Fundamentals of Life

Chapter 1.2

Is life reducible to complexity?

David L. Abel

What exactly is complexity? Is complexity an adequate measure of 'genetic instructions' and 'code'? How do complex stochastic ensembles such as random biopolymers come to 'specify' function? All known life is instructed and managed by bio-information. The first step in understanding bio-information is to enumerate the different types of complexity. Since biopolymers are linear sequences of monomers, emphasis in this chapter is placed on different types of sequence complexity. Sequence complexity can be 1) random (RSC), 2) ordered (OSC), or 3) functional (FSC). OSC is on the opposite end of the spectrum of complexity from RSC. FSC is paradoxically close to the random end of the complexity scale. FSC is the product of non-random selection pressure. FSC results from the equivalent of a succession of algorithmic decision node 'switch settings.' FSC alone instructs sophisticated metabolic function. Self-ordering processes preclude both complexity and sophisticated function. Bio-information is more than mere complexity or a decrease in comparative uncertainty in an environmental context. Life is also more than the self-replication of gibberish. Life is the 'symphony' of dynamic and highly integrated algorithmic processes which yields homeostatic metabolism, development, growth, and reproduction. Apart from our non-empirical protolife models, algorithmic processes alone produce the integrated biofunction of metabolism. All known life and artificial life are program-driven. Shannon-based 'information theory' should have been called 'signal theory.' It cannot distinguish 'meaningful' signals from gibberish. In biology, meaningful signals are metabolically functional signals. Shannon theory lacks the ability to recognize whether a sequence is truly instructional. It cannot distinguish quantitatively between introns and exons. Nucleic acid is the physical matrix of recordation of the switch settings that constitute genetic programming. Progress in understanding the derivation of bioinformation through natural process will come only through elucidating more detailed mechanisms of selection pressure 'choices' in biofunctional decision-node sequences. The latter is the subject of both 'BioFunction theory' and the more interdisciplinary 'instruction theory'.

1. Introduction

At a colloquium at NASA's Goddard Space Flight Center near the beginning of 2000,[1] Freeman Dyson defined 'life' as "a material system that can acquire, store, process, and use information to organize its activities." This refreshing distillation of life was especially noteworthy since Prof. Dyson for many years has been a proponent of the 'metabolism first' model (or at least a two-step model) rather than an 'information first' model.

Hubert P. Yockey has been investigating the meaning of complexity with mathematical precision for forty years. His recent distillation of life is also rather remarkable:

"It is highly relevant to the origin of life that the genetic code is constructed to confront and solve the problems of communication and recording by the same principles found both in the genetic information system and in modern computer and communication codes. There is nothing in the physico-chemical world that remotely resembles reactions being determined by a sequence and codes between sequences. The existence of a genome and the genetic code divides living organisms from non-living matter.[2] "

Thirty years ago Leslie Orgel wrote[3a]**:**

It is possible to make a more fundamental distinction between living and nonliving things by examining their molecular structure and molecular behavior. In brief, living organisms are distinguished by their specified complexity[a]. Crystals are usually taken as the prototype of simple, well-specified structures, because they consist of a very large number of identical molecules packed together in a uniform way. Lumps of granite or random mixtures of polymers are examples of structures that are complex but not specified. The crystals fail to qualify as living because they lack complexity; the mixtures of polymers fail to qualify because they lack specificity."

This thirty-year-old passage suggests that Orgel may have been using specified to mean 'deterministically ordered,' 'uniform,' or 'repetitive.' But clearly nucleic acid and protein polymers do not fit such a description. If they were deterministicallyordered, uniform, or repetitive, they could not retain any significant amount of information in their physical matrix. It is the freedom of association and resortability of monomers that make biopolymers ideal information retaining matrices. An update by Prof. Orgel of his current perspective on specified complexity would be most welcome and helpful.

Later in Orgel's book[3b] he goes on to describe information as it
relates to complexity:

"Roughly speaking, the information content of a structure is the minimum number of instructions needed to specify the structure. One can see intuitively that many instructions are needed to specify a complex structure. On the other hand, a simple repeating structure can be specified in a rather few instructions. Complex but random structure, by definition, need hardly be specified at all.

Orgel was right in differentiating mere complexity from specified complexity. He was also right in questioning the need to enumerate and quantify gibberish. Unfortunately, information theory continues to ignore the meaning and significance of 'specification.' The instructive aspect of information was sacrificed in our zeal to mathematically quantify complexity and signal transmission

[a]. It is impossible to find a simple catch phrase to capture this complex idea. "Specified and, therefore, repetitive complexity" gets a little closer (see later).

success. Current information theory is purely mathematical and oblivious to function. Its applicability to biology is therefore limited. A whole new approach is needed to link Kolmogorov-Chaitin-Yockey mathematical sequence theory with metabolic instructions.

The environment selects for neither complexity nor code. Selection pressure favors what 'works.' It is not clear how a prebiotic environment could select for anything beyond mere molecular stability or self-replication. But the rise of any protometabolism would have depended on selection for protophenotypic function, not linear symbol sequences. Self-replication is only one type of function. Any primordial metabolism would have depended largely on molecular shapes. While our ability is poor (65 % success rate[4]) to predict three-dimensional shapes from sequence data, we can still afford to ignore Yockey's above-quoted point. The sequence of amino acids results directly from translated code of a completely different set of 'alphabetical' symbols. Problematic as it is, responsible investigators must address the reality of the specification provided by nucleic acid symbol sequences.

