COMPLEXITY INTERNATIONAL

Volume 08

http://www.csu.edu.au/ci/

Evolution of the molecular biological interpreter

P. R. Wills

Department of Physics, University of Auckland,
Private Bag 92019, Auckland, New Zealand

Email: p.wills@auckland.ac.nz

Abstract

A macromolecular system may be said to be self-constructing if the relationship between structural components and the functional operations they perform to synthesise one another is reflexive. Reflexivity is a property of the catalytic structure-function relationship, that is, the embedding of catalytic functions in the space of polymeric structures. An operational algorithm for determining functions from structural information, such as the genetic code, is an interpreter. Molecular biological interpreters are continuously inherited products of historical processes of self-organisation. The evolution of interpreters is the hallmark of increasing functional complexity in self-constructing molecular biological systems.

1. Introduction

Many modern discussions of the molecular origins of life focus on the problem of molecular replication and the accumulation of genetic information in prebiotic systems. Consequently attempts to explain the evolution of functional complexity in molecular biological systems have often been caught up with considerations of the mutation and selection of information (Eigen, 1971; Eigen and Schuster, 1979). More recently the intrinsically autocatalytic character of collective biochemical processes has been elaborated as the physico-chemical basis of life's origin (Kauffman, 1993), but there has been scant attention paid to the conceptual divide between these alternative visions. Here I present an overview and interpretation of the results of collaborative work, carried out during the last decade or so, that has aimed to find a unified framework for understanding some of the central problems in theoretical biology. Much has of necessity been left out, but it is hoped that the presentation will contribute to the efforts of others to develop a new framework for elucidating the character of molecular biological processes.

Received: 15 Jul 2000
Accepted: 01 Nov 2000

Section 2 commences by making a simple distinction between the material entities and temporal events that make up biochemical processes. Some special features of the relationship between the structural and dynamic features of a system that allow it to be self-constructing are described. In Section 3 these special microscopic properties are analysed in terms of the structure-function relationship of functional biochemical components, with particular reference to proteins. In Section 4 the idea that this relationship can be thought of as an algorithm that is executed by a molecular biological *interpreter* is introduced. There is a discussion of how protein folding and genetic coding fit into this conception. In Section 5, autocatalytic selection and the decomposition of a molecular alphabet are offered as crude examples of the emergence or refinement of an interpreter and as providing an explanation of the evolution of molecular biological complexity. Some implications are discussed in Section 6.

2. Self-Construction and Selection

We consider natural dynamic systems in which we can identify operations $\{O_i: i = 1, 2, 3 ... N\}$ that are carried out by a molecular component or components $\{M_j: j = 1, 2, 3 ... M\}$. The rates $\mathbf{w} = (w_1, w_2, \ldots. w_N)$ at which operations O_i can be carried out depend on the population numbers $\mathbf{x} = (x_1, x_2, \ldots. x_M)$ of the various components:

$$w_i = f_i(\mathbf{x}) \tag{1}$$

We further restrict consideration to systems in which any component can only be synthesised as a result of system operations being performed, in which case the rate of synthesis of each component has the dependence

$$dx_j/dt = g_j(\mathbf{w}) \tag{2}$$

The functions $\mathbf{f}(\mathbf{x}) = \left(f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots f_M(\mathbf{x})\right)$ imply the *embedding* of functional operations O_i in the field of molecular components M_j. The simplest situation occurs when the embeddings are specified by a set of vectors $(w_{i1}, w_{i2}, .. w_{ij} ..w_{iM})$ in which $w_{ij} = w^o$ if the molecular component M_j can perform the operation O_i, or zero otherwise, and f_i is the scalar product. In that case

