DNA and the Origin of Life: Information, Specification, and Explanation

By Stephen C. Meyer

ABSTRACT—Many origin-of-life researchers now regard the origin of biological information as the central problem facing origin-of-life research. Yet, the term 'information' can designate several theoretically distinct concepts. By distinguishing between *specified* and unspecified information, this essay seeks to eliminate definitional ambiguity associated with the term 'information' as used in biology. It does this in order to evaluate competing explanations for the origin of biological information. In particular, this essay challenges the causal adequacy of naturalistic chemical evolutionary explanations for the origin of *specified* biological information, whether based upon "chance," "necessity," or the combination. Instead, it argues that our present knowledge of causal powers suggests intelligent design or agent causation as a better, more causally adequate, explanation for the origin of specified biological information.

1. INTRODUCTION

Discussions of the origin of life necessarily presuppose knowledge of the attributes of living cells. As historian of biology Harmke Kamminga has observed, "At the heart of the problem of the origin of life lies a fundamental question: What is it exactly that we are trying to explain the origin of?" [1, p. 1]. Or as the pioneering chemical evolutionary theorist Alexander Oparin put it, 'the problem of the nature of life and the problem of its origin have become inseparable' [2, p. 7]. Origin-of-life researchers want to explain the origin of the first and presumably simplest—or, at least, minimally complex—living cell. As a result, developments in fields that explicate the nature of unicellular life have historically defined the questions that origin-of-life scenarios must answer.

Since the late 1950s and 1960s origin-of-life researchers have increasingly recognized the complex and specific nature of unicellular life and the biomacromolecules upon which such systems depend. Furthermore, molecular biologists and origin-of-life researchers have characterized this complexity and specificity in informational terms. Molecular biologists routinely refer to DNA, RNA and proteins as carriers or repositories of 'information' [3-6]. Further, many origin-of-life researchers now regard the origin of the information in these biomacromolecules as the central question facing origin-of-life research. As Bernd-Olaf Kuppers has stated, "the problem of the origin of life is clearly basically equivalent to the problem of the origin of biological information" [7, pp. 170-72].

This essay will evaluate competing explanations for the origin of the biological information necessary to build the first living cell. Yet, to do so will require determining what biologists have meant by the term 'information' as it has been applied to bio-

macromolecules. As many have noted, 'information' can denote several theoretically distinct concepts. Thus, this essay will attempt to eliminate this ambiguity and to determine precisely what type of information origin-of-life researchers must explain 'the origin of.' Thus, what follows comprises two clear divisions. The first will seek to *characterize* the information in DNA, RNA and proteins as an *explanadum*—a fact in need of explanation; the second will *evaluate* the efficacy of competing classes of explanation for the origin of biological information— that is, the competing *explanans*.

Part One will seek to show that molecular biologists have used the term 'information' consistently to refer to the joint properties of 'complexity' and functional 'specificity' or 'specification.' This part will contrast the biological usage of the term with its classical information-theoretic usage and show that 'biological information' entails a richer sense of information than the classical mathematical theory of Shannon and Wiener. It will also argue *against* attempts to treat biological 'information' as a metaphor that lacks empirical content and/or ontological status [8-10]. Instead, it will show that the term biological information refers to two real features of living systems—indeed, ones that jointly do require explanation.

Part Two will evaluate competing types of explanation for the origin of specified biological information. In so doing, it will employ the categories of 'chance' and 'necessity.' These categories provide a helpful heuristic for understanding the recent history of origin-of-life research. From the 1920s to the mid-1960s origin of life researchers relied heavily on theories that emphasized the creative role of random events—'chance'—often in tandem with some form of pre-biotic natural selection. Since the late 1960s, theorists have instead emphasized deterministic self-organizational laws or properties, i.e., 'necessity.' Part Two will critique the causal adequacy of chemical evolutionary theories based upon 'chance,' 'necessity,' and their combination. Instead, a concluding third part will suggest that the phenomenon of *specified* complexity or *specified* information requires a radically different explanatory approach. In particular, I will argue that our present knowledge of causal powers suggests intelligent design or agency as a better, more causally adequate, explanation for the origin of specified information, including that present in large biomolecules such as DNA, RNA and proteins.

2.1 Simple to Complex: Defining the Biological *Explanandum*

After Darwin published the *Origin of Species* in 1859, many scientists began to think about a problem that Darwin had not addressed,[i] namely, how life had arisen in the first place. While Darwin's theory purported to explain how life could have grown gradually more complex starting from "one or a few simple forms," it did not explain, nor did it attempt to explain, how life had first originated. Yet evolutionary biologists in the 1870s and 1880s such as Ernst Haeckel and Thomas Huxley assumed that devising an explanation for the origin of life would be fairly easy in large part because Haeckel and Huxley assumed life was, in its essence, a chemically simple substance called

'protoplasm.' Both thought protoplasm could be easily constructed by combining and recombining simple chemicals such as carbon dioxide, oxygen and nitrogen.

Over the next sixty years biologist and biochemists gradually revised their view of the nature of life. Whereas many biologists during the 1860s and 70s saw the cell, in Ernst Haeckel's words, as an undifferentiated and "homogeneous globule of plasm", by the 1930s most biologist had come to see the cell as a complex metabolic system [11, p. 111; 12]. Origin of life theories reflected this increasing appreciation of cellular complexity. Whereas 19th century theories of abiogenesis envisioned life arising almost instantaneously via a one or two-step processes of chemical 'autogeny,' Alexander Oparin's theory of *evolutionary* abiogenesis envisioned a multi-billion year process of transformation from simple chemicals to a complex metabolic system [13, pp. 64-103; 14, pp. 174-212]. Even so, most scientists during the late 1930s (whether those studying the nature of life or its origin) still vastly underestimated the complexity and specificity of the cell and its key functional components—as developments in molecular biology would soon make clear.

2.2. The Complexity and Specificity of Proteins

During the first half of the twentieth century biochemists had come to recognize the centrality of proteins to the maintenance of life. Many mistakenly believed that proteins also contained the source of heredity information. Nevertheless, throughout the first half of twentieth century biologists repeatedly underestimated the complexity of proteins. For example, during the 1930s the English X-ray crystallographer William Astbury elucidated the molecular structure of certain fibrous proteins, such as keratin, the key structural protein in hair and skin [15; 16, p. 80; 17, p. 63]. Keratin, exhibits a relatively simple, repetitive structure, and Astbury was convinced that all proteins, including the mysterious globular proteins so important to life, represented variations on the same primal and regular pattern. Similarly, the biochemists Max Bergmann and Carl Niemann of the Rockefeller Institute argued in 1937 that the amino acids in proteins occurred in regular, mathematically expressible proportions [17, p. 7]. Other biologists imagined that insulin and hemoglobin proteins, for example, "consisted of bundles of parallel rods" [17, p. 265].

Beginning in the 1950s, however, biologists made a series of discoveries that caused this simplistic view of proteins to change. From 1949-1955 the molecular biologist Fred Sanger determined the structure of the protein molecule insulin. Sanger showed that insulin comprised a long and irregular sequence of the various proteineous amino acids, rather like a string of differently colored beads arranged without any discernible pattern [16, pp. 213, 229-35, 255-61, 304, 334-35; 18]. His work showed for a single case what subsequent work in molecular biology would establish as a norm: amino acid sequencing in functional proteins generally defies expression by any simple rule and is characterized, instead, by aperiodicity or complexity [16, pp. 213, 229-35, 255-61, 304, 334-35]. Later in the 1950s, work by Andrew Kendrew on the structure of the protein myoglobin showed that proteins also exhibit a surprising three-dimensional complexity. Far from

the simple structures that biologists had imagined earlier, Kendrew's work revealed an extraordinarily complex and irregular three-dimensional shape—a twisting, turning, tangle of amino acids. As Kendrew explained in 1958, "the big surprise was that it was so irregular....the arrangement seems to be almost totally lacking in the kind of regularity one instinctively anticipates, and it is more complicated than has been predicted by any theory of protein structure" [19; 16, pp. 562-63].

By the mid-1950s, biochemists recognized that proteins possess another remarkable property. In addition to their complexity, proteins also exhibit specificity, both as one-dimensional arrays and three-dimensional structures. Whereas proteins are built from chemically rather simple amino acid 'building blocks,' their function (whether as enzymes, signal transducers or structural components in the cell) depends crucially upon the complex but specific arrangement of these building blocks [20, pp. 111-12, 127-31]. In particular, the specific sequencing of amino acids in a chain, and the resultant chemical interactions between amino acids, (largely) determine the specific three-dimensional structure that the chain as a whole will adopt. These structures or shapes in turn determine what function, if any, the amino acid chain can perform in the cell.

For a functioning protein, its three-dimensional shape gives it a 'hand-in-glove' fit with other molecules in the cell, enabling it to catalyze specific chemical reactions or to build specific structures within the cell. Because of this three-dimensional specificity, one protein can usually no more substitute for another, than one tool can substitute for another. A topoisomerase can no more perform the job of a polymerase, than a hatchet can perform the function of soldering iron. Instead, proteins perform functions only by virtue of their three-dimensional specificity of fit either with other equally specified and complex molecules or with more simple substrates within the cell. Moreover, this three dimensional specificity derives in large part from the one-dimensional specificity of sequencing in the arrangement of the amino acids that form proteins. Indeed, even slight alterations in sequencing often result in the loss of protein function.

2.3 The Complexity and Sequence Specificity of DNA

During the early part on the twentieth century, researchers also vastly underestimated the complexity (and significance) of nucleic acids such as DNA and RNA. By the early part of the twentieth century, biologists knew the chemical composition of DNA. Chemists knew that in addition to sugars (and later phosphates), DNA was composed of four different nucleotide bases, called adenine, thyamine, cytosine and guanine. In 1909, the chemist P.A. Levene had shown (incorrectly as it later turned out) that these four different nucleotide bases always occurred in equal quantities within the DNA molecule [16, p. 30]. He formulated what he called the "tetranucleotide hypothesis" to account for this putative fact. According to the tetranucleotide hypothesis, the four nucleotide bases in DNA link together in repeating sequences of the same four chemicals in the same sequential order. Since Levene envisioned these sequential arrangements of nucleotides as repetitive and invariant, their potential for expressing any genetic diversity seemed inherently limited. To account for the heritable differences

between species, biologists needed to discover some source of variable or irregular specificity—some source of information—within the germ lines of different organisms. Yet in so far as DNA was seen as an uninterestingly repetitive molecule most biologists assumed that DNA could play little if any role in the transmission of heredity.

