutines whose operation produces modular entities with higher order functionalities. Emergence of this sort has been demonstrated in the evolution of (i) digital circuits that perform elementary logical operations (Lenski et al. 2003) and (ii) the complex task of arithmetic multiplication (Füchslin et al. 2006).

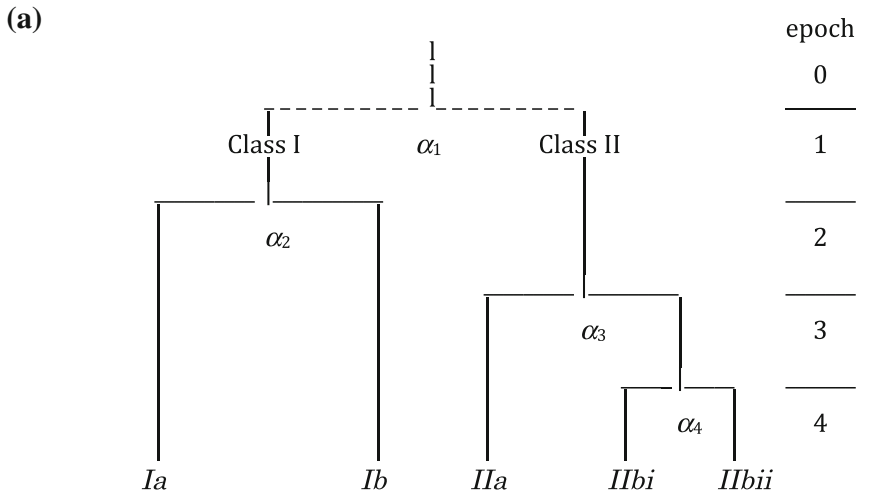
The tension between specific and general problem-solving accompanies the evolution of biological organisms throughout their history. Optimizing the assembly of molecular machinery to solve a specific task (e.g., the synthesis of an organic catalyst) is contrasted with optimizing the machinery economically to support a whole range of changing tasks (e.g., catalyzing a whole network of reactions with time varying structure resulting from varying chemical resources). Further examples of general problems include: recognizing arbitrary molecular shapes (solved by encoding antibodies); replicating arbitrary molecules (solved by encoding their synthesis in a replicable molecule such as DNA); at a higher organismic level, sexual reproduction (recognizing individuals from the same species with varying properties and recombinative inheritance); and the recognition of patterns of light (solved by retinal fields with the construction of spatial topology-preserving neural networks). In each of these cases, specific problems can be solved more efficiently by using highly customized machinery. But as the number of problem variants and the rate of change of the problem set both increase, there comes a point where it is more efficient to assemble machinery that supports a more rapid

solution and optimization for new problem instances, often by operating in an iterative fashion. This is a clear case where issues of cooperation and exploitation also play a role: such machinery can also be exploited by viruses or other organisms to produce optimal solutions for exploitation of the host organism – e.g., all viruses exploit the protein translation machinery provided by the cell. The digital character of nucleotide-base information provides for its transmission as well as the easy stepwise spanning of evolutionary problem space in an RNA World with a rich phenomenology (Takeuchi and Hogeweg 2012), but not at the level of chemical sophistication and specificity possible with proteins. Substantial machinery can be stably evolved if the problem space involves sufficiently many entities that can be optimized via a common mechanism, enabled by informatic encoding, as has been shown in the case of protein translation (Füchslin and McCaskill 2001). When exactly it is sufficiently advantageous for an organism to invest in general purpose problem-solving machinery, given the extra overhead cost in terms of resources and the dangers of exploitation, will be the subject of a separate investigation.

It is evident that genetic coding conferred a selective advantage on the proto-biological system in which it emerged in such a way that all surviving organisms employ the same near-universal genetic code. Coding enabled heredity in physical systems in exactly the manner envisaged by Schrödinger (1944): every organism carries a stable, compact, reproducible description (computational "codescript"), which can, to a limited extent, act as a naked representation of itself and its construction, as demonstrated by the genome substitution exercise conducted by Gibson et al. (2010). But what enabled a computational system as sophisticated and specific as the 64 codon to 21 amino acid code of the normal genetic code (counting "stop" as a null amino acid) progressively to emerge as the dominant executive mechanism of chemical change in the prebiotic world? Our answer to this question is that it was driven by coupled transitions in the specificity of protein autocatalysis and the accuracy of RNA replication. The transitions, whose underlying mechanism is illustrated in Fig. 1 of the accompanying paper (Wills 2015), were enabled by coincidental instabilities in both the chemical dynamics (autocatalytic amplification of molecular species) and the computational dynamics (recursive amplification of coding assignments) of co-dependent peptide and RNA synthesis.