$$w_i = \Sigma\, w_{ij}\, x_j \tag{3}$$

is the overall rate at which O_i is performed within the system. In similar vein, the functions $\mathbf{g}(\mathbf{w}) = \left(g_1(\mathbf{w}), g_2(\mathbf{w}), \ldots g_N(\mathbf{w})\right)$ imply a set of constructive outcomes from the joint performance of some sequences of operations. The simplest situation occurs when a unique sequence (of length ν) of operations produces a unique component of the system and g_j is a simple product of magnitudes. In that case, with certain other conditions satisfied,

$$dx_j/dt = \Pi w_{n_j(k)} \tag{4}$$

where the $n_j(k) \in \{1, 2, \ldots N\}$ are the sequence of operations ($k = 1, 2, \ldots \nu$) needed to construct the component M_j. For $\nu=1$ Eq. (4) reduces to a replicator-type equation, giving the quasi-species rate equations of Eigen (1971) for the linear growth rates specified by Eq. (3).

We say that selected subsets of operations O_i and molecular components M_j are related to one another *reflexively* and the system is *self-constructing* if the mathematical forms of **f** and **g** are such that there are dynamic fixed points of Eqs. (1) and (2) satisfying certain conditions.

1. All of the selected components required to carry out the selected operations can be constructed by carrying out operations from within the selected subset of operations.

2. All of the operations required to construct the selected components can be carried out by components from within the selected subset of components.

The emergence of a self-constructing molecular system through dynamic selection requires the existence of a reflexive relationship (arising from the mathematical forms of the **f** and **g**) between some operations and components. Reflexivity is a formal, semiotic relationship between operations and components that is dependent on physical principles at most only implicitly. Criteria for the existence of reflexivity in systems defined by Eqs. (3) & (4) have been discussed in relation to molecular biological coding (Wills, 1993; Wills, 1994; Nieselt-Struwe & Wills, 1997) and ligation autocatalysis (Wills & Henderson, 1997; Wills *et al*., 1998). A reflexive relationship between a set of polymers {*aba*, *aab*, *baa*} and the ligation reactions required for their synthesis {*a*+*a* ¬ *ab*, *a*+*b* ¬ *ab*, *b*+*a* ¬ *ba*} is illustrated in Figure 1. Note that the simple replication of *bbb* through catalysis of the reaction *b*+*b* ¬ *bb* also instantiates a reflexive relationship between structure and function.

3. Structure-Function Relationships

In order to determine that the condition of reflexivity is satisfied in relation to a subset of operations and components comprising a self-constructive system, it is necessary to know the form of the embedding of operations in the field of material components. An embedding is a specification of a *structure-function* relationship. Knowing the form of the embedding, and thus **f**, amounts to the specification of a rule (or rules) for determining whether or not a component with a certain structure has the functional capability of performing any of the operations. We are most familiar with this idea in relation to the catalytic properties of proteins. The primary structure (amino acid sequence) of a protein and its folding under standard conditions are determinants, in some cases practically complete determinants, of the protein's three-dimensional molecular structure and its functional catalytic properties. From a broad-brush point of view it is reasonable to think of the protein structure-function relationship as a set of catalytic functions embedded in a sequence space. The embedding, in a 3-dimensional binary sequence space *xyz*, of catalysis of the four simple ligation reactions *x*+*y* ¬ *xy* that can occur between the two monomers comprising a binary alphabet {*a*, *b*}, is illustrated in Figure 1.

For a polymer sequence space of resonable size (having λ^{ν} points, where λ is the number of letters in the monomer alphabet and ν is the polymer length) and a limited number N of operations, the number of embeddings (structure-function relationships) that are conceptually possible is hyper-astronomical (Nieselt-Struwe & Wills, 1997). On the other hand, it could be argued that only one such is consistent with, in fact dictated by, the laws of physics and chemistry. In terms of our example of proteins, the laws of quantum and statistical mechanics determine how a protein with a particular amino acid sequence folds and what reaction(s), if any, it is able to catalyse. From the fact that complex self-constructing systems (biological organisms) exist, we can conclude that a reflexive relationship exists between certain physico-chemical components and operations in the quantum-thermodynamic world that we occupy. One of the tasks of theoretical biology is to describe that relationship of reflexivity and how it has been manifested at various points in the process of biological evolution, starting with the origin of life.