In molecular biology, message 'meaning' corresponds to 'biofunction.' As Yockey has pointed out many times,[2,5] current information theory compartmentalizes meaning from its measure of transmission 'success.'[6] Herein lies the problem with the application of so-called information theory to biology. We can ill afford to isolate 'biofunction' from life-origin theory and bio-informatics. Bio-information cannot be quantified independent of its meaning (the function it instructs). To do so would render all nucleic acid strands/segments of the same length, both exons and introns, equal in quantifiable information. It would be a mistake to equate introns with gibberish or junk. Our understanding is still too limited. But if base ratios were equal, a totally random sequence of nucleotides would contain the same number of bits of 'information' as an exon with the same number of nucleotides.

For complexity to specify structure or function, it needs to guide decision node switch settings toward a productive endpoint. Only one of the three kinds of sequence complexity described below does this: functional sequence complexity (FSC). Biochemical pathways are instructed by strings of decision node selections recorded in biopolymer strands, The key to life-origin research lies in uncovering the mechanisms whereby these productive algorithmic programming choices were made and recorded in nucleic acid.

2. What is complexity?

There are many kinds of complexity. They can usually be grouped into two general classes: (1) static and (2) dynamic. Static complexity pertains to physical and structural arrangements or states. Dynamic complexity reflects the degree of computational effort required to describe and know the uncertainty reflected by that object or state.

It is unfortunate that the term 'dynamic' was seized to describe our human computational manipulations. Dynamic complexity ought not to be limited to human epistemology. The term dynamic could also be used to describe a kind of natural process complexity (e.g., algorithmic biochemical pathways and metabolism) that exists in nature independent of our knowledge and mental mathematical gymnastics.

Life-origin research presupposes that physical molecules interacting through chemical evolution spontaneously acquired a state of 'organized' complexity sufficient to produce protolife. Such an event would have predated humans, their consciousness, and their computational pursuits by at least[7] 3.8 to $4.4.10^9$ yr.[8] Dynamic computational complexity is a mental construction of *Homo sapiens*. It therefore has no causative relevance to life origin. Whatever the physical mechanisms of self-organization, they were independent of afterthe-fact human cognitive analysis. Human knowledge applications such as pattern matching, data compression, and mining had no role in the dynamics of life origin. All that matters is the question, "How did inanimate physicality acquire sufficient where-with-all to self-organize algorithmic metabolism? How did inanimacy generate integrative genetic instructions?"

In living organisms, static complexity exists not only in the form of linear biopolymers, but also in the form of their three-dimensional shapes. Other aspects of static complexity include architecture of such structures as the cell membrane and microtubules. Thus, some might question defining static complexity solely in terms of linear sequence theory. Why not define complexity in terms of three-dimensional shapes, for example? In addition, the primacy of amino acid sequencing has been questioned. Amino acid substitutions can sometimes be made without destroying tertiary structure and function. Occasionally completely different sequences, both homologous and non-homologous, can yield roughly the same shape, grove, and function.

Despite the above considerations, amino acid sequences are nonetheless instructed in the ribosomes by linear, segregatable, digital sequences of nucleotides in mRNA.[2] Thus, the primary structure of proteins (the sequence of amino acids) is determined by the sequence of nucleotide codons. While exceptions have been pointed out,[9] the 'central dogma' of molecular biology remains fundamentally intact.[10] Regulatory proteins and prions notwithstanding, directive information normally flows from nucleic acid to proteins. Linear sequences having one alphabet dictate the linear sequences of a completely different alphabet. The genetic code stands in between. This code conceptually relates one sequence to another.

The tertiary structure of proteins is primarily determined by amino acid sequence. Despite the exceptions and problems, in all known molecular biological life, the best approach to defining and quantifying complexity is still through sequence theory. Even regulatory proteins and viral RNA are basically linear sequences. Cell, organ, and body differentiation in eukaryotes, for example, are dictated by linear biopolymers. In prokaryotes, metabolism, growth, and reproduction are all dependent upon instructions encoded in linear 'recipe,' linear 'messenger molecules,' and linear structural and catalytic proteins. It is the sequence of monomers more than any other factor that determines what sophisticated life can and will be. Any attempt to sidestep sequence theory as the major contributor to the complexity of current life is folly.

There has been an explosion of literature in the past decade affirming the primary importance of sequencing in molecular biology. In the genome projects, for example, "sequencing is everything." The importance of sequence complexity has been reemphasized of late by Thomas Schneider's work on molecular machines.[11] Christoph Adami's artificial life research is based on sequence theory with a strong emphasis on environmental contexts of aboutness.[12]

A few notable exceptions exist to emphasizing the linear, se§regatable, digital nature of genetic information emphasized by Yockey.[2,5,10,13] The best example is probably,the life-origin research of Segre and Lancet on 'GARD' and 'compositional genomes.'[14] But it remains to be seen just how much bio-information could be generated and retained in the complexity of nonlinear compositional genomes. Also unexplained thus far is the mechanism of 'genetic takeover' that would make the transition from compositional genomes to empiricallinear genomes. The 'continuity principle' of life-origin research would require such explanation. Failure to provide a convincing mechanism for genetic takeover was a major factor in the stasis of the Cairns-Smith model.[15] Another problem was the failure to explain the derivation of an instructive clay template.