According to this view, the first bits of information to acquire referential meaning, to function computationally in a prebiotic system, were found in the complementary strands of a double-stranded RNA quasi-species, perhaps about 290 bases long (Li and Carter 2013; Carter et al. 2014), which was replicated with an accuracy above the threshold of Eigen (1971). And the meaning of the information in the two strands was manifest as proteins with distinguishable Class I and II aminoacylation functionalities. Peptides with these distinct functionalities are envisaged to have been synthesized through the ordering, along the strands of the information-carrying RNA, of aminoacylated tRNA-like species produced by them (Wills 2015). Such a process could transfer information from the RNA to the protein domain as long as the specificity of aminoacylation remained above the error catastrophe threshold of Orgel (1963; 1970). The proteins would have been "statistical" in the sense of Woese (1965), but there would have been sufficient bimodality in the overall protein population to define distinguishable Class I and II functionalities. Coding would not have needed to be particularly strict – the sets of amino acids utilized by the original Class I and II urzymic statistical proteins would not have needed to be completely disjoint, just sufficiently different to maintain the distinction in functionality between the two protein subpopulations produced from the separate strands of the information-carrying RNA.

This initial scenario of a primitive binary code (Epoch 1 in Fig. 1a) can be reached through co-dependent RNA quasi-species selection (Eigen 1971) and protein coding self-organisation (Wills 1993), starting from a chemically homogeneous system (Füchslin and McCaskill 2001).




**(b)**

$$R_1 = \begin{pmatrix} I & II \\ - & \alpha_1 \\ \alpha_1 & - \end{pmatrix} \quad R_2 = \begin{pmatrix} Ia & Ib & II \\ - & \alpha_2 & \alpha_1 \\ \alpha_2 & - & \alpha_1 \\ \alpha_1 & \alpha_1 & - \end{pmatrix} \quad R_3 = \begin{pmatrix} Ia & Ib & IIa & IIb \\ - & \alpha_2 & \alpha_1 & \alpha_1 \\ \alpha_2 & - & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & - & \alpha_3 \\ \alpha_1 & \alpha_1 & \alpha_3 & - \end{pmatrix} \quad R_4 = \begin{pmatrix} Ia & Ib & IIa & IIbi & IIbii \\ - & \alpha_2 & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_2 & - & \alpha_1 & \alpha_1 & \alpha_1 \\ \alpha_1 & \alpha_1 & - & \alpha_3 & \alpha_3 \\ \alpha_1 & \alpha_1 & \alpha_3 & - & \alpha_4 \\ \alpha_1 & \alpha_1 & \alpha_3 & \alpha_4 & - \end{pmatrix}$$

**Fig. 1** **a** Epochs of increasing coding specificity in the phylogenetic tree of aaRS subclasses. The tree is artificially rooted above the Class I/II coalescence point in a random population of proteins displaying some overall undifferentiated aaRS activity. **b** Minimally parameterized amino acid substitution matrices $R_i$ for each epoch $i$=1,2,3,4. The substitution rates $\alpha_i$ are arranged hierarchically so that they are relevant only to amino acid alphabet distinctions that post-date emergent distinctions in aaRS specificities. The complete aaRS tree of extant organisms comprises epochs 1-19 and predates the Last Universal Common Ancestor

Subsequent epochs can be reached through a similar process, each transition corresponding to a further bifurcation in the codependent populations of the information-carrying RNA quasispecies and the functional urzymes. The iterated process whereby the alphabets of distinguishable codons and amino acids increased, step by step, could usefully be called *quasispecies symmetry breaking* – a non-Darwinian process of epigenetic self-organisation, akin to sympatric speciation (Wills 2014). At each branchpoint in the joint Class I & II aaRS phylogenies the very opposite of information expression occurs: a macroscopic event (a non-equilibrium phase transition) creates, *de novo*, a formal distinction, which previously did not exist in the system, between two classes of entities. That distinction then enables the selection of encoded information that is used to control choices between different possible events in the system. We therefore propose that recreation of the universal phylogenies of aaRSs (O'Donoghue and Luthey-Schulten 2003; Caetano-Anollés et al. 2013) should now be conducted using substitution matrices, such as those shown in Fig. 1b, which reflect the limited specificities of the amino acid (and codon) alphabets that were functional in different epochs (Wills 2014). In this way we will achieve a deeper understanding of how systems of computation bootstrap themselves into existence. We will also be better placed to enable nanotechnologies that are

in important ways "self-designed" for the controlled accomplishment of selected general tasks (McCaskill et al. 2012).