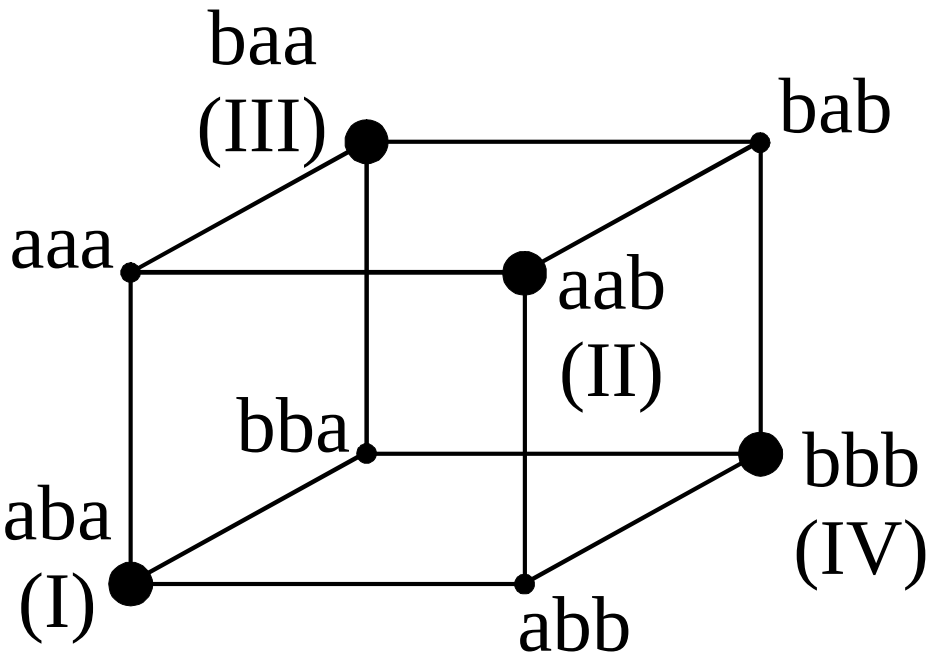

Figure 1. A three-dimensional sequence space of polymers xyz *comprised of letters from a binary alphabet {*a, b*} is represented as a cube. Specified polymers* xyz *catalyse one of the four possible ligation reactions* x+y ¬ xy *between isolated monomers* x *and* y, *representing the set {*a+a, a+b, b+a, b+b*} according to enumeration of the set with Roman numerals {I, II, III, IV}. The subsets of structures {*aab, aba, baa*} and reactions {*a+a, a+b, b+a*} are related to one another reflexively. The structure-function relationship can be described by an algorithm: reaction* x+y ¬ xy *is catalysed by* xyz *if and only if* x#z *where "#" means "≠" if* y=a *and "=" if* y=b.

However, we need not take the point of view that the relationship of reflexivity between biological structures and functions that makes possible the very existence of organisms is a simple feature of macromolecular structures and properties. The three-dimensional structures of proteins and the effects of their interactions with one another and other types of molecules in cells cannot generally be regarded as autonomous properties of the molecules themselves, perhaps ascribable to their primary sequences alone. Rather it is found that proteins have many possible folded states and that special folded states of some proteins are selected as a result of dynamically controlled processes that take place during or after their synthesis. The cellular properties and functions of such proteins are determined through interactions that depend on the populations of other molecules, including other proteins, already present in the cell. This means that the biological structure-function relationship of proteins is contingent upon the state of the cell in which the proteins in question are synthesised and is not, as we may have thought at first, some general, immediate consequence of the quantum mechanics of isolated molecules, characterised only by their primary structures. An extreme example of the mutability of the protein structure-function relationship is provided by the existence of prion protein variants which are determinants of not only their own function but also the phenotype of the cells in which they occur.