The larger body of literature on complexity, unfortunately, still manifests a chaos of its own. 'Order' is often confused with 'organization' which in turn is confused with complexity. To make matters worse, complexity is then equated with instructional information. Crucial definitions conflict from one paper to the next. Opinions are strong and sometimes almost antithetical. It is easy for investigators of biological complexity and information to study the literature with great diligence, but to come away thoroughly confused as to what is sequence entropy, regular entropy (Maxwell-Boltzmann-Gibbs entropy),[5] order, complexity, instructive information, and organization. The results of this linguistic confusion and cognitive dissonance can be catastrophic.

Information theorists treat an information or data source as though it were a random sequence.[16] This is a perfectly suitable approach for Shannon signal theory. But there is something fundamentally wrong with the entire field of information theory treating information as 'random sequences.' Orgel, we might recall, stated above,[3b] "Complex but random structure, by definition, need hardly be specified at all." In other words, why bother enumerating and quantifying nonsense (non sense)? We could distinguish the bit string of this paper from gibberish based on pattern recognitions. But we would derive no instructional meaning from the signal alone. Agreement on symbolic representation between source and destination would be required for the signal to become a true message. Just as in language, both semantics and syntax are required for bio-information and biomessenger molecules to produce function in a cell.

In mathematics and linguistics, we arbitrarily assign meaning to certain symbols and to certain syntactic alphanumeric sequences. But biology is not arbitrary. Only certain monomeric sequences and the physical shapes they produce work. Certain sequences of nucleotides in mRNA give rise to certain sequences of amino acids. The two sequences are enumerated in completely different languages. They must be coordinated by representational code to translate from nucleotide sequence language into amino acid sequence language/shape. How did nature write this translation code such that the needed three-dimensional shape would result in the new language?

In a paper published on, November 17, 2000, on PNAS USA online,[4] Ronneberg, Landweber and Freeland conclude that the coevolution theory of J.T.F. Wong[17] is not viable: the theory's definition of 'precursor-product' amino acid pairs is unjustified biochemically; the theory neglects important biochemical constraints; and it cannot adequately explain the structure of the genetic code. Another recent paper by Shapiro[18] concludes with surprising dogmatism that a replicator was not involved in the origin of life.

In an Oct-2000-paper by Weiss, Jimenez-Montano and Herzel,[19] the complexity of large sets of non-redundant protein sequences was measured. Both Shannon entropy and compression algorithms were used to determine complexity. Proteins were found to be close to random sequences, with entropy reduction due to correlations "being around 1 %. Compression algorithms also suggested a redundancy of approximately 1 %. The authors concluded that proteins can be regarded mathematically as only slightly edited random strings (strands). This is not a picture of crystal-like, uniform, or repetitive order. And protein specificity is about as far from random as we can get. Something is missing in our description of bio-information and specified complexity and their ability to instruct function.

Among the most perplexing questions for life-origin research is, "How does a certain sequence of nucleotide monomers 'program' amino acid sequence so as to produce needed three-dimensional shapes?" Our poor track record at predicting tertiary structure from primary structure further compounds the mystery. As if this weren't enough of a problem, we must further ask, "How do hundreds, even thousands, of these sequences spontaneously organize themselves into pathways and an integrated metabolic network?"

Only one of the three kinds of sequence complexity can provide sophisticated bio-instructions for function: functional sequence complexity (FSC). To understand why, we must take a close look at all three types of sequence complexity.

3. Random sequence complexity (RSC)

Random sequence complexity can be simplistically defined as a mathematical function of the number of equiprobable potential alphanumeric symbols that could occupy each locus times the number of loci in that sequence of symbols. This function can be easily adjusted to allow for cases in which each 'alphabetical symbol' or monomer is not equiprobable. Such is the situation in prebiotic chemical evolution models of the RNA world. Each nucleotide may not be equally available to fill the next spot in a forming polyribonucleotide.

Random sequence complexity (RSC) has three components:

- the number of symbols in the alphabet that could potentially occupy each locus of the sequence (bit string) (e.g., there are four potential nucleotides that could occupy each monomeric position in a forming polynucleotide);
- the probabilistic availability (frequency) of each symbol to each locus (e.g., the frequency of adenine is not the same as that of guanine, cytosine, or uracil to each position in a randomly forming polyribonucleotide);
- the number of loci in the sequence (e.g., the number of mers must be adequate for a ribozyme to acquire minimal happenstantial function).

The sequence complexity of random alphanumeric symbol sequences can be precisely quantified using straight Shannon theory. No. discussions of aboutness [12] or 'before and after' differences of 'knowledge' [11e, i, ,j k, 1] or 'shared' anything are relevant to a measure of RSC.

It has long since become customary in information theory to refer to RSC as information. The enumeration of any specific sequence of alphanumeric symbols itself becomes a kind of information. We become aware or know what the sequence is by naming and quantifying all of the possibilities the sequence might have been in 'bits.' Such knowledge is then called information. But information theory does not address whether that specific sequence means or does anything. Is the sequence instructive to its environment? Is the strand physically functional? Does it contribute in any way to metabolism, or are we just naming its monomeric sequence and calling that bioinformation?

Bits were supposed to refer to binary choices or switches in an algorithm.

Unfortunately, bits' has come to mean little more than the number of binary possibilities. This does not measure the full meaning of specific instructional information.