This view began to change in the mid-1940s for several reasons. First, Avery Oswald's famous experiments on virulent and non-virulent strains of pneumococcus identified DNA as the key factor in accounting for heritable differences between these different bacterial strains [16, pp. 30-31, 33-41, 609-10; 21]. Second, work by Erwin Chargaff of Columbia University in the late 1940s undermined the "tetranucleotide hypothesis." Chargaff showed, contradicting Levene's earlier work, that nucleotide frequencies actually do differ between species, even if they often hold constant within the same species or within the same organs or tissues of a single organism [22, p. 21; 16, pp. 95-96]. More importantly, Chargaff recognized that even for nucleic acids of exactly "the same analytical composition"—meaning those with precisely the same relative proportions of A, T, C, and G—"enormous" numbers of variations in sequencing were possible. Indeed, as he put it, different DNA molecules or parts of DNA molecules might "differ from each other. . .in the sequence, [though] not the proportion, of their constituents" [22, p. 21]. As he realized, for a nucleic acid consisting of 2500 nucleotides (roughly the length of a long gene) the number of sequences "exhibiting the same molar proportions of individual purines [A,G] and pyrimidines [T,C] . . . is not from from 10^{1500}" [22, p. 21]. Thus, Chargaff showed that, contrary to the tetranucleotide hypothesis, base sequencing in DNA might well display a high degree of improbability, complexity and aperiodicity—as required by any potential carrier of heredity.

Thirdly, the elucidation of the structure of DNA by Watson and Crick in 1953 made clear that DNA could function as a carrier of hereditary information [3]. The model that Watson and Crick proposed envisioned a double-helix structure to explain the maltese cross pattern derived from X-Ray crystallographic studies of DNA by Franklin, Wilkins and Bragg in the early 1950s. According to the now well-known Watson and Crick model, the two strands of the helix were made of sugar and phosphate molecules linked by phosphodiester bonds. Nucleotide bases were linked horizontally to the sugars on each strand of the helix and to a complementary base on the other strand to form an internal 'rung' on the twisting 'ladder.' For geometric reasons, their model required the pairing (across the helix) of adenine with thyamine and cytosine with guanine, respectively. This complementary pairing helped to explain a significant regularity in composition ratios that Chargaff had discovered. Though Chargaff had shown that none of the four nucleotide bases appear with the same frequency as all the other three, he did discover that the molar proportions of adenine and thyamine, on the one hand, and cytosine and guanine, on the other, do consistently equal each other [16, p. 96]. Watson and Crick's model explained this regularity as Chargaff had expressed it in his famous "ratios."

Yet the Watson-Crick model also made clear that DNA might possess an impressive chemical and structural complexity. Not only did the double helix structure presuppose

(as was then widely known) that DNA constituted an extremely long and high molecular weight structure, but the Watson and Crick model also implied that the sugar molecules in the sugar-phosphate backbone would allow (from a chemical point of view) any of the four nucleotide bases to attach to them. This chemical freedom suggested that the sequencing of bases would (in all probability) defy reduction to any rigidly repeating pattern, thus allowing DNA to possess an impressive potential for variability and complexity in sequencing. As Watson and Crick explained, "The sugar-phosphate backbone in our model is completely regular but any sequence of base pairs can fit into the structure. It follows that in a long molecule many different permutations are possible, and it, therefore, seems likely that the precise sequence of bases is the code which carries genetic information" [4].

As with proteins, subsequent discoveries soon showed that DNA sequencing was not only complex, but also highly specific relative to the requirements of biological function. Indeed, the discovery of the complexity and specificity of proteins led researchers to suspect a functionally specific role for DNA. Molecular biologists, working in the wake of Sanger's results, assumed that proteins were much too complex (and yet also functionally specific) to arise by chance *in vivo*. Moreover, given their irregularity, it seemed unlikely that a general chemical law or regularity could explain their assembly. Instead, as Jacques Monod has recalled, molecular biologists began to look for some source of information or 'specificity' within the cell that could direct the construction of these highly specific and complex structures. To explain the presence of the specificity and complexity in the protein, as Monod would later explain, "you absolutely needed a code" [16, p. 611].

The structure of the DNA as elucidated by Watson and Crick suggested a means by which information or 'specificity' might be encoded along the spine of DNA's sugar-phosphate backbone [3,4]. Their model suggested that variations in sequencing of the nucleotide bases might find expression in the sequencing of the amino acids that form proteins. In 1955 Francis Crick proposed this idea as the so-called "sequence hypothesis" [16, pp. 245-46]. According to Crick's hypothesis, the specificity of arrangement of amino acids in proteins derives from the specific arrangement of the nucleotide bases on the DNA molecule [16, pp. 335-36]. The sequence hypothesis suggested that the nucleotide bases in DNA functioned like letters in an alphabet or characters in a machine code. Just as alphabetic letters in a written language may perform a communication function depending upon their sequencing, so too might the nucleotide bases in DNA result in the production of a functional protein molecule depending upon their precise sequential arrangement. In both cases, function depends crucially upon sequencing. Thus, the sequence hypothesis implied not only the complexity, but also the functional specificity of DNA base sequencing.

By the early 1960s, a series of experiments had confirmed that DNA base sequencing plays a critical role in determining amino acid sequencing during protein synthesis [16, pp. 470-89; 23; 24]. Further, by this time, molecular biologists had determined (at least in outline) the processes and mechanisms by which DNA sequences determine key stages

of this process. Protein synthesis or 'gene expression' proceeds as long chains of nucleotide bases are first copied during a process known as 'transcription.' The resulting copy, a 'transcript' made of single-stranded 'messenger RNA,' comprises a sequence of RNA bases that precisely reflects the sequence of bases on the original DNA strand [20, pp. 106-08; 25, pp. 574-82, 639-48]. This transcript is then transported to a complex organelle called a ribosome. At the ribosome, the transcript is 'translated' (with the aid of highly specific adaptor molecules called transfer-RNAs) and specific enzymes (called amino-acyl t-RNA synthetases) to produce a growing amino acid chain [20, pp. 108-10; 25, pp. 650-84]. (See Figure 1). Whereas the function of the protein molecule derives from the specific arrangement of twenty different types amino acids, the function of DNA depends upon the arrangement of just four kinds of bases. This lack of one-to-one correspondence means that a group of three DNA nucleotides (a triplet) are needed to specify a single amino acid. In any case, the sequential arrangement of the nucleotide bases in DNA does determine (in large part)[ii] the one-dimensional sequential arrangement of amino acids during protein synthesis. Moreover, since protein function depends critically upon amino acid sequencing, and amino acid sequencing depends critically upon DNA base sequencing, DNA base sequences (in the coding regions of DNA) themselves possess a high degree of specificity relative to the requirements of protein (and cellular) function.

2.4. Information Theory and Molecular Biology

From the beginning of the molecular biological revolution, biologists have ascribed information-bearing properties to DNA, RNA and proteins. In the parlance of molecular biology, DNA base sequences contain the 'genetic information' or the 'assembly instructions' necessary to direct protein synthesis. Yet the term 'information' can denote several theoretically distinct concepts. It will, therefore, be necessary to clarify which sense of 'information' applies to large biomacromolecules such as DNA and protein in order to clarify what kind of information origin-of-life researchers must explain 'the origin of.' This will prove particularly important because, as we shall see, molecular biologists employ both a stronger conception of information than mathematicians and information-theorists, and a (slightly) weaker conception of the term than linguists and ordinary users.

During the 1940s, Claude Shannon at Bell Laboratories developed a mathematical theory of information [26]. His theory equated the amount of information transmitted with the amount of uncertainty reduced or eliminated by a series of symbols or characters [27, pp. 6-10]. For example, before one rolls a six-sided die, there are six possible outcomes. Before one flips a coin there are two. Rolling a die will thus eliminate more uncertainty and, on Shannon's theory, convey more information, than flipping a coin. Equating information with the reduction of uncertainty implied a mathematical relationship between information and probability (or its inverse, complexity). Note that for a die each possible outcome has only a 1 in 6 chance of occurring, compared to a 1 in 2 chance for each side of the coin. Thus, in Shannon's theory the occurrence of the more

improbable event conveys more information. Shannon generalized this relationship by stating that the amount of information conveyed by an event is inversely proportional to the prior probability of its occurrence. The greater the number of possibilities, the greater the improbability of any one being actualized, and thus, the more information is transmitted when a particular possibility occurs.

Moreover, information increases as improbabilities multiply. The probability of getting four heads in a row when flipping a fair coin is 1/2 X 1/2 X 1/2 X 1/2 or $(1/2)^4$. Thus, the probability of attaining a specific sequence of heads and/or tails decreases exponentially as the number of trials increases. The quantity information increases correspondingly. Even so, information theorists found it convenient to measure information additively rather than multiplicatively. Thus, the common mathematical expression ($I = -\log_2 p$) for calculating information converts probability values into informational measures through a negative logarithmic function (where the negative sign expresses an inverse relationship between information and probability) [26, 27, pp. 6-10].

Shannon's theory applies most easily to sequences of alphabetic symbols or characters that function as such. Within any given alphabet of **x** possible characters, the placement of a specific character eliminates **x-1** other possibilities and thus a corresponding amount of uncertainty. Or put differently, within any given alphabet or ensemble of **x** possible characters, (where each character has an equi-probable chance of occurring), the probability of any one character occurring is 1/**x**. The larger the value of **x**, the greater the amount of information that is conveyed by the occurrence of a specific character in a sequence. In systems where the value of **x** can be known (or estimated), as in a code or language, mathematicians can easily generate quantitative estimates of information carrying capacity. The greater the number of possible characters at each site, and the longer the sequence of characters, the greater is the information carrying capacity (or Shannon information) associated with the sequence.

The functionally alphabetic character of the nucleotide bases in DNA and the amino acid residues in proteins enabled molecular biologists to calculate the information carrying capacity (or syntactic information) of these molecules using the new formalism of Shannon's theory. Because at every site in a growing amino acid chain, for example, the chain may receive any one of twenty proteineous amino acids, the placement of a single amino acid in the chain eliminates a quantifiable amount of uncertainty and increases the (Shannon or syntactic) information of a polypeptide by a corresponding amount. Similarly, since at any given site along the DNA backbone any one of four nucleotide bases may occur (with equal probability [28], the **p** value for the occurrence of a specific nucleotide at that site equals 1/4 or .25 [28, p. 364]. The information carrying capacity of a sequence of a specific length **n** can then be calculated using Shannon's familiar expression ($I = -\log_2 p$) once one computes a **p** value for the occurrence of a particular sequence **n** nucleotides long where $\mathbf{p}=(1/4)^{\mathbf{n}}$. This **p** value yields a corresponding measure of information carrying capacity or syntactic information for a sequence of **n** nucleotide bases [5].[iii]

2.5 COMPLEXITY, SPECIFICITY AND *BIOLOGICAL* INFORMATION

Though Shannon's theory and equations provided a powerful way to measure the amount of information that could be transmitted across a communication channel, it had important limits. In particular, it did not, and could not distinguish merely improbable sequences of symbols from those that conveyed a message. As Warren Weaver made clear in 1949, "the word information in this theory is used in a special mathematical sense that must not be confused with its ordinary usage. In particular, information must not be confused with meaning" [29, p. 8]. Information theory could measure the "information carrying capacity" or the "syntactic information" of a given sequence of symbols, but could not distinguish the presence of a meaningful or functional arrangement of symbols from a random sequence (e.g. "we hold these truths to be self-evident. . ." v. "ntnyhiznlhteqkhgdsjh"). Thus, Shannon information theory could quantify the amount of functional or meaningful information that *might be present* in a given sequence of symbols or characters, but it could not distinguish the status of a functional or message-bearing text from random gibberish. Thus, paradoxically, random sequences of letters often have more syntactic information (or information carrying capacity) as measured by classical information theory, than do meaningful or functional sequences that happen to contain a certain amount of intentional redundancy or repetition.