4. Molecular Biological Interpreters

Rewriting Eqs. (1) & (2) in the alternative forms

$$\mathbf{w} = f\left(\int g(\mathbf{w})dt\right) \tag{5a}$$

$$d\mathbf{x}/dt = g\left(f(\mathbf{x})\right) \tag{5b}$$

allows the analytical difficulty due to the nonlinearity in the dynamics of self-constructing systems to be perceived. It would be of some advantage if we could describe the manner in which **f** and **g** are related, an algorithm for determining the rates at which components M_j of the system perform operations O_i based on information about the components and their construction.

If the folded three-dimensional structure and all of the biological properties of a protein in its cellular milieu were determined solely by its covalently bonded structure (primary sequence), then we could regard the quantum-thermodynamic rules that govern the folding process as a kind of *interpreter* specifying a structure-function relationship. Given an amino acid sequence, which can be regarded as information, the interpreter translates it into a cellular function, instantiated as the folded protein. The structure-function relationship illustrated in Figure 1 can be specified in terms of an algorithm which is an elementary example of an interpreter: reaction $x+y \neg xy$ is catalysed by xyz if and only if $x\#z$ where "#" means "$\neq$" if $y=a$ and "=" if $y=b$.

There has been an enormous, largely unsuccessful, research effort, especially during the last two decades, to find an algorithmic description of nature's protein-folding interpreter more simple than the laws of quantum-thermodynamics. It cannot be determined whether such a description will ever be achievable and should it prove impossible it would be reasonable to abandon use of the term "interpreter" in this context because the mapping from information to folded structure would then appear to be almost totally arbitrary and of irreducible complexity, bearing none of the hallmarks of the processes of language. The algorithm describing the structure-function relationship illustrated in Figure 1 qualifies as an example of an interpreter because it can potentially be specified using about 3 bits of information, which is considerably less than the approximately 9 bits required to specify an arbitrary embedding of 4 functions in a sequence space of 8 points.

However, if we look at protein folding in cells we find that it does not proceed spontaneously as a simple thermodynamically driven process. Rather, it is regulated by the presence of various cofactors, usually called chaperones. From the point of view of cellular function there is a "correctly" folded form of a protein which in many cases can only be produced through the intervention of the chaperones. It is not unreasonable to expect that a general description of the action of chaperones could be constructed in which case they could be thought of as parts of the machinery that operates as a protein-folding interpreter. What we notice is that we are now talking about cellular components, pieces of matter that have to be constructed, not physical principles, that act as an interpreter, or part thereof. This is true of chaperones, but the archetypal molecular biological entities that collectively act as an interpreter are the components (ribosomes, amino-acyl synthetases, tRNAs, initiation factors, etc.) that comprise the machinery of translation. The language of interpretation is given by the assignments from nucleotide codon triplets to amino acids, the genetic code, supplemented by the mRNA motifs that function to specify translation initiation and termination sites.

In recent years many colleagues and I have concerned ourselves with attempting to demonstrate the means whereby interpreters come into existence and increase in complexity. We have shown, for example, that if protein production is nucleic acid sequence-dependent

and requires protein assignment catalysts (like the amino acyl synthetases) then fully coded protein production can arise spontaneously in systems where protein production is initially completely random (Wills, 1993; Wills, 1994). What is required is that the available nucleic acid information has the property of reflexivity with respect to the protein structure-function relationship. It is necessary that the proteins that effect the assignments belonging to the chosen code (as opposed to proteins that effect the many possible assignments not belonging to the code) construct themselves when they interpret the available nucleic acid information. All organisms have genomes that have this property of reflexivity because when the machinery of translation acts on the information provided to it, it constructs itself. The required property of reflexivity is not intrinsic to any nucleic acid sequence. It is not encoded in the genes. It is contingent upon the operational mechanism of protein synthesis and the chemical properties of the protein products involved in the process (Nieselt-Struwe & Wills, 1997).