The best place to begin to understand real information (instruction) is in a primordial bio-information context. This context is inseparable from proto-metabolism. Functional information conveys message and meaning. It does more than just exist as a specific sequence. Specific does not mean 'specified to function.' Quantifying the number of possibilities only provides less than half of the information we would most like to gain. The real information we desire is, which sequence works? Functional sequences do something useful. They are truly informational (well beyond the Shannon theory definition of informational) because they are either instructive of other biosubsystems or because their sequences provide direct physical catalytic or architectural function.

4. Ordered sequence complexity (OSC)

Ordered sequence complexity is exampled by polymers such as polysaccharides. OSC is so ruled by redundant 'necessity' that it affords the least complexity of the three types of sequences. The mantra-like matrix of OSC has little capacity to retain information. OSC would limit so severely information retention that the sequence could not direct the simplest of biochemical pathways, let alone integrate metabolism.

Appealing to 'unknown laws' as life-origin explanations is nothing more than an appeal to cause-and-effect necessity. The latter only produces OSC with greater order, less complexity, and less potential for eventual information retention *(Table I).* In addition, appealing to unknown laws as a mechanism for information generation is a logical category error. Laws don't cause anything. They are merely human mental constructions. Laws are cognitive generalizations of human minds. They have no physical agency to produce effects or to serve as mechanisms.

Table I. The difference between sequence 'order' and 'complexity'.

Order	Complexity
Regular	irregular
repeating	nonrepeating
redundant	nonredundant
predictable	nonpredictable
symmetrical	asymmetrical
periodic	aperiodic
monotonous	variable
crystal-like patterning	linguistic-like patterning°
reducible	largely irreducible
compressible	noncompressible*

°Random complexity lacks true patterns. *Linguistic-like patterning permits some degree of compressibility. Random complexity does not. The paradox of Kolmogorov-Chaitin-Yockey algorithmic information theory is that orderliness lies at the opposite end of the complexity scale from information. Even more paradoxical is that random complexity contains the maximum number of non-c:ompressible bits of information. Here information is defined in the tradition of Shannon, ignoring meaning.

Self-ordering is often confused with self-organizing. *Table /I* provides helpful clarification of the differences between the two. All known life depends upon genetic instructions. While we have many different metabolism first and twostep models of life-origin, no hint of metabolism has ever been observed independent of an oversight and management information system. We use the term bioengineering for good reason. Holistic, sophisticated, integrative processes such as metabolism don't just happen stochasticallb". Self-ordering does. But the dissipative structures of Prigogine's chaos theory[20] are a far cry from the kind of self-organization that would be required to generate genetic instructions or stand-alone metabolism. We can hypothesize that metabolism '~ust happened,' independent of directions, in a prebiotic environment $3.9 \cdot 10^9$ yr ago. But we can hypothesize anything. The question is whether such hypotheses are plausible.

Random sequences are theoretically the most complex (the least compressible). Yet empirical evidence of randomness producing sophisticated functionality is virtually nonexistent. Neither RSC nor OSC possesses the characteristics of informing or directing highly integrative metabolism. Bits of complexity alone can not adequately measure functional (meaningful) bioinformation. Information theory is not succeeding in quantifying the kind of information on which life depends. We call it information, but in reality all we are quantifying is Shannon's signal complexity. It is true that sophisticated bio-information involves considerable complexity. But complexity is not synonymous with bioinformation.

Bio-information has been selected to instruct metabolic function. Apart from actually producing function, information has little or no value. No matter how many bits of possible combinations it has, there is no reason to call it

Table II. The difference between spontaneously "self-ordering" and "self-organizing" systems in nature.

SELF-ORDERING	SELF-ORGANIZING
Increases redundancy	Decreases redundancy
Increases predictability	Decreases predictability
Increases symmetry	Decreases symmetry
Increases periodicity	Decreases periodicity
Increases monotony	Decreases monotony
Produces crystal-like patterns	Produces linguistic-like patterns
Decreases complexity	Increases complexity
Short-lived (highly dissipative)	Long-lasting (minimal dissipation)
Produced by cause-and-effect	Still lacking natural process mechanism
Observed	Unobserved
Consistent with 2nd Law	Seems inconsistent with the 2nd Law
Non integrative	Integrative
Non conceptual	"Conceptual"
Not particularly functional	Produces extraordinary function

information if it doesn't at least have the potential of effecting something useful. What kind of information produces function? In computer science, we call it a program. Another name for computer software is an algorithm.

5. Functional sequence complexity (FSC)

Functional sequence complexity (FSC) is a succession of algorithmic selections leading to function. Bits of functional information represent binary choices at successive algorithmic decision nodes. Algorithms are processes or procedures that produce a needed result, whether it is computation or the end-products of biochemical pathways. Such strings of decision node selections are anything but random. And they are certainly not self-ordered by ~edundant cause-andeffect necessity. Every successive nucleotide is a meaningful quaternary switch setting effected by selection pressure. There is a cybernetic aspect of life processes that is directly analogous to that of computer programming. We are not paying enough attention to the reality and mechanisms of selection at the decision nodes of very real biological algorithms.