In essence, therefore, Shannon's theory provides a measure of complexity or improbability, but remains silent upon the important question of whether a sequence of symbols is functionally specific or meaningful. Nevertheless, in its application to molecular biology, Shannon information theory did succeed in rendering rough quantitative measures of the "information carrying capacity" or "syntactic information" (where these terms correspond to measures of brute complexity) [5; 30, pp. 58-177]. As such, information theory did help to refine biologists' understanding of one important feature of the crucial biomolecular components upon which life depends: DNA and proteins are highly complex, and quantifiably so. Nevertheless, information theory by itself did not, and could not, establish whether base sequences (in DNA) or amino acid sequences (in proteins) possessed the property of functional specificity. Information theory could measure the amount of "syntactic information" that DNA and proteins possess, it could not determine whether these molecules possessed "functional" or "semantic" information. Information theory could help to establish that DNA and proteins *could* carry large amounts of functional information, it could not establish whether or not they did.

The ease with which information theory applied to molecular biology (to measure information carrying capacity), has created considerable confusion about the sense in which DNA and proteins contain "information." Information theoretic analyses of DNA and proteins strongly suggested that these molecules possess vast information carrying capacities or large amounts or "syntactic information," as defined technically by Shannon's theory. Nevertheless, in their descriptions of DNA as the carrier of hereditary information, for example, molecular biologists have meant much more by the term "information" than these technically limited terms. Instead, as Sarkar points out, leading

molecular biologists defined biological information so as to incorporate the notion of specificity of function (as well as complexity) as early 1958 [31, p. 196; 32]. Molecular biologists such as Monod and Crick understood biological information—indeed, the information stored in DNA and proteins—as something more than mere complexity (or improbability). While their notion of information did associate both biochemical contingency and combinatorial complexity with DNA sequences (thus, allowing its carrying capacity to be calculated), they also recognized that sequences of nucleotides and amino acids in functioning biomacromolecules possessed a high degree of *specificity* relative to the maintenance of cellular function. As Crick would explain in 1958, "By information I mean the specification of the amino acid sequence in protein. . . Information means here the *precise* determination of sequence, either of bases in the nucleic acid or on amino acid residues in the protein" [32, pp. 144, 153].

Since the late 1950s, biologists have equated the "*precise* determination of sequence" with the extra-information theoretic property of specificity or specification. Biologists have defined specificity tacitly as 'necessary to achieve or maintain function.' They have determined that DNA base sequences (for example) are specified, not by applying information theory, but by making assessments (experimentally) of the function of DNA sequences within the overall apparatus of gene expression.[iv] Similar experimental considerations established the functional specificity of proteins. Even so, developments in complexity theory have now made possible a fully general theoretical account of specification—indeed, one that applies readily to biological systems (see below). In particular, recent work by the mathematician William Dembski has employed the notion of a rejection region from statistics to provide a formal complexity-theoretic account of specification. According to Dembski, a specification occurs when (a) an event or object falls within an independently given pattern or domain or (b) when an object or event "matches" or exemplifies a (conditionally) independent pattern or (c) meets a conditionally independent set of functional requirements [33, pp. 1-35, 136-74].

To illustrate Dembski's notion of specification consider these two strings of characters:

"iuinsdysk]idfawqnzkl,mfdifhs"

"Time and tide wait for no man."

Given the number of possible ways of arranging the letters and punctuation marks of the English language for sequences of this length, both these two sequences constitute highly improbable arrangements of characters. Thus, both have a vast and quantifiable information carrying capacity. Nevertheless, only the second of these two sequences exhibits a specification on Dembski's account. To see why consider the following. Within the set of combinatorially possible sequences only a very few will convey meaning. This smaller set of meaningful sequences, therefore, delimits a domain or pattern within the larger set of the totality of possibilities. Moreover, this set constitutes a "conditionally independent" pattern. Roughly speaking, a conditionally independent pattern corresponds to a pre-existing pattern or set of functional requirements, not one

contrived after the fact of observing the event in question, specifically, in this case, the event of observing the two sequences above [33, pp. 136-74]. Since the smaller domain distinguishes functional from non-functional English sequences, and the functionality of alphabetic sequences depends upon the pre-existing or independently given conventions of English vocabulary and grammar, the smaller set or domain qualifies as a conditionally independent pattern.[v] Since the second string of characters ("Time and tide wait. . .") falls within this smaller conditionally independent domain, (or "matches" one of the possible meaningful sentences that fall within it), the second sequence exhibits a specification according to Dembski's complexity-theoretic account of the concept. The second sequence, therefore exhibits the joint properties of complexity and specification, and possesses not just "information carrying capacity," but both "specified" and "semantic" information.

Biological organisms also exhibit specifications, though not necessarily semantic or subjectively "meaningful" ones. The nucleotide base sequences in the coding regions of DNA are highly specific relative to the independent functional requirements of protein function, protein synthesis and cellular life. To maintain viability the cell must regulate its metabolism, pass materials back and forth across its membranes, destroy waste materials, and many other specific tasks. Each of these functional requirements in turn necessitates specific molecular constituents, machines or systems (usually made of proteins) to accomplish these tasks. As noted, for a protein to perform a particular function within the cell it must have a very specific three-dimensional shape and a specific arrangement of amino acids. To build functional proteins in turn requires specific arrangements of nucleotide bases on the DNA molecule.

Nevertheless, the chemical properties of DNA allow a vast ensemble of combinatorially possible arrangements of nucleotide bases. Thus, any particular sequence will necessarily be highly improbable and rich in (Shannon) information or information carrying capacity. Yet within this set of possible sequences a very few will (given the multimolecular system of gene expression within the cell) produce functional proteins [34-36]. Those that do are thus, not only improbable, but also functionally "specified" or "specific" as molecular biologists use the terms. Indeed, the smaller set of functionally-efficacious sequences again delimits a domain or pattern within the larger set of combinatorial possibilities. Moreover, this smaller domain constitutes a conditionally independent pattern, since (as with the English sequences above) it distinguishes functional from non-functional sequences, and the functionality of nucleotide bases sequences depends upon the independent requirements of protein function. Thus, any actual nucleotide sequence that falls within this domain (or "matches" one of the possible functional sequences that fall within it), exhibits a specification. Or put differently, any nucleotide base sequence that produces a functional protein clearly meets certain independent functional requirements, in particular, those of protein function. Thus, any sequence that meets such requirements (or "falls within the smaller subset of functional sequences"), is again, not only highly improbable, but also specified relative to that independent pattern or domain. Thus, the nucleotide sequences

in the coding regions of DNA not only possess "syntactic information;" they also have "specified" information.

One final note of definitional clarity must offered about the relationship between "specified" information and "semantic information." Though both natural languages and the DNA base sequences are specified, only natural language conveys meaning. If one defines "semantic information" as 'subjectively meaningful information that is conveyed syntactically (as string of phonemes or characters) and that is understood by a conscious agent,' then clearly the information in DNA does not qualify as semantic. Indeed, unlike a written or spoken natural language, DNA does not convey "meaning" to a conscious agent.

Rather the coding regions of DNA function in much the same way as a software program or machine code, directing operations within a complex material system via highly complex yet specified sequences of characters. As Richard Dawkins has noted, "The machine code of the genes is uncannily computer-like" [37, p. 10]. Or as the software developer Bill Gates has noted, "DNA is like a computer program, but far, far more advanced than any software we've ever created" [38, p. 228]. Just as the specific arrangement of two symbols (0 and 1) in a software program can perform a function within a machine environment, so too can the precise sequencing of the four nucleotide bases in DNA perform a function within the cell.

Thus, though DNA sequences do not convey "meaning," they do exhibit specificity or specification. Moreover, as in a machine code, the sequence specificity of DNA occurs within a syntactic (or functionally alphabetic) domain. Thus, DNA possesses both syntactic and specified information. In any case, since the late 1950s the concept of information as employed by molecular biologists has comprised the joint notions of complexity (or improbability) and specificity (of function). The crucial biomolecular constituents of living organisms possess, therefore, not only Shannon or syntactic information, but also "*specified* information" or "*specified* complexity" [39, p. 189]. Biological information so defined, therefore, constitutes a salient feature of living systems that any origin-of-life scenario must explain "the origin of." Further, as we will discuss below (in 3.1-3.5), though DNA and proteins do not convey meaningful or semantic information, the kind of information that DNA does possess—namely, functionally "specified" information—has more than sufficed to defy explanation by reference to naturalistic chemical evolutionary theories.

2.6 Information as Metaphor: Nothing to Explain?

Though most molecular biologists would regard the characterization of DNA and proteins as "information-bearing" molecules as noncontroversial, some historians and philosophers of biology have recently challenged this description. Before evaluating competing types of explanation for the origin of biological information, this challenge must be addressed. Recently, historian of science Lily Kay has characterized the application of information theory to biology as a failure (in particular) because classical information theory could not capture the idea of meaning [8-10]. She suggests, therefore,

that the term 'information' as used in biology constitutes nothing more than a metaphor. Since, in Kay's view, the term does not designate anything real, it follows that the origin of 'biological information' does not require explanation. [8-10]. Instead, only the origin of the *use* of the term 'information' within biology requires explanation. As a social constructivist, Kay explains this as the result of various social forces operating within the "Cold War Technoculture" [8, pp. 611-12, 629; 9; 10]. In a different but related vein, Sahotra Sarkar has argued that the concept of information has little theoretical significance in biology because it lacks predictive and explanatory power [31, pp. 199-202]. He, like Kay, seems to regard the concept of information as a superfluous metaphor that lacks empirical reference and ontological status.