Even further, it is important to emphasise that the structure-function relationship with respect to which information may be said to be reflexive is not a simple mapping between sequences and operations, because the functionally significant chemical properties of a polymer with a particular sequence, including catalytic capabilities, may depend on how the folding of that sequence is influenced by other protein components of the system. The biological structure-function relationship is not something that is dictated altogether in advance by physical principles. Rather, it is a historical product of many events of self-organisation that have occurred during evolution.

We cannot realistically hope ever to specify fully all of the elements of the molecular biological interpreter which confers the property of reflexivity on an organism's genome. Organisms' molecular biological interpreters are intricate products of millennia of evolution. In a sense, they evolve independently of genomes, adapting to whatever information is available to them, often conferring selective advantage or disadvantage on mutations that occur in the genomes with which they are associated. Thus, trying to unravel the complete algorithm of a cell's molecular processes would be like trying to deduce all of the mutations that have occurred since two disparate species had a common anscestor and the historical reasons for each selection event.

5. Evolution of Molecular Biological Complexity

These considerations lead us to the following conclusion. While it is evident that cells are self-constructing in terms of covalently bonded molecules and the processes of metabolism, including RNA-dependent protein synthesis, the reflexivity between selected operations and components existing at this level of molecular structure does not exhaustively determine the biological identity of a system. In some cases, the reflexivity required for self-construction could be dependent on the arbitrary historical contingency of past events of selective self-organisation. It could exist in the realm of very subtle molecular features and interactions that are not accessible to simple biochemical or genetic-informational description. We then have to ask what molecular features could ever qualify as constituting the biologically significant structural "components" of a cell and what sort of induced changes could ever be defined as a biologically relevant "operations" on those components, bearing in mind that quantum mechanics provides an irreducible description of the states of any isolated molecular subsystem.

The traditional approach to the understanding of evolution in biological systems has been to find the form of functions **f** and **g** that provide a satisfactory representation of a system's

dynamics and the processes of selection that take place within it. What we now seem to be suggesting is that in reality the form of these functions and their relationship to one another (the "interpreter") undergoes evolutionary change that could be as significant to biology as the accumulation of encoded information. In other words, we must consider the unsettling possibility that no specification in purely physical terms of functions **f** and **g**, necessitating as they do the definition of "components" and "operations", can give a complete description of a biological system, because what qualifies as a "component" or "operation" depends on the context and the context changes as the biological system evolves. [This problem occurs in different forms in the work of Rosen (1991), Pattee (1997) and Kauffman (*in press*).]

What simple concepts might be useful for describing the evolutionary development of a biological interpreter? One approach has been to define classes of components and operations that undergo progressive internal differentiation as a result of dynamic selection. Such an approach has been applied to the problems of coding evolution (Nieselt-Struwe & Wills, 1997; Wills, *in press*) and ligation autocatalysis (Wills & Henderson, 1997). In essence one considers an alphabet of subcomponents (for example, a binary alphabet, {*a*, *b*}) which undergoes progressive decomposition (perhaps first to a ternary alphabet, {α, β, γ}). This process of selection, if it occurs, allows the degree of differentiation of both components and operations, and thus the internal functional complexity of the system, to increase. Actually, the initial differentiation of subcomponents, {*a*, *b*}, as distinguishable members of a general class can be regarded as decomposition of a unitary alphabet. As discussed elsewhere (Wills, *in press*), the autocatalytic process of selection that constitutes the dynamics of decomposition requires that there be a reflexive relationship between the subsets of structures and functions in terms of which the decomposed alphabet of subcomponents is eventually defined. Outstanding problems concerning the dynamic stability of systems in which mutable information is interpreted to produce catalytic function are currently being addressed. It has been demonstrated recently that replicating molecular information can be stably maintained by functions to which it is transmitted in systems that are spatially differentiated as a result of simple diffusion processes (Altmeyer *et al*., 1999).