Unfortunately, bits of information have come to represent nothing more than a measure of sequence possibilities. And decision node choices have come to mean the equivalent of nothing more than coin tosses of probabilistic or combinatorial uncertainty. Shannon theory has no way of distinguishing options that produce function from those that do not produce function. Shannon theory cannot address the utility (or lack thereof) of any sequence. Bits are currently being used to simultaneously describe both RSC and FSC. Such a state of affairs is insufferable. The result is 'a fool's errand.' We wind up dutifully quantifying meaningless nonsense as though it were equal in bit value to truly instructional information. Shannon theory of course has great utility in many areas. But it remains anemic without supplementation.

In cybernetics, there is never any question as to the connection between malfunction and a programming bug. To fix the problem, we always look for a bad choice at some decision node. In biology, the mechanism of selection at each decision node is far more obscure. But malfunction still results from the same basic cause: a defective choice in the algorithm at some decision node.

To quantify functional bits, or 'fits,' we would need special new equations.

But notice that mathematics alone can never tell us which sequence of decision node choices works. We cannot gain the most important piece of information computationally. The situation is similar to Godel's 'incompleteness theorem.' The question of meaning is undecidable within Shannon theory. We can only gain the information we need (which ensembles function) from arational, nonmathematical sensory observation external to computation. Empirical input into our knowledge provides the missing component of FSC and functional information. Without empirical input, we don't even know that one branch of the algorithmic dendrogram works! For our knowledge to be fully informed about the nature of instructive information (such as genetic information), the particular sequence must be specified through confirmatory phenomenological experience. Such sensorY experience alone resolves 'the halting problem' of computer science.[2,5,21] We cannot know whether an algorithm will complete a computation without actually running it to see if it halts. Observation alone provides the missing element to complete the functional information picture. Current information theory fails to explain the phenomenon of instructions.

Quantification using Shannon theory does not give us what we most want and need. Instructive complexity must signify (to use Shannon's term)[6] or specify (to use Orgel's term),[3a] or select (to use instruction theory and BioFunction theory's term). Not just any complexity will do. More specifically, OSC and RSC will not do to describe or quantify sophisticated bioinformation. Genetic instructions are true algorithmic programming-sequential switch settings that alone produce biofunction.

Selection, specification, or signification of choices in FSC sequences result only from nonrandom selection pressure. We need a way of measuring selection value. Functional bits - 'fits' - are only those bits that measure selected sequences with a known specific metabolic role. But to avoid losing the benefits of current Shannon theory, fits are measured by a ratio of sequences that work to the total number of possible sequences that could occur times the number of Shannon bits contained in that sequence. The 'certainty' of known biofunction is weighted against Shannon 'uncertainty.' Presentation and derivation of these equations are presented in a separate journal manuscript. But it is important to note that FSC can only be quantified in fits relative to a certain specified function in a certain environment.[12b, c] And we can only quantify the ratio based on those sequence combinations that are thus far known to be functional.

The uncertainty (H) of Shannon is an epistemological term. It is an expression of our 'surprisal'[22] or knowledge uncertainty. But humans can also gain definite after-the-fact empirical knowledge of which specific sequences work. Such knowledge comes closer to certainty than uncertainty. More often than not in everyday life, when we use the term information, we are referring to a relative certainty of knowledge rather than uncertainty. Stand-alone Shannon equations represent a very limited knowledge system. But functional bio-information is ontological, not epistemological. Genetic instructions exist in objective reality independent of any knowers. The twentieth century was marked by a consistent problem. Quantum paradoxes left us confusing our own sentience and epistemology with external objective reality. The two were particularly smeared into oneness by postmodern 'science' and quantum quackery. Life origin investigators, of all scientists, should be most critical of this error. Primordial life was not subject to nor affected by human consciousness, observation, or knowledge.

Shannon uncertainty is actually more of a measure of the lack of knowledge.

The closest Shannon information comes to positive knowledge is when the difference is taken between before and after uncertainties. Schneider [11 l] considers this reduction in uncertainty to represent an increase in positive knowledge. But we are seeking a more ontological definition and quantification of genetic instructions. Biopolymers directed metabolism before we existed. FSC matrices provide a highly specific object of our knowledge and relative certainty. But FSC is far more than a reduction in humans uncertainty. It is an effectual physical sequence in its own right. It knows nothing itself because it is inanimate, but it does plenty by virtue of its sequential switch settings ('decision node selections').

Stochastic ensembles could happenstantially acquire functional sequence significance. But a stochastic ensemble is more likely, by many orders of magnitude, to be useless than accidentally functional. Apart from nonrandom selection pressure, we are left with the statistical prohibitiveness of a purely chance metabolism and spontaneous generation.

Shannon's uncertainty equations alone will never explain this phenomenon. They lack meaning, choice, and function. FSC, on the other hand, can be counted on to work. FSC becomes the object of our relative epistemological certainty. Its quantification in fits is based on the fact of its known function. Its specifically enumerated sequence coupled with observed function is regarded as the equivalent of a proven halting program.

Selection is exactly what is found in computer algorithms. Correct choices at each successive decision node alone produce sophisticated software. RSC strings are pragmatically distinguished from FSC strings by virtue of the fact that RSC strings are almost never observed to do anything useful in any context. FSC strings, on the other hand, can be counted on to contribute specific utility.

Yockey's mathematical precision is more than commendable. But the sequence theory of Koimogorov-Chaitin-Yockey fails to provide explanation of algorithmic programming of biochemical pathways. Koimogorov-ChaitinYockey sequence theory is called algorithmic because it employs compression algorithms in an effort to quantify minimum sequence complexity. But compression algorithms tell us nothing about biofunction algorithms.