Of course, insofar as the term 'information' connotes semantic meaning, it does function, as a metaphor within biology. Nevertheless, this does not mean that the term *only* functions metaphorically or that origin-of-life biologists have nothing to explain. Though information theory had a *limited* application in describing biological systems, it has succeeded in rendering quantitative assessments of the complexity of biomacromolecules. Further, experimental work established the functional specificity of the sequencing of monomers in DNA and proteins. Thus, the term 'information' as used in biology does refers to two real and contingent properties—complexity and specificity. Indeed, since scientists began to think seriously about what would be required to explain the phenomenon of heredity, they have recognized the need for some feature or substance in living organisms possessing precisely these two properties together. Thus, Schrodinger envisioned an "aperiodic crystal" [40]; Chargaff perceived DNA's capacity for "complex sequencing" [20, p. 21]; Watson and Crick equated complex sequencing with "information," which Crick in turn equated with "specificity" [3, 4, 32]; Monod equated irregular specificity in proteins with the need for "a code" [14, p. 611]**;** and Orgel characterized life as a "specified complexity" [39, p. 189]. Further, Davies has recently argued that the "specific randomness" of DNA base sequences constitutes the central mystery surrounding the origin of life [41, p. 120]. Whatever the terminology, scientists have recognized the need for, and now know the location of, a source of complex specificity in the cell in order to transmit heredity and maintain biological function. The incorrigibility of these descriptive concepts suggests that complexity and specificity constitute real properties of biomacromolecules—indeed, properties that could be otherwise but only to the detriment of cellular life. As Orgel notes:

> Living organisms are distinguished by their specified complexity. Crystals. . .fail to qualify as living because they lack complexity; mixtures of random polymers fail to qualify because they lack specificity. [39, p. 189]

The origin of specificity and complexity (in combination), to which the term 'information' in biology commonly refers, therefore, does require explanation, even if it connotes only complexity in classical information theory, and even if the concept of information does not have any explanatory or predictive value in itself. Instead, as a descriptive (rather than an explanatory or predictive) concept, the term 'information' helps to define (either in conjunction with the notion of "specificity," or by subsuming it)

the character of the effect that origin of life researchers must explain. Thus, *only* where information connotes subjective meaning does it function as a metaphor in biology. Where it refers to an analogue of meaning, namely, functional specificity, it defines an essential feature of living systems that biologists must (in conjunction with complexity) explain "the origin of."

3.1 Naturalistic Explanations for the Origin Of Specified Biological Information

The discoveries of molecular biologists during the 1950s and 1960s raised the question of the ultimate origin of the specified complexity or specified information in both DNA and proteins. Since at least the mid-1960s many scientists have regarded the origin of information (so defined) as the central question facing origin-of-life biology [6; 41; 5; 42, p. 190; 43, pp. 287-340; 30, pp. 178-293; 7, pp. 170-72; 44, pp. 59-60, 88; 45; 39, p. 189; 46, pp. 199-211, 263-66; 2, pp. 146-47; 47]. Accordingly, origin-of-life researchers have proposed three broad types of naturalistic explanation to explain the origin of specified genetic information: those emphasizing chance, necessity, or the combination of the two.

3.2 Beyond the Reach of Chance

Perhaps the most common popular view about the origin of life is that it happened exclusively by chance. A few serious scientists have also voiced support for this view, at least, at various points during their careers. In 1954 the physicist George Wald, for example, argued for the causal efficacy of chance in conjunction vast expanses of time. As he explained, "Time is in fact the hero of the plot. . . . Given so much time, the impossible becomes possible, the possible probable, and the probable virtually certain" [48; 49, p. 121]. Later in 1968 Francis Crick would suggest that the origin of the genetic code—i.e., the translation system—might be a "frozen accident" [50, 51]. Other theories have invoked chance as an explanation for the origin of genetic information though often in conjunction with pre-biotic natural selection. (see below 3.3)

While outside origin-of-life biology some may still invoke 'chance' as an explanation for the origin of life, most serious origin-of-life researchers now reject it as an adequate causal explanation for the origin of biological information [52; 44, pp. 89-93; 47, p. 7]. Since molecular biologists began to appreciate the sequence specificity of proteins and nucleic acids in the 1950s and 1960s, many calculations have been made to determine the probability of formulating functional proteins and nucleic acids at random. Various methods of calculating probabilities have been offered by Morowitz, Hoyle and Wickramasinghe, Cairns-Smith, Prigogine, Yockey, and more recently, Robert Sauer [53, pp. 5-12; 54, pp. 24-27; 55, pp. 91-96; 56; 30, pp. 246-58; 57; 34; 35; 36; 49, pp. 117-31]. For the sake of argument, these calculations have often assumed extremely favorable prebiotic conditions (whether realistic or not), much more time than was actually available on the early earth, and theoretically maximal reaction rates among constituent

monomers (i.e., the constituent parts of proteins, DNA and RNA). Such calculations have invariably shown that the probability of obtaining functionally sequenced biomacromolecules at random is, in Prigogine's words, "vanishingly small . . .even on the scale of . . .billions of years" [56]. As Cairns-Smith wrote in 1971:

> Blind chance...is very limited. Low-levels of cooperation he [blind chance] can produce exceedingly easily (the equivalent of letters and small words), but he becomes very quickly incompetent as the amount of organization increases. Very soon indeed long waiting periods and massive material resources become irrelevant. [55, p. 95]

Consider the probabilistic hurdles that must be overcome to construct even one short protein molecule of one hundred amino acid in length. (A typical protein consists of about 300 amino acid residues, and many crucial proteins are very much longer.) [18, p. 118].

First, all amino acids must form a chemical bond known as a peptide bond so as to join with other amino acids in the protein chain. Yet in nature many other types of chemical bonds are possible between amino acids; in fact, peptide and non-peptide bonds occur with roughly equal probability. Thus, at any given site along a growing amino acid chain the probability of having a peptide bond is roughly 1/2. The probability of attaining four peptide bonds is: (1/2 x 1/2 x 1/2 x 1/2)=1/16 or $(1/2)^4$. The probability of building a chain of 100 amino acids in which all linkages involve peptide linkages is $(1/2)^{99}$ or roughly 1 chance in 10^{30}.

Second, in nature every amino acid has a distinct mirror image of itself, one left-handed version or L-form and one right-handed version or D-form. These mirror-image forms are called optical isomers. Functioning proteins tolerate only left-handed amino acids, yet the right-handed and left-handed isomers occur in nature with roughly equal frequency. Taking this into consideration compounds the improbability of attaining a biologically functioning protein. The probability of attaining at random only L-amino acids in a hypothetical peptide chain 100 amino acids long is $(1/2)^{100}$ or again roughly 1 chance in 10^{30}. The probability of building a 100 amino acid length chain at random in which all bonds are peptide bonds and all amino acids are L-form is, therefore, roughly 1 chance in 10^{60}.

Functioning proteins have a third independent requirement, the most important of all; their amino acids must link up in a specific sequential arrangement just as the letters in a meaningful sentence must. In some cases, even changing one amino acid at a given site can result in loss of protein function. Moreover, because there are twenty biologically occurring amino acids, the probability of getting a specific amino acid at a given site is small, i.e. 1/20. (Actually the probability is even lower because there are many non-proteineous amino acids in nature). On the assumption that all sites in a protein chain require one particular amino acid, the probability of attaining a particular protein 100

amino acids long would be $(1/20)^{100}$ or roughly 1 chance in 10^{130}. We know now, however, that some sites along the chain do tolerate several of the twenty proteineous amino acids, while others do not. The biochemist Robert Sauer of M.I.T has used a technique known as "cassette mutagenesis" to determine how much variance among amino acids can be tolerated at any given site in several proteins. His results have shown that, even taking the possibility of variance into account, the probability of achieving a functional sequence of amino acids[vi] in several known (roughly 100 residue) proteins at random is still "vanishingly small," about 1 chance in 10^{65}—an astronomically large number [36; 58: 59; 60; 30, pp. 246-58]. (There are 10^{65} atoms in our galaxy) [60]. Recently, Doug Axe of Cambridge University has used a refined mutagenesis technique to measure the sequence specificity of the protein Barnase (a bacterial RNase). Axe's work suggests that previous mutagenesis experiments actually underestimated the functional sensitivity of proteins to amino acid sequence change because they presupposed (incorrectly) the context independence of individual residue changes [58]. If, in addition to the improbability of attaining proper sequencing, one considers the need for proper bonding and homochirality, the probability of constructing a rather short functional protein at random becomes so small (no more than 1 chance in 10^{125}) as to appear absurd on the chance hypothesis. As Dawkins has said, "we can accept a certain amount of luck in our explanations, but not too much" [37, pp. 54, 139].

Of course, this assertion begs a quantitative question, namely, "how improbable does an event, sequence or system have to be before the chance hypothesis can be reasonably eliminated?" This question has recently received a formal answer. William Dembski, following and refining the work of earlier probabilists such as Emile Borel, has shown that chance can be eliminated as a plausible explanation for specified systems of small probability, whenever the complexity of a specified event or sequence exceeds available probabilistic resources [33, pp. 175-223; 61, p. 28].[vii] He then calculates a (conservative estimate for the) universal probability bound of 1 in 10^{150} corresponding to the probabilistic resources of the known universe. This number provides a theoretical basis for excluding appeals to chance as the best explanation for specified events of probability less than $1/2 \times 1/10^{150}$. Dembski, thus, answers the question: "how much luck is, in any case, too much to invoke in a explanation?"

Significantly, the improbability of assembling and sequencing even a short functional protein approaches this universal probability bound—the point at which appeals to chance become absurd given the "probabilistic resources" of the entire universe [33, pp. 175-223]. Further, making the same kind of calculation for even moderately longer proteins pushes these measures of improbability well beyond this limit. For example, the improbability of generating a protein of only 150 amino acids in length exceeds (using the same method as above)[viii] 1 chance in 10^{180}, well beyond the most conservative estimates of the small probability bound given our multi-billion year old universe [33, pp. 67-91, 175-214; 61, p. 28]. Thus, given the complexity of proteins, it is extremely unlikely that a random search through the space of combinatorially possible amino acid sequences could generate even a single relatively short functional protein in the time

available since the beginning of the universe (let alone the time available on the early earth). Conversely, to have a reasonable chance of finding a short functional protein in a random search of combinatorial space would require vastly more time than either cosmology or geology allows.