**Acknowledgments** This work received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 318671 (MICREAgents). PRW thanks Nobuto Takeuchi for his helpful suggestions.

## References

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:e72225

Carter CW Jr, Li L, Weinreb V et al. (2014) The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene. Biol Direct 9:11

Crick FHC (1958) On protein synthesis. Symp Soc Exp Biol 12:138–163

Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. Naturwissenchaften 58:465–523

Füchslin RM, McCaskill JS (2001) Evolutionary self-organization of cell-free genetic coding. Proc Natl Acad Sci U S A 98:9185–9190

Füchslin RM, Maeke T, Tangen U, McCaskill JS (2006) Evolving inductive generalization via genetic self-assembly. Adv Complex Sys 9:1–29

Gánti T (2003) The principles of life. Oxford University Press, Oxford

Gibson DG, Glass JI, Lartigue C et al. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329:52–56

Hogeweg P (2011) The roots of bioinformatics in theoretical biology. PLoS Comput Biol 7(3):e1002021

Hordijk W, Steel M, Kauffman SA (2012) The structure of autocatalytic sets: evolvability, enablement, and emergence. Acta Biotheor 60:379–392

Hordijk W, Wills PR, Steel M (2014) Autocatalytic sets and biological specificity. Bull Math Biol 76:201–224

Joyce GF (1994) Foreword. In: Deamer DW, Fleischaker GR (eds) Origins of life: the central concepts. Jones and Bartlett, Boston, pp xi–xii

Kauffman SA (1986) Autocatalytic sets of proteins. J Theor Biol 119:1–24

Lenski RE, Ofria C, Pennock RT, Adami C (2003) The evolutionary origin of complex features. Nature 423:139–145

Li L, Carter CW Jr (2013) Full implementation of the genetic code by tryptophanyl-tRNA synthetase requires intermodular coupling. J Biol Chem 288:34736–34745

Maturana HR, Varela FJ (1980) Autopoiesis and cognition: the realization of the living, 42nd edn, Boston studies in the philosophy of science. D Reidel Publishing Company, Dordrecht, Holland

McCaskill JS, von Kiedrowski G, Oehm J et al. (2012) Microscale chemically reactive electronic agents. Int J Unconv Comput 8:289–299

O'Donoghue P, Luthey-Schulten Z (2003) On the evolution of structure in aminoacyl-tRNA synthetases. Microbiol Mol Biol Rev 67:550–573

Orgel LE (1963) The maintenance of the accuracy of protein synthesis and its relevance to ageing. Proc Natl Acad Sci U S A 49:517–521, 67:1476

Schrödinger E (1944) What is life? Cambridge University Press, Cambridge

Takeuchi N, Hogeweg P (2012) Evolutionary dynamics of RNA-like replicator systems: a bioinformatic approach to the origin of life. Phys Life Rev 9(219–263):279–284

Takeuchi N, Hogeweg P, Koonin EV (2011) On the origin of DNA genomes: evolution of the division of labor between template and catalyst in model replicator systems. PLoS Comput Biol 7:e1002024

Wills PR (1993) Self-organisation of genetic coding. J Theor Biol 162:267–287

Wills PR (1994) Does information acquire meaning naturally? Ber Bunsenges Phys Chem 98:1129–1134

Wills PR (2014) Genetic information, physical interpreters and thermodynamics; the material-i nformatic basis of biosemiosis. Biosemiotics 7:141–165

Wills PR (2015) Spontaneous mutual ordering of nucleic acids and proteins. Orig Life Evol Biosph. doi:10.1007/s11084-014-9396-z

Wills PR, Henderson L (2000) Self-organisation and information-carrying capacity of collectively autocatalytic sets of polymers: ligation systems. In: Bar-Yam Y (ed) Unifying themes in complex systems. Perseus Books, New York, pp 613–623

Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci U S A 54:1546–1552