Compression algorithms are something we humans do to the sequence itself. Such tasks are done on and internal to the bit string. Biofunction algorithms, on the other hand, are accomplished external to the sequence. They are performed by the sequence on its environment. The function such sequences produce predates humans with their conscious computational pursuits altogether. So we must be careful not to confuse the two kinds of algorithms. Kolmogorov-Chaitin-Yockey algorithmic complexity deals with compression algorithms, not with the functional algorithms performed by messenger molecules. Exon equivalents are instructional sequences with undeniable meaning. They are not just stochastic ensembles all having an equal number of bits. Some sequences engineer cellular metabolism. Others do not. What's the difference?

Sometimes compression formulae permit greatly shortened sequences to retain high functionality. A very short algorithm for 1r can replace an endless string of seemingly random integers. Yet no one would doubt the high functionality of TC. While complexity is normally a component of any sophisticated information-retaining linear matrix, it is not synonymous with bio-information.

6. Principles of BioFunction theory

6.1. Bio-information is fundamentally ontological, not epistemological

Information objectively exists independent of knowledge, though it often becomes the object and pursuit of human knowledge. Systems of knowing information are epistemological. These systems must never be confused with information itself.

6.2. Shannon theory is a somewhat crippled knowledge system

Set theory is a knowledge system. Combinatorics, probabilism, and set theory are all functions of cognition. Human notions of "before-and-after uncertainty differences" and "aboutness in certain environments" are altogether sentient and epistemological.

The knowledge of information by sentient beings allows them to undertake sophisticated tasks 'artificially' through their 'agency.' But molecular biological information produces its biofunction independent of knowers or their artificial agency. Life predates knowers, their ideas of aboutness, and their combinatorial and probabilistic computations. Any attempt to cram bio-information into Shannon's knowledge-system box will invariably yield an inadequate sense of what functional bio-information objectively is.

6.3. The meaning of bio-information is the biofunction instructed by that information

Information with no meaning was never functional information in the first place. At best, it was a non-instructive enumeration of specific sequence. But so what? Any sequence is specific. Life does not depend upon humans' specific enumerations of sequences or their measurements. life exists on its own. The biopolymers of life behave as though every nucleotide selection were a meaningful decision node choice. Even introns once thought to be junk will likely be found to serve crucial function in proper spacing and architecture. Each monomer is not just specific. It is specified to function. This is why bioengineering is the only appropriate term for what we observe in molecular biology. Genetic instructions cannot adequately be measured using probabilistic and combinatorial Shannon theory.

6.4. Objective bio-information is positive, nonrandom, specifically instructive, and functional

Shannon theory treats sequences as though they were random. Genetic instructions are anything but random sequences. Before-and-after differences in uncertainty serve only to shrink the sample space of known possibilities in our minds. Such differences tell us nothing positive and specific within the remaining sample space about what works. Nonmathematical knowledge of functional information (such as that contributing to the genome and proteome projects) allows precise and sophisticated predictions of relative certainty of biofunction. Standalone Shannon equations provide nothing but relative uncertainty.

Only when Shannon math is informed by human observation that a certain sequence works (e.g., a known exon) can we get true genetic information. It is specific enumeration of an instructional sequence that matters. The latter must be weighted against the Shannon bits for that sequence. There must be a marriage of specification with 'total possibilities'. It is fits that we want and need, not just bits. Fits are a measure of functional bits. Almost all of the bits of so-called information are useless in biology. Only one or a very few algorithmic strings of nucleotide switch settings successfully produce the needed catalyst. 100 % of fits work. Fits give us a measurement of specified complexity with proven biofunction.

6.5. Quantification of bio-information cannot be achieved with mathematics alone. Empirical knowledge of instructed function must educate our mathematical axioms

The situation is similar to Kurt Goedel's Incompleteness theorem in mathematics,[23] and to the pragmatic halting problem in computer proprammming.[2] The meaning of information is undecidable in Shannon theory.[2,,5] Function is not addressed mathematically. Observation alone tells us what sequence works and whether a computational program will halt (complete its task).

6.6. Order and complexity are inversely related

High compressibility means a sequence is more ordered and therefore redundant. High compressibility means a sequence is less complex. Random sequences are the most complex of all. Complexity alone has nothing to do with meaning or function.

6.7. Bio-information and the biofunction it instructs are as 'conceptual' as the general mathematical workings of nature

All three predate human discovery and knowledge. None was generated by human minds. Mathematical concept in physics is objectively existent in nature independent of human mentation.

We would have made no progress in physics had we disallowed 'concept' in nature.

Mathematics is purely conceptual. Physics presupposes a cosmic mathematical rationality. Disallowing concept in nature would have precluded approaching physics mathematically. Mathematics affords the best description and predictability we have in physics. We must revisit "the unreasonable effectiveness of mathematics"[24] in describing a nature that we claim is chaotic and arational. Mathematics is the ultimate expression of rationality. Why, then, does mathematics work so well to describe reality? Could it be that reality is not as chaotic and a-rational as we metaphysically presupposed?

We will never begin to understand bio-information, genetic instructions, biofunction, or metabolism so long as the science of biology lives in denial of observable concept (emergence) within innumerable natural life processes. Continuing to evaluate biopolymer sequence theory and metabolic pathways using nothing but uncertainty equations will only preclude progress.