Yet more realistic calculations (taking into account the probable presence of non-proteineous amino acids, the need for vastly longer functional proteins to perform specific functions such as polymerization, and the need for multiple proteins functioning in coordination) only compound these improbabilities—indeed, almost beyond computability. For example, recent theoretical and experimental work on the so-called "minimal complexity" required to sustain the simplest possible living organism suggests a lower bound of some 250-400 genes and their corresponding proteins [62, 63, 64]. The nucleotide sequence space corresponding to such a system of proteins exceeds 4^{300000}. The improbability corresponding to this measure of molecular complexity again vastly exceeds 1 chance in 10^{150}, and thus the 'probabilistic resources' of the entire universe [33, pp. 67-91, 175-223, 209-10]. Thus, when one considers the full complement of functional biomolecules required to maintain minimal cell function and vitality, one can see why chance-based theories of the origin of life have been abandoned. What Mora said in 1963 still holds:

> Statistical considerations, probability, complexity, etc., followed to their logical implications suggest that the origin and continuance of life is not controlled by such principles. An admission of this is the use of a period of practically infinite time to obtain the derived result. Using such logic, however, we can prove anything. [65, pp. 212-19]

Though the probability of assembling a functioning biomolecule or cell by chance alone is exceedingly small, it is important to emphasize that scientists have not generally rejected the chance hypothesis merely because of the vast improbabilities associated with these events. Very improbable things do occur by chance. Any hand of cards or any series of rolled dice, will represent a highly improbable occurrence. Observers often justifiably attribute such events to chance alone. What justifies the elimination of the chance is not just the occurrence of a highly improbable event, but the occurrence of an improbable event that also conforms to a discernible pattern, (indeed, to a conditionally *independent* pattern, see section 2.5). If someone repeatedly rolls two dice and turns up a sequence such as: 9, 4, 11, 2, 6, 8, 5, 12, 9, 2, 6, 8, 9, 3, 7, 10, 11, 4, 8 and 4, no one will suspect anything but the interplay of random forces, though this sequence does represent a very improbable event given the number of combinatorial possibilities that correspond to a sequence of this length. Yet rolling twenty (or certainly 200) consecutive sevens will justifiably arouse suspicion that something more than chance is in play. Statisticians have long used a method for determining when to eliminate the chance hypothesis that involves pre-specifying a pattern or "rejection region" [66, pp. 74-75]. In the dice example above one could pre-specify the repeated occurrence of seven as such a pattern in order to detect the use of loaded dice, for example. Dembski has generalized this method to show how the presence of any conditionally independent pattern, whether

temporally prior to the observation of an event or not, can help (in conjunction with a small probability event) to justify rejecting the chance hypothesis [33, pp. 47-55].

Origin of life researchers have tacitly, and sometimes explicitly, employed this kind of statistical reasoning to justify the elimination of scenarios that rely heavily on chance. Christian de Duve, for example, has recently made this logic explicit in order to explain why chance fails as an explanation for the origin of life:

> A single, freak, highly improbable event can conceivably happen. Many highly improbable events—drawing a winning lottery number or the distribution of playing cards in a hand of bridge—happen all the time. But a string of improbable events—drawing the same lottery number twice, or the same bridge hand twice in a row—does not happen naturally. [67, p. 437]

De Duve and other origin-of-life researchers have long recognized that the cell represents not only a highly improbable, but also a functionally specified system. For this reason, by the mid-1960s most researchers had eliminated chance as a plausible explanation for the origin of the specified information necessary to build a cell [47, p. 7]. Many have instead sought other types of naturalistic explanations (see below).

3.3 Pre-Biotic Natural Selection: A Contradiction in Terms

Of course, even early theories of chemical evolution did not rely exclusively on chance as a causal mechanism. For example, A.I. Oparin's original theory of evolutionary abiogenesis first published in the 1920s and 30s invoked prebiotic natural selection as a complement to chance interactions. Oparin's theory envisioned a series of chemical reactions that he thought would enable a complex cell to assemble itself gradually and naturalistically from simple chemical precursors.

For the first stage of chemical evolution, Oparin proposed that simple gases such as ammonia (NH_3), methane (CH_4), water (H_20), carbon dioxide (CO_2) and hydrogen (H_2) would have rained down to the early oceans and combined with metallic compounds extruded from the core of the earth [13, pp. 64-103; 14, pp. 174-79, 194-98, 211-12]. With the aid of ultraviolet radiation from the sun, the ensuing reactions would have produced energy-rich hydrocarbon compounds [13, pp. 107-08]. These in turn would have combined and recombined with various other compounds to make amino acids, sugars, phosphates and other 'building blocks' of the complex molecules (such as proteins) necessary to living cells [13, pp. 133-35]. These constituents would eventually arrange themselves by chance into primitive metabolic systems within simple cell-like enclosures that Oparin called coacervates [13, pp. 148-59]. Oparin then proposed a kind of Darwinian competition for survival among his coacervates. Those that, by chance, developed increasingly complex molecules and metabolic processes would have survived to grow more complex and efficient. Those that did not would have dissolved [13, pp. 195-96]. Thus, Oparin invoked differential survival or natural selection as a mechanism

for preserving complexity-increasing events, thus allegedly helping to overcome the difficulties attendant pure chance hypotheses.

Nevertheless, developments in molecular biology during the 1950s cast doubt on Oparin's scenario. Oparin originally invoked natural selection to explain how cells refined primitive metabolism once it had arisen. His scenario relied heavily, therefore, on chance to explain the initial formation of the constituent biomacromolecules upon which any cellular metabolism would depend. The discovery of the extreme complexity and specificity of these molecules during the 1950s undermined the plausibility of this claim. For this and other reasons, Oparin published a revised version of his theory in 1968 that envisioned a role for natural selection earlier in the process of abiogenesis. His new theory claimed that natural selection acted upon random polymers as they formed and changed within his coacervate protocells [2, pp. 146-47]. As more complex and efficient molecules accumulated, they would have survive and reproduce more prolifically.

Even so, Oparin's concept of *pre-biotic* natural selection acting on initially unspecified biomacromolecules remained problematic. For one thing, it seemed to presuppose a pre-existing mechanism of self-replication. Yet self-replication in all extant cells depends upon functional and, therefore, (to a high degree) sequence-specific proteins and nucleic acids. Yet the origin of specificity in these molecules is precisely what Oparin needed to explain. As Christian de Duve has explained, theories of pre-biotic natural selection "need information which implies they have to presuppose what is to be explained in the first place" [68, p. 187]. Oparin attempted to circumvent this problem by claiming that the first polymers need not have been highly sequence specific. But this claim raised doubts about whether an accurate mechanism of self-replication (and thus, natural selection) could have functioned at all. Oparin's scenario did not reckon on a phenomenon known as "error catastrophe" in which small errors, or deviations from functionally necessary sequencing, are quickly amplified in successive replications [69, pp. 8-13].

Thus, the need to explain the origin of specified information created an intractable dilemma for Oparin. On the one hand, if he invoked natural selection late in his scenario, then he would need to rely on chance alone to produce the highly complex and specified biomolecules necessary to self-replication. On the other hand, if Oparin invoked natural selection earlier in the process of chemical evolution, before functional specificity in biomacromolecules would have arisen, he could give no account of natural selection could even function. Natural selection presupposes self-replication system, but self-replication requires functioning nucleic acids and proteins (or molecules approaching their complexity)—the very entities Oparin needed to explain. Thus, Dobzhansky would insist that, "prebiological natural selection is a contradiction in terms" [72, 73].

While some rejected the hypothesis of pre-biotic natural selection as question begging, others dismissed it as indistinguishable from implausible chance-based hypotheses [70; 71, p. 82]. The work of the mathematician Von Neumann supported this judgment. Von Neumann showed during 1960s that any system capable of self-replication would require sub-systems that were functionally equivalent to the

information storage, replicating and processing systems found in extant cells [74]. His calculations established a very high minimal threshold of biological function as would later experimental work [62, 63, 64]. These minimal complexity requirements pose a fundamental difficulty for natural selection. Natural selection selects for functional advantage. It can play no role, therefore, until random variations produce some biologically advantageous arrangement of matter. Yet, Von Neuman's calculations (and similar ones) by Wigner, Landsberg, and Morowitz, showed that random fluctuations of molecules in all probability (to understate the case) would not produce the minimal complexity needed for even a primitive replication system [75; 76; 77; 53, pp. 10-11]. As noted above, the improbability of developing a functionally integrated replication system vastly exceeds the improbability of developing the protein or DNA components of such a system. Given this improbability, and the high functional threshold it implies, many origin-of-life researchers came to regard pre-biotic natural selection as both inadequate and essentially indistinguishable from appeals to chance.

Nevertheless, during the 1980s Richard Dawkins and Bernd-Olaf Kuppers attempted to resuscitate pre-biotic natural selection as an explanation for the origin of biological information [37, pp. 47-49; 28] . Both accept the futility of naked appeals to chance and invoke what Kuppers calls a "Darwinian optimization principle." Both use a computer to demonstrate the efficacy of pre-biotic natural selection. Each selects a target sequence to represent a desired functional polymer. After creating a crop of randomly constructed sequences, and generating variations among them at random, their computers select those sequences that match the target sequence most closely. The computers then amplify the production of those sequences, eliminate the others (to simulate differential reproduction) and repeat the process. As Kuppers puts it, "Every mutant sequence that agrees one bit better with the meaningful or reference sequence. . .will be allowed to reproduce more rapidly" [28, p. 366]. In his case, after a mere 35 generations, his computer succeeds in spelling his target sequence, "NATURAL SELECTION."

Despite superficially impressive results, these 'simulations' conceal an obvious flaw: molecules *in situ* do not have a target sequence 'in mind.' Nor will they confer any selective advantage on a cell, and thus differentially reproduce, until they combine in a functionally advantageous arrangement. Thus, nothing in nature corresponds to the role that the computer plays in selecting functionally non-advantageous sequences that happen to agree 'one bit better' than others with a target sequence. The sequence 'NORMAL ELECTION' may agree more with 'NATURAL SELECTION' than does the sequence 'MISTRESS DEFECTION,' but neither of the two yield any advantage in communication over the other, if, that is, we are trying to communicate something about 'NATURAL SELECTION.' If so, both are equally ineffectual. Even more to the point, a completely non-functional polypeptide would confer no selective advantage on a hypothetical proto-cell, even if its sequence happens to 'agree one bit better' with an unrealized target protein than some other nonfunctional polypeptide.

And, indeed, both Kuppers's and Dawkins's published results of their simulations show the early generations of variant phrases awash in non-functional gibberish [28, p.

366; 37, pp. 47-49; 78]. In Dawkins's simulation, not a single functional English word appears until after the tenth iteration (unlike the more generous example above that starts with actual, albeit incorrect, words). Yet to make distinctions on the basis of function among sequences that have no function whatsoever would seem quite impossible. Such determination can only be made if considerations of proximity to possible future function are allowed, but this requires foresight that natural selection does not have. But a computer, programmed by a human being, can perform these functions. To imply that molecules can as well only illicitly personifies nature. Thus, if these computer simulations demonstrate anything, they subtly demonstrate the need for intelligent agents to elect some options and exclude others—that is, to create information.