The reflexive relationship between components and operations that allows selection to take place as a result of autocatalysis cannot always be formulated in such simple terms, but the simple examples are of heuristic value and therefore worthy of analysis. Furthermore, it is only possible to give an algorithmic description of the relevant reflexive relationship characterising the structure-function embedding (and therefore speak of an "interpreter") if it is possible to describe components and their construction in terms of a *class* (or alphabet) of more elementary structural features. There could be no such thing as genetic information *per se*, nor a code for its translation, unless nucleotides and amino acids each constituted a finite class (or alphabet) of differentiated substructures comprising types of cellular components (nucleic acids and proteins).

The decomposition of an elementary class of molecular structures can occur only as a result of autocatalytic selection. Selection of some possibilities and the exclusion of others represents an *ordering*, in this case very crude, of the performance of possible operations, so that the components that carry out the selected operations are constructed and other components are not constructed. A system in which the performance of operations is ordered carries information to the extent that the ordering determines a sequences of choices, or selection among alternatives, that is made at each step. A fully-fledged molecular biological interpreter (like the machinery of translation operating according to the rules of the universal code) provides for a unique "information-based" choice to be made (one amino acid from among twenty) at each step in a defined sequence of molecular operations.

6. Conclusion

The overarching conceptual problem faced by the theoretical biologist is that in the world of real cellular biochemistry classes of entities are constituted operationally and the operations that constitute classes are themselves the historical products of selection. Nucleotide triplets and amino acids form classes between which a mapping corresponding to the genetic code can be specified only because amino acyl transferases, ribosomes and the other components of the machinery of translation (that operationally differentiate between members of both classes of molecular structures - nucleic acid codons and amino acids) exist and endure as continuously replicated products of evolution. The components of the interpreter and the operations they perform are what constitutes information as it resides in nucleic acid strands. Understanding the origin of life and its increasing complexity requires that we enquire into the semiotics of molecular interpreters, not just the mechanisms of molecular replication.

Acknowledgements

The author is grateful to Stuart Kauffman, John McCaskill and Lee Smolin for helpful discussions.

References

Altmeyer S., Füchslin R., Tangen U., Ackermann J. & McCaskill, J.S. (1999), *Molekulare Konstructionssyteme - die Evolution eines Codes*, Der GMD-Spiegel 3/4, 11-13, URL: http://www.gmd.de/de/GMD-Spiegel/Spiegel3-499/Sites/02HMasthis.html

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

Eigen M. & Schuster, P. (1979), *The Hypercycle* (Springer-Verlag, Berlin)

Kauffman S.A. (1993), *The Origins of Order* (Oxford University Press, New York)

Kauffman S.A. (in press), *Investigations* (Oxford University Press, New York)

Nieselt-Struwe K. & Wills P.R. (1997), *The Emergence of Genetic Coding in Physical Systems*, J. Theor. Biol. 187, 1-14

Pattee H.H. (1997), *Causation, control, and the evolution of complexity*, in: Downward Causation, P.B. Anderson, P.V Christiansen, C. Emmeche, & N.O. Finnemann (eds.)

Rosen R. (1991), *Life Itself* (Columbia University Press, New York)

Wills P.R. (1993), *Self-organisation of genetic coding*, J. Theor. Biol. 162, 267-287

Wills P.R. (1994), *Does Information Acquire Meaning Naturally?*, Ber. Bunsenges. Phys. Chem. 98, 1129-1134

Wills P.R. (in press), *Autocatalysis, Information and Coding*, BioSystems

Wills P.R. & Henderson L. (1997), *Self-organisation and information-carrying capacity of collectively autocatalytic sets of polymers: ligation systems*, NESCI Interjournal, Ms 102, URL: http://interjournal.org