The following questions are impossible to answer using current information theory, physics, chemistry, and mathematics:

- what exactly is the difference between an intron and an exon from the standpoint of mathematical sequence theory?
- what exactly is the difference between an intron and an exon from the standpoint of physicochemical determinism (necessity)?
- how do nucleotide sequences acquire functional significance when protein tertiary structures are coded in a completely different language from nucleotide language?

Questions relating to the origin of FSC are among the most difficult in biology, if not all of science. But perhaps we can answer the question asked by this chapter: "Is life reducible to complexity?" In the case of OSC and RSC, the best answer would probably be the slang expression, "No way!"

FSC does indeed specify function. It specifically enumerates through transcription/translation the sequences of amino acids that will work. FSC allows precise predictions. The tertiary structures of both catalytic and structural proteins are primarily determined by their one-dimensional amino acid sequence. Both nucleic acid and proteins example FSC rather than OSC or RSC. If life were reducible to any physical form of complexity, it would be FSC.

The problem is that 100 % of a cell's FSC is still intact nanoseconds after cell death. If life is reducible to FSC, why then is the cell now dead?

The answer perhaps lies once again in the analogy of computer algorithms. Software programs and their.switches are intact on a turned-off hard drive. But it is only the dynamic algorithmic process that provides computer function. With computers, algorithmic process requires an energy flow into the system from its environment along with the system's instruction set. Selected (specified) decision node sequences along with hardware engineering transduce that energy and utilize it for its function and computational output. Computers depend on FSC, but their operation is more than just the recorded instructions in their FSC. The same is true in biology.

Life, then, is not only not reducible to complexity; it is not even reducible to FSC! Life is a symphony of dynamic, highly integrated, algorithmic processes yielding homeostatic metabolism, development, growth, and reproduction (ignoring the misgivings of those few life-origin theorists with mule fixations!).

But as Yockey argues, it remains to be seen whether such highly sophisticated algorithmic processes can exist apart from the linear, segregatable, digital, FSC instructions observed at the helm of all known empirical life.

Acknowledgements

The author greatly appreciates the invitation and encouragement of Gyula Palyi to contribute this chapter, and the time-consuming critiques and edits of Robert Bass, J. M. Kaper, Richard Clark, Mike O'Neill, Hubert Yockey, and Lowell Owens.

References

1) Dyson F., Life in the Universe: Is life Digital or Analog?, NASA Goddard Space Flight Center CoIloquiem, Greenbelt, MD, 1999.

2) Yockey H.P., Origin of life on Earth ond Shannon's theory of communication, Camp. Chem. 24/1 (2000) 105-123.

3) (a) Orgel L.E., The Origins of life: Molecules and Natural Selection, John Wiley, New York, 1973, p 189. (b) Ibid 1973, p 190.

4) Ronneberg TA., Landweber L.F., Freeland S.J., Testing a biosynthetic theory of the genetic code: Fact or artifact?, Proc. Natl. Acad. Sci. USA 21 November (2000).

5) Yockey H.P., Information Theory and Molecular Biology, Cambridge University Press, Cambridge, 1992, 408 p ..

6) Shannon C, Parts I and II: A mathematical theory of communication, The Bell System Techn. J. XXVII (3 July) (1948) 379-423

7) (a) Mojzsis S.J, et al., Evidence for life on Earth before 3,800 million years ago. Nature, 384 (1996) 55-59. (b) Mojzsis S., Krishnamurthy R., Arrhenius G., Before RNA and After: Geophysical and Geochemical Constraints on Molecular Evolution, in: Mojzsis S., Krishnamurthy R., Arrhenius G. (Eds.), The RNA World, Harbor laboratory Press, Cold Spring, 1999.

8) (a) Wilde S.A., et al., Evidence from detrital zircons for the existence of continental crust and oceans on the Earth 4.4 Gyr ago, Nature 409 (II January) (2001) 175-178. (b) Mojzsis S.J., Harrison M., Pidgeon R.I, Oxygen-isotope evidence from ancient zircons for liquid water at the Earth's surface 4 300 Myr ago, Nature 409 (II January) (2001) 178-181.

9) (a) Jukes TH., Quakes hit California and, perhaps, the central dogma, J. Mol. Evol. 30/1 (1990) 1-2. (b) Stuart K., RNA editing: trypanosomes rewrite the genetic code, Verh. K. Acad. Geneeskd Belg. 60/1 (1998) 163-74.

10) (a) Yockey H.P., An application of information theory to the Central Dogma and the Sequence Hypothesis, J. Theor. Bioi. 46/2 (1974) 369-406. (b) Yockey H.P, Can the central dogma be derived from information theory?, Theor. Bioi. 74/1 (1978) 149-152.