3.4 Self-Organizational Scenarios

Because of the difficulties with chance-based theories, including those that rely upon pre-biotic natural selection, most origin-of-life theorists after the mid-1960s attempted to address the problem of the origin of biological information in a completely different way. Researchers began to look for self-organizational laws and properties of chemical attraction that might explain the origin of the specified information in DNA and proteins. Rather than invoking chance, these theories invoked necessity. Indeed, if neither chance nor pre-biotic natural selection acting on chance explains the origin of specified biological information, then those committed to finding a naturalistic explanation for the origin of life necessarily must rely on physical or chemical necessity. Given a limited number of broad explanatory categories, the inadequacy of chance (with or without pre-biotic natural selection), has, in the minds of many researchers, left only one option. Christian de Duve articulates the logic:

> a string of improbable events—drawing the same lottery number twice, or the same bridge hand twice in a row—does not happen naturally. All of which lead me to conclude that life is an obligatory manifestation of matter, bound to arise where conditions are appropriate. [67, p. 437]

By the late 1960s origin-of-life biologists began to consider the self-organizational perspective that de Duve describes. At that time, several researchers began to propose that deterministic forces (stereochemical 'necessity') made the origin of life not just probable, but inevitable. Some suggested that simple chemicals might possess "self-ordering properties" capable of organizing the constituent parts of proteins, DNA and RNA into the specific arrangements they now possess [53, pp. 5-12]. Steinman and Cole, for example, suggested that differential bonding affinities or forces of chemical attraction between certain amino acids might account for the origin of the sequence specificity of proteins [79, 80, 81]. Just as electrostatic forces draw sodium (Na+) and chloride ions (Cl-) together into a highly-ordered patterns within a crystal of salt (NaCl), so too might amino acids with special affinities for each other arrange themselves to form proteins. Kenyon and Steinman developed this idea in a book entitled *Biochemical Predestination*

in 1969. They argued that life might have been "biochemically predestined" by the properties of attraction that exist between its constituent chemical parts, particularly between the amino acids in proteins [46, pp. 199-211, 263-66].

In 1977, another self-organizational theory was proposed by Prigogine and Nicolis based on a thermodynamic characterization of living organisms. In *Self Organization in Nonequilibrium Systems*, Prigogine and Nicolis classified living organisms as open, nonequilibrium systems capable of "dissipating" large quantities of energy and matter into the environment [82, pp. 339-53, 429-47]. They observed that open systems driven far from equilibrium often display self-ordering tendencies. For example, gravitational energy will produce highly ordered vortices in a draining bathtub; thermal energy flowing through a heat sink will generate distinctive convection currents or "spiral wave activity." Prigogine and Nicolis argued that the organized structures observed in living systems might have similarly "self-originated" with the aid of an energy source. In essence, they conceded the improbability of simple building blocks arranging themselves into highly ordered structures under normal equilibrium conditions. But they suggested that, under non-equilibrium conditions, where an external source of energy is supplied, biochemical building blocks might arrange themselves into highly ordered patterns.

More recently, Kauffman and de Duve have proposed self-organizational theories with somewhat less specificity, at least with regard to the problem of the origin of specified genetic information [43, pp. 285-341; 67; 83]. Kauffman invokes so-called "autocatalytic properties" to generate metabolism directly from simple molecules. He envisions this autocatysis occurring once very particular configurations of molecules have arisen in a rich "chemical minestrone." De Duve also envisions proto-metabolism emerging first with genetic information arising later as a by-product of simple metabolic activity.

3.5 Order v. Information

For many current origin-of-life scientists self-organizational models now seem to offer the most promising approach to explaining the origin of specified biological information. Nevertheless, critics have called into question both the plausibility and the relevance of self-organizational models. Ironically, a prominent early advocate of self-organization, Dean Kenyon, has now explicitly repudiated such theories as both incompatible with empirical findings and theoretically incoherent [84, pp. v-viii; 85; 86; 87; 81].

First, empirical studies have shown that some differential affinities do exist between various amino acids (i.e., particular amino acids do form linkages more readily with some amino acids than others) [79, 80]. Nevertheless, these differences do not correlate to actual sequencing in large classes of known proteins [81]. In short, differing chemical affinities do not explain the multiplicity of amino acid sequences that exist in naturally occurring proteins or the sequential arrangement of amino acids in any particular protein.

In the case of DNA this point can be made more dramatically. Figure 2 shows that the structure of DNA depends upon several chemical bonds. There are bonds, for

example, between the sugar and the phosphate molecules that form the two twisting backbones of the DNA molecule. There are bonds fixing individual (nucleotide) bases to the sugar-phosphate backbones on each side of the molecule. There are also hydrogen bonds stretching horizontally across the molecule between nucleotide bases making so-called complementary pairs. These bonds, which hold two complementary copies of the DNA message text together, make replication of the genetic instructions possible. Most importantly, however, notice that there are *no* chemical bonds between the bases along the vertical axis in the center of the helix. Yet it is precisely along this axis of the molecule that the genetic information in DNA is stored [18, p. 105].

Further, just as magnetic letters can be combined and recombined in any way to form various sequences on a metal surface, so too can each of the four bases A, T, G, and C attach to any site on the DNA backbone with equal facility, making all sequences equally probable (or improbable). Indeed, there are no significant differential affinities between any of the four bases and the binding sites along the sugar-phosphate backbone. The same type of ('n-glycosidic') bond occurs between the base and the backbone regardless of which base attaches. All four bases are acceptable, none is preferred. As Kuppers has noted, "the properties of nucleic acids indicates that all the combinatorially possible nucleotide patterns of a DNA are, from a chemical point of view, equivalent" [28, p. 364]. Thus, 'self-organizing' bonding affinities can not explain the sequentially specific arrangement of nucleotide bases in DNA because: (1) there are *no* bonds between bases along the message-bearing axis of the molecule and, (2) there are no *differential* affinities between the backbone and the specific bases that could account for variations in sequencing. Because the same holds for RNA molecules, researchers who speculate that life began in an 'RNA world,' have also failed to solve the sequencing problem[ix]—i.e., the problem of explaining how information in all functioning RNA molecules could have arisen in the first place.

For those who want to explain the origin of life as the result of self-organizing properties intrinsic to the material constituents of living systems, these rather elementary facts of molecular biology have decisive implications. The most obvious place to look for self-organizing properties to explain the origin of genetic information is in the constituent parts of the molecules that carry that information. But biochemistry and molecular biology make clear that forces of attraction between the constituents in DNA, RNA and proteins do not explain the sequence specificity of these large information-bearing biomolecules.

We know this, in addition to the reasons already stated, because of the multiplicity of variant polypeptides and gene sequences that exist in nature and can be synthesized in the laboratory. The properties of the monomers constituting nucleic acids and proteins simply do not make a particular gene, let alone life as we know it, inevitable. Yet if self-organizational scenarios for the origin of biological information are to have any theoretical import, they must claim just the opposite. And, indeed, they often do, albeit without much specificity. As de Duve has put it, "the processes that generated life" were "highly deterministic" making life as we know it "inevitable" given "the conditions that

existed on the prebiotic earth" [67, p. 437]. Yet imagine the most favorable prebiotic conditions. Imagine a pool of all four DNA nucleotides, and all necessary sugars and phosphates; would any particular genetic sequence have to arise? Given all necessary monomers, would any particular functional protein or gene, let alone a specific genetic code, replication system or signal transduction circuitry, have to arise? Clearly not.

In the parlance of origin-of-life research, monomers are 'building blocks.' And building blocks can be arranged and rearranged in innumerable ways. The properties of blocks do not determine their arrangement in the construction of buildings. Similarly, the properties of *biological* building blocks do not determine the arrangement of functional polymers. Instead, the chemical properties of the monomers allow a vast ensemble of possible configurations, the overwhelming majority of which have no biological function whatsoever. Functional genes or proteins are no more inevitable given the properties of their "building blocks" than the palace of Versailles, for example, was inevitable given the properties of the bricks and stone used to construct it. To anthropomorphize, neither bricks and stone, nor letters in a written text, nor nucleotide bases 'care' how they are arranged. In each case, the properties of the constituents remain largely indifferent to the many specific configurations or sequences that they may adopt. Conversely, the properties of nucleotide bases and amino acids do not make any specific sequences 'inevitable' as self-organizationalists must claim.

Significantly, information theory makes clear that there is a good reason for this. If chemical affinities between the constituents in the DNA determined the arrangement of the bases, such affinities would dramatically diminish the capacity of DNA to carry information. Recall that classical information theory equates the reduction of uncertainty with the transmission of information, (whether specified or unspecified). The transmission of information, therefore, requires physical-chemical contingency. As Robert Stalnaker has noted, "[information] content requires contingency" [88, p. 85]. If, therefore, forces of chemical necessity completely determine the arrangement of constituents in a system, that arrangement will not exhibit complexity or convey information.

Consider, for example, what would happen if the individual nucleotide 'bases' (A, T, G, C) in the DNA molecule *did* interact by *chemical* necessity (along the information-bearing axis of DNA). Every time adenine (A) occurred in a growing genetic sequence, it would attract thymine (T) to it.[x] Every time cytosine (C) appeared, guanine (G) would likely follow. As a result, the longitudinal axis of DNA would be peppered with repetitive sequences of A's followed by T's and C's followed by G's. Rather than a genetic molecule capable of virtually unlimited novelty and characterized by unpredictable and aperiodic sequencing, DNA would contain sequences awash in repetition or redundancy—much like the sequences in crystals. In a crystal the forces of mutual chemical attraction do determine, to a very considerable extent, the sequential arrangement of its constituent parts. As a result, sequencing in crystals is highly ordered and repetitive, but neither complex nor informative. Once one has seen 'Na' followed by 'Cl' in a crystal of salt, for example, one has seen the extent of the sequencing possible.

In DNA, however, where any nucleotide can follow any other, a vast array of novel sequences are possible, corresponding to a multiplicity of amino acid sequences.

The forces of chemical necessity produce redundancy (roughly, law or rule generated repetition) or monotonous order, but reduce the capacity to convey information and express novelty. Thus, as the chemist Michael Polanyi noted:

> Suppose that the actual structure of a DNA molecule were due to the fact that the bindings of its bases were much stronger than the bindings would be for any other distribution of bases, then such a DNA molecule would have no information content. Its code-like character would be effaced by an overwhelming redundancy. . . Whatever may be the origin of a DNA configuration, it can function as a code only if its order is not due to the forces of potential energy. It *must be* as physically indeterminate as the sequence of words is on a printed page. [89, emphasis added]

In other words, if chemists had found that bonding affinities between the nucleotides in DNA produced nucleotide sequencing, they would have also found that they had been mistaken about DNA's information-bearing properties. Or, to put the point quantitatively, to the extent that forces of attraction between constituents in a sequence determine the arrangement of the sequence, to that extent will the information carrying capacity of the system be diminished or effaced (by redundancy).[xi] As Dretske has explained:

> As p(si) [the probability of a condition or state of affairs] approaches 1 the amount of information associated with the occurrence of si goes to 0. In the limiting case when the probability of a condition or state of affairs is unity [p(si) =1], no information is associated with, or generated by, the occurrence of si. This is merely another way to say that no information is generated by the occurrence of events for which there are no possible alternatives. [27, p. 12]

Bonding affinities, to the extent they exist, inhibit the maximization of information because they determine that specific outcomes will follow specific conditions with high probability [57, p. 18]. Yet information carrying capacity is maximized when just the opposite situation obtains, namely, when antecedent conditions allow many improbable outcomes.

Of course, as noted in 2.4, the bases sequences in DNA do not just possess information carrying capacity or syntactic information or as measured by classical Shannon information theory. These sequences store functionally specified information—that is, they are specified as well as complex. Clearly, however, a sequence cannot be both specified and complex, if it is not at least complex. Therefore, the self-organizational forces of chemical necessity that produce redundant order and

preclude complexity, also preclude the generation of specified complexity (or specified information) as well. Chemical affinities do not generate complex sequences. Thus, they cannot be invoked to explain the origin of information, whether specified or otherwise.

The tendency to conflate the qualitative distinctions between 'order' and 'complexity' has characterized self-organizational research efforts and calls into question the relevance of such work to the origin of life. As Yockey has argued, the accumulation of structural or chemical order does not explain the origin of biological complexity or genetic information. He concedes that energy flowing through a system may produce highly ordered patterns. Strong winds form swirling tornados and the 'eyes' of hurricanes; Prigogine's thermal baths do develop interesting 'convection currents'; and chemical elements do coalesce to form crystals. Self-organizational theorists explain well what does not need explaining. What needs explaining in biology is not the origin of order (defined as symmetry or repetition), but the specified information—the highly complex, aperiodic, and (yet specified) sequences that make biological function possible. As Yockey warns:

> Attempts to relate the idea of order …with biological organization or specificity must be regarded as a play on words which cannot stand careful scrutiny. Informational macromolecules can code genetic messages and therefore can carry information because the sequence of bases or residues is affected very little, if at all, by [self-organizing] physico-chemical factors. [90]

In the face of these difficulties, some self-organizational theorists have claimed that we must await the discovery of new natural laws to explain the origin of biological information. As Manfred Eigen has argued, "our task is to find an algorithm, a natural law, that leads to the origin of information" [91, p. 12]. But this suggestion betrays confusion on two counts. First, scientific laws don't generally explain or cause natural phenomena, they describe them. For example, Newton's law of gravitation described, but did not explain, the attraction between planetary bodies. Second, laws necessarily describe highly deterministic or predictable relationships between antecedent conditions and consequent events. Laws describe patterns in which the probability of each successive event (given the previous event and the action of the law) approaches unity. Yet information mounts as *im*probabilities multiply. Thus, to say that the that scientific laws describe complex informational patterns, is essentially a contradiction in terms. Instead, scientific laws describe (almost by definition) highly predictable and regular phenomena—i.e., redundant order, not complexity (whether specified or otherwise).

Though the patterns that natural laws describe display a high degree of regularity, and thus lack the complexity that characterizes information-rich systems, one could argue that we might someday discover a very particular configuration of *initial conditions* that routinely generates high informational states. Thus, while we cannot hope to find a law that describes a information-rich *relationship* between antecedent and consequent variables, we might find a law that describes how a very particular set of *initial*

conditions routinely generates a high information state. Unfortunately, however, even the statement of this hypothetical seems itself to beg the question of the ultimate origin of information, since "a very particular set of initial conditions" sounds precisely like an information rich—indeed, a highly complex and specified—state. In any case, everything we know experientially suggests that the amount of specified information present in a set of antecedent conditions necessarily equals or exceeds that of any system produced from these conditions.

3.6 Other Scenarios and the Displacement of the Information Problem

In addition to the general categories of explanation already examined, origin-of-life researchers have proposed many more specific scenarios, each emphasizing random variations (chance), self-organizational laws (necessity) or both. Some of these scenarios purport to address the information problem, while others attempt to by-pass it altogether. Yet on closer examination, even scenarios that appear to alleviate the problem of the origin of specified biological information merely shift the problem elsewhere. Genetic algorithms can "solve" the information problem, but only if programmers providing informative target sequences and selection criteria. Simulation experiments can produce biologically relevant precursors and sequences, but only if experimentalists manipulate initial conditions or select and guide outcomes—that is, only if they add information themselves. Origin of life theories can leapfrog the problem altogether, but only by presupposing the presence of information in some other pre-existing form. Such approaches "solve" the information problem only by shifting it elsewhere.

Any number of theoretical models for the origin of life have fallen prey to this difficulty. For example, in 1964 Henry Quastler, an early pioneer in the application of information theory to molecular biology, proposed a DNA-first model for the origin of life. He envisioned the initial emergence of a system of unspecified polynucleotides capable of primitive self-replication via the mechanisms of complementary base pairing. The polymers in this system would have, on Quastler's account, initially lacked specificity (which he equated with information) [47, p. ix]. Only later when this system of polynucleotides had come into association with a fully functional set of proteins and ribosomes would the specific nucleotide sequences in the polymers take on any functional significance. He likened this process to the random selection of a combination for a lock in which the combination would only later acquire functional significance once particular tumblers had been set to allow the combination to open the lock. In both the biological and the mechanical case, the surrounding context would confer functional specificity on an initially unspecified sequence. Thus, he characterized the origin of information in polynucleotides as an "accidental choice remembered."

Though this way of conceiving of the origin of specified biological information did allow "a chain of nucleotides [to] become a [functional] system of genes without necessarily suffering any change in structure" [47, p. 47], it did have an overriding difficulty. It did not account for the origin of the complexity and specificity of the system of molecules whose association with the initial sequence gave the initial sequence

functional significance. In Quastler's combination lock example, conscious agents choose the tumbler settings that made the initial combination functionally significant. Yet Quastler expressly precluded conscious design as a possibility for explaining the origin of life [47, p. 1]. Instead, he seems to suggest that the origin of the biological context—that is, the complete set of functionally specific proteins (and the translation system) necessary to create a "symbiotic association" between polynucleotides and proteins—would arise by chance. He even offered some rough calculations to show that the origin of this multi-molecular context, though improbable, would have been probable enough to expect it to occur by chance in the prebiotic soup. Quastler's calculations now seem extremely implausible in light of the discussion of minimal complexity in 3.2 [30, p. 247]. More significantly, Quastler only "solved" the problem of the origin of complex specificity in nucleic acids by transferring the problem to an equally complex and specified system of proteins and ribosomes. Whereas, admittedly, *any* polynucleotide sequence would suffice initially, the subsequent proteins and ribosomal material constituting the translation system would have to possess an extreme specificity *relative to the initial polynucleotide sequence* and relative to any proto-cellular functional requirements. Thus, Quastler's attempt to by-pass the sequencing problem merely shifted it elsewhere.

Self-organizational models have fallen prey to similar difficulties. For example, chemist J. C. Walton has argued (echoing earlier articles by Mora) that even the self-organizational patterns produced in Prigogine-style convection currents do not exceed the organization or structural information represented by the experimental apparatus used to create the currents [92; 70, p. 41]. Similarly, Maynard-Smith, Dyson, and Speigelman have shown that Manfred Eigen's so-called hypercycle model for generating biological information actually shows how information tends to degrade over time [93; 94, pp. 9-11, 35-39, 65-66, 78; 49, p. 161]. They note that Eigen's hypercycles presuppose a large initial contribution of information in the form of a long RNA molecule and some forty specific proteins, (and thus, does not attempt to explain the ultimate origin of biological information). They also show that because hypercycles lack an error-free mechanism of self-replication, this mechanism succumbs to various 'error-catastrophes' that ultimately diminish, not increase, the (specified) information content of the system over time.

Stuart Kauffman's self-organizational theory also subtly transfers the information problem. In *The Origins of Order*, Kauffman attempts to leapfrog the sequence specificity problem by proposing a means by which metabolism might emerge directly from molecules in a pre-biotic soup. He suggests that large ensembles of molecules in solution (in a so-called 'chemical minestrone') may have 'auto-catalytic' properties that might directly generate the integrated complexity of living cells [43, pp. 285-341]. He acknowledges, however, that such autocatalysis (for which there is as yet no experimental evidence) would not occur unless the molecules in the chemical minestrone achieve a very specific spatial-temporal relationships to one another. In other words, for the direct autocatalysis of integrated biological complexity to occur, a system of molecules must first achieve a very specific molecular configuration, or a low configurational entropy

state [84, pp. 127-43]. Yet this claim is isomorphic with the claim that the system must start with a high (specified) information content. Thus, to explain the origin of specified biological complexity at the systems-level, Kauffman must presuppose the existence of a highly specific and complex—i.e., an information-rich—arrangement of matter at the molecular level. Therefore, his work—if it has any relevance to the actual behavior of molecules—assumes rather than explains the ultimate origin of specified complexity or information.

Others have claimed that the so-called "RNA World" scenario offers a promising approach to origin of life problem, and with it, presumably, the problem of the origin of the first genetic information. Yet this claim is problematic on several counts. First, the RNA world was not proposed as an explanation for the sequencing or information problem. Rather it was proposed as an explanation for the origin of the interdependence of nucleic acids and proteins in the cell's information processing system. In extant cells, building proteins requires genetic information from DNA, but information on DNA cannot be processed without many specific proteins and proteins complexes. This poses a "chicken-or-egg" problem. The discovery that RNA (a nucleic acid) possesses some limited catalytic properties (similar to those of proteins) suggested a way to solve this problem. "RNA first" advocates proposed an early state in which RNA performed both the enzymatic functions of modern proteins and the information storage function of modern DNA, thus allegedly making the interdependence of DNA and proteins unnecessary in the earliest living system.

Nevertheless, there are many fundamental difficulties with the RNA world scenario. First, synthesizing (and/or maintaining) many essential building blocks of RNA molecules under realistic conditions has proven either difficult or impossible [95, 96]. Further, the chemical conditions required for the synthesis of ribose sugars are decidedly incompatible with the conditions required for synthesizing nucleoside bases [97, 85]. Yet both are necessary constituents of RNA. Second, naturally occurring RNA possesses very few of the specific enzymatic properties of the proteins that are necessary to extant cells. Third, RNA world advocates offer no plausible explanation for how primitive RNA replicators might have evolved into modern cells that do rely (almost exclusively) on proteins to process genetic information and regulate metabolism [98]. Fourth, attempts to enhance the limited catalytic properties of RNA molecules, inevitably have involved extensive investigator manipulation in so-called "Ribozyme engineering" experiments [99], thus simulating, if anything, the need for intelligent design, not the adequacy of an undirected chemical evolutionary process.