11) (a) Schneider TD., et al., Information content of binding sites on nucleotide sequences, j. Mol. BioI. 188/3 (1986) 415-431. (b) Schneider TD., in: Erickson G.., Smith CR. Eds, Maximum-Entropy and Bayesian Methods in Science and Engineering, Kluwer Academic, Dordrecht, The Netherlands, 1988, p. 147-154. (c) Schneider TD., Stormo G.D., Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique, Nucleic Acids Res. 17/2 (1989) 659-674. (d) Schneider T.D., Stephens R.M., Sequence logos: a new way to display consensus sequences, Nucleic Acids Res. 18/20 (1990) 6097-6100. (e) Schneider TD., Theory of molecular mochines. I. Channel capacity of molecular machines, J Theor. Biol. 148/1 (1991) 83-123. (f) Schneider TD., Theory of molecular machines. II. Energy dissipation from molecular machines, J Theor. Bioi. 148/1 (1991) 125-137. (g) Schneider TD., Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers; a review of the theory of molecular machines, Nanotechnology 5 (1994)1-18. (h) Schneider TD., Reading of DNA sequence logos: prediction of mjior groove binding by information theory, Methods Enzymol. 274 (1996) 445-455. (i) Schneider TD., Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences [published erratum appears in Nucleic Acids Res. 26/4 (1998) following 1134], Nucleic Acids Res. 25/21 (1997) 4408-4415. (i) Schneider TD., Information content of individual genetic sequences, J Theor. BioI. 189/4 (19971 427-441. (k) Schneider TD., Measuring molecular information, J Theor. BioI. 201/1 (1999) 87-92. (I) Schneider TD., Evolution of biological information, Nucleic Acids Res. 28/14 (2000) 2794-2799. (m) Schneider TD., The bottle, Nature, 406/6794 (2000) 351.

12) (a) Adami C lEd.), Introduction to Artificial Life. Springer, New York, 1998. (b) Adami C, Cerf Nj., Physicol complexity of symbolic sequences, Physica D 137 (2000) 62-69. (cl Adami C, Ofria C, Collier TC, Evolution of biological complexity,. Proc. Natl. Acad. Sci. USA, 97/9 (2000) 4463-4468.

13) (a) Yockey H.P., A prescription which predicts functionally equivalent residues at given sites in protein sequences, J Theor. BioI. 67/3 (1971) 337-343. (b) Yockey H.P., On the information content of cytochrome c, J Theor. Bioi. 67/3 (1977) 345-376. (c) Yockey H.P., A calculation of the probability of spontaneous biogenesis by information theory, J Theor Bioi. 67/3 (1977) 377-398. (d) Yockey H.P., Do overlapping genes violate molecular biology and the theory of evolution?, J Theor. BioI. 80/1 (1979) 21-26. (e) Yockey H.P., Self organization origin of life scenarios and information theory, J Theor. BioI. 91/1 (19811 13-31. (f) Yockey H.P., Rebuttal of "overlapping genes and information theory" [letter], J Theor. BioI. 91/2 (1981) 381-382. (g) Yockey H.P., Children of choice [letter; comment], Nature 364/6432 (1993) 10. (h) Yockey H.P, Comments on "Let there be life; thermodynamic reflections on biogenesis and evolution" by Elitzur AC [comment], J Theor. BioI. 176/3 (1995) 349-355. (i) Yockey H.P., Color blindness results from an alteration in the linear, segregatable, digital sequence of DNA monomers, 2000.

14) (0) Segre D., et aI., Graded autocatalysis replication domain (GARD): kinetic analysis of self-replication in mutually catalytic sets, Origin Life Evol. Biosph. 28/4-6 (1998) 501-514. (b) Segre D., Ben-Eli D., Lancet D., Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies, Proc. Natl. Acad. Sci. USA 97/8 (2000) 4112-4117.

15) (a) Cairns-Smith AG., Walker G.L., Primitive metabolism, Curr. Mod. Biol. 5/4 (1974) 173-186. (b) Cairns-Smith AG., Takeover mechanisms and early biochemical evolution, Biosystems 9/2-3 (1977) 105-109. (c) Cairns-Smith AG., Seven Clues to the Origin of life, Cambridge University Press, Cambridge, Canto Ed, 1990, 130 P

16) Wyner AD., Typical sequences and all that: entropy, pattern matching, and data compression, in: 1994 Shannon Lecture, Murray Hill, New Jersey 07974 USA, AT&T Bell Labs,1994

17) Wong J-T, A co-evolution theory of the genetic code, Proc. Natl. Acad. Sci. USA. 72/5 (1975) 1909-1912.

18) Shapiro R., A replicator was not involved in the origin of life, IUBMB Life 49/3 (2000) 173-176.

19) Weiss 0., Jimenez-Montano M.A, Herzel H., Information content of protein sequences, J Theor. Bioi. 206/3 (2000) 379-386.

20) Prigogine I., From Being to Becoming, W. H. Freeman and Co., San Francisco, (1980).

21) Turing AM., On computable numbers, with an application to the Entscheidungs problem. Proc. Roy. Soc. London, 42 (1936) 230-265 [correction in 243, 544-546].

22) Tribus M., Thermostatics and Thermodynamics. D. van Nostrand Company, Inc. Princeton, NJ, 1961.

23) (a) Godel K., Ober formal unentscheidbare Satze der Principia Mathematica und verwandte Systeme I (On formal undecidable propositions of Principia Mathematica and related systems I), Monat: Math. Phys. 38, (1931) 173- 198. (b) Chaitin G., Goedel's theorem and information, Internat. J Theor. Phys., 22 (1982) 941-945.

24) (a) Wigner E.P., The Unreasonable Effectiveness of Mathematics in the Natural Sciences, Comun. Pure Applied Math. XIII (1960) 1-14. (b) Hamming R.W., The unreasonable effectiveness of mathematics, Am. Mathemat: Month. 87 (2 February). 1980.