Most importantly for our present purposes, the RNA World hypothesis presupposes, but does not explain, the origin of sequence specificity or information in the original functional RNA molecules. Some RNA world theorists seem to envision leapfrogging the sequence specificity problem. They envision oligimers of RNA arising by chance on the pre-biotic earth and then later acquiring the ability to polymerize copies of themselves, that is, to self-replicate. In this scenario, the capacity to self-replicate would favor the survival of those RNA molecules that could do so, and would thus favor

the specific sequencing that the first self-replicating molecules happened to have. Thus, sequencing that originally arose by chance would subsequently acquire a functional significance as "an accidental choice remembered."

Like Quastler's DNA first model, however, this suggestion merely shifts the specificity problem out of view. First, for strands of RNA to perform enzymatic functions (including enzymatically-mediated self-replication) they must, like proteins, have very specific arrangements of constituent building blocks (in the RNA case, the nucleoside bases). Further, they must be long enough to fold into complex three-dimensional shapes (to form so-called tertiary structure). Thus, any RNA molecule capable of enzymatic function must have the same properties of complexity and specificity that DNA and proteins have. Indeed, such molecules must possess considerable (specified) information content. Nevertheless, explaining how the building blocks of RNA might have arranged themselves into functionally specified sequences has proven no easier than explaining how the constituent parts of DNA might have done so, especially given the high probability of destructive cross reactions between desirable and undesirable molecules in any realistic pre-biotic soup. As Christian de Duve has noted in critique of the RNA world hypothesis, "hitching the components together in the right manner raises additional problems of such magnitude that no one has yet attempted to do so in a prebiotic context" [83, p. 23].

Second, for a single stranded RNA-catalyst to self-replicate (which is the only function that could be selected in a pre-biotic environment) it must find an identical RNA molecule in close vicinity to function as a template, since a single stranded RNA cannot function as both enzyme and template. Thus, even if an originally unspecified RNA sequence might later acquire functional significance by chance, it could only perform a function if another RNA molecule—i.e., one with a highly specific sequence relative to the original—arose in close vicinity to it. Thus, the attempt to bypass (albeit unsuccessfully, see above) the need for specific sequencing in an original catalytic RNA, only shifts the specificity problem elsewhere, namely, to a second and necessarily highly specific RNA sequence. Put differently, in addition to the specificity required to give the first RNA molecule self-replicating capability, a second RNA molecule with an extremely specific sequence—indeed, one with precisely the same sequence as the original—would also have to arise. Yet RNA World theorists do not explain the origin of the requisite specificity in either the original molecule or its twin. Indeed, Joyce and Orgel [69, pp. 1-25, esp. p. 11] have calculated that to have a reasonable chance of finding two identical RNA molecules of a length sufficient to perform enzymatic functions would require a RNA library of some 10^{54} RNA molecules. The mass of such a library vastly exceeds the mass of the earth, suggesting the extreme implausibility of the chance origin of a primitive replicator system. Yet one cannot invoke natural selection to explain the origin of such primitive replicators, since natural selection only ensues once self-replication has arisen. Likewise, RNA bases, like DNA bases, do not manifest self-organizational bonding affinities that can explain their specific sequencing. In short, the same kind of evidentiary and theoretical problems emerge whether one proposes that

genetic information arose first in RNA or DNA molecules. Further, the attempt to leapfrog the sequencing problem by starting with RNA replicators only shifts the problem to the specific sequences that would make such replication possible.

4.1 The Return of The Design Hypothesis

If attempts to solve the information problem only relocate it, and if neither chance, nor physical-chemical necessity, nor the two acting in combination, explain the ultimate origin of specified biological information, what does? Do we know of any entity that has the causal powers to create large amounts of specified information? We do. As Henry Quastler recognized, the "creation of new information is habitually associated with conscious activity" [47, p. 16].

Experience affirms that specified complexity or information (so defined) routinely arises from the activity of intelligent agents. When a computer user traces the information on a screen back to its source, he invariably comes to a mind—a software engineer or programmer. Similarly, the information in a book or newspaper column ultimately derives from a writer—from a mental, not a material, cause. Our experience-based knowledge of information flow confirms that systems with large amounts of specified complexity or information (especially codes and languages) invariably[xii] originate from an intelligent source—i.e., from mental or personal agents. Moreover, this generalization holds not only for (the semantically) specified information present in natural languages, but also for other forms of specified complexity or information whether present in machine codes, machines or works of art. Like the letters in a section of meaningful text, the parts in a working engine represent a highly improbable and yet functionally specified configuration. Similarly, the highly improbable shapes in the rock on Mount Rushmore conform to an independently given pattern—the faces of American presidents known from books and paintings. Thus, both these systems have a large amount of *specified* complexity or information. Not coincidentally, they also originated by intelligent design, not by chance and/or physical-chemical necessity.

This generalization—that intelligence is the only known cause of specified complexity or information (at least, starting from a non-biological source, see endnote **xii** above)—has received support from origin-of-life research itself. During the last forty years, every naturalistic model proposed has failed precisely to explain the origin of the specified genetic information required to build a living cell [100; 30, pp. 259-93; 84, pp. 42-172; 42, pp. 193-97; 49]. Thus, mind or intelligence, or what philosophers call "agent causation," now stands as the only cause known to be capable of generating large amounts[xiii] of specified information (starting, at least, from a non-living system). As a result, the presence of a specified information-rich sequence or system provides a basis for inferring design.[xiv]

Recently, a formal theoretical account of such reasoning has been developed. In *The Design Inference*, mathematician and probability theorist William Dembski notes that rational agents often infer or detect the prior activity of other designing minds by the character of the effects they leave behind. Archaeologists assume, for example, that

rational agents produced the inscriptions on the Rosetta Stone. Insurance fraud investigators detect certain "cheating patterns" that suggest intentional manipulation of circumstances rather than "natural" disasters. Cryptographers distinguish between random signals and those that the carry encoded messages. Dembski's work shows that recognizing the activity of intelligent agents constitutes a common and fully rational mode of inference [33, pp. 1-35].

Moreover, Dembski provides a rational reconstruction of how such inferences are made. In the process, he identifies two criteria that typically enable human observers to recognize intelligent activity and to distinguish the effects of such activity from the effects of strictly material causes. He notices that we invariably attribute systems, sequences or events that have the joint properties of "high complexity" (or low probability) and "specification" [see section 2.5] to intelligent causes—to design—not chance or physical-chemical laws [33, pp. 1-35, 136-223]. By contrast, he notes that we typically attribute to chance those low or intermediate probability events that do not conform to discernable patterns. And we attribute to necessity highly probable events that result from natural regularities or laws. Furthermore, these inference patterns reflect our knowledge of the way the world works. Since, experience teaches, for example, that complex and specified events or systems invariably arise from intelligent causes, we invariably infer intelligent design when we encounter events that exhibit the joint properties of complexity and specificity. Dembski's work thus outlines a comparative evaluation process that provides criteria for decide between natural and intelligent causes based on the probabilistic features or "signatures" they leave behind [33, pp. 36-66]. This evaluation process constitutes, in effect, a method for detecting the activity of intelligence in the echo of its effects.

A homespun example illustrates this method as well as Dembski's theoretical criteria of design detection. When visitors first enter Victoria Harbor in Canada from the sea, they notice a hillside awash in red and yellow flowers. As they get closer, they naturally, and correctly, infer design. Why? Observers quickly recognize a complex and specified pattern—an arrangement of flowers spelling 'Welcome to Victoria.' They infer the past activity of an intelligent cause—in this case, the careful planning of gardeners. Had the flowers been more haphazardly scattered so as to defy pattern recognition, observers might have justifiably attributed the arrangement to chance—random gusts of wind scattering the seed, for example. Had the colors been segregated by elevation, the pattern might have been explained by some natural necessity—such as certain types of plants requiring particular environments or soil types. But since the arrangement exhibits both complexity (meaningful arrangements are highly improbable given the space of possible arrangements) and specificity (the pattern conforms to the independent requirements of English grammar and vocabulary), observers naturally infer intelligent design. As it turns out, these twin criteria are equivalent (or "isomorphic") with the notion of information as used in molecular biology. Thus, Dembski's theory, when applied to molecular biology, implies that intelligent design played a role in the origin of specified biological information.

In any case, even a pre-theoretic awareness of the connection between information and intelligence is sufficient to justify design as an inference to the best (or only causally adequate) explanation. Since, in our experience, mind or intelligent design is the only known cause of functionally-specified information-rich sequences, one can detect (or, retrodict) the past action of an intelligence from an information-rich effect, (even lacking a theory of design detection) and even if the cause itself cannot be directly observed [14, pp. 77-140]. Logically, one can infer a cause from its effect, (or an antecedent from a consequent), when the cause (or antecedent) is known to be necessary to produce the effect in question. If it's true that 'where's there's smoke there's fire' then the presence of smoke billowing over the hillside will allows us to infer a fire beyond our view. Since information requires an intelligent source, the pattern flowers spelling 'welcome to Victoria' will lead visitors to infer the activity of intelligent agents—even if they did not see the flowers planted or arranged. Similarly, the specified and complex arrangement of nucleotide sequences—the functionally specified information—in DNA implies the past action of an intelligent mind, even if the past action of such mental agency cannot be directly observed.

The logical calculus underlying such inferences follows a valid and well-established method used in all historical and forensic sciences. In historical sciences, knowledge of the present causal powers of various entities and processes enables scientists to make inferences about possible causes in the past. When a thorough study of various possible causes turns up just a single adequate cause for a given effect, historical or forensic scientists can make fairly definitive inferences about the past [14, pp. 77-140**;** 101, pp. 4-5; 102, pp. 249-50]. Several years ago, for example, one of the forensic pathologists from the original Warren Commission that investigated the assassination of President Kennedy spoke out to quash rumors about a second gunman firing from in front of the motorcade. Apparently, the bullet hole in the back of President Kennedy's skull evidenced a distinctive beveling pattern that clearly indicated its direction of entry. In particular, it revealed that the bullet had entered from the rear. The pathologist called the beveling pattern a "distinctive diagnostic" to indicate a necessary causal relationship between the direction of entry and the angle of the beveling [103]. Inferences based on knowledge of empirically necessary conditions or causes ("distinctive diagnostics") are common in historical and forensic sciences, and often lead to the detection of intelligent as well as natural causes and events. Since criminal X's fingers are the only known cause of criminal X's fingerprints, X's prints on the murder weapon incriminate him with a high degree of certainty. In the same way, since intelligent design is the only known cause of large amounts of specified information, the presence of such information implies an intelligent source.

Scientists in many fields recognize the connection between intelligence and specified information and make inferences accordingly. Archaeologists assume a mind produced the inscriptions on the Rosetta Stone. Evolutionary anthropologists argue for the intelligence of early hominids by showing that certain chipped flints are too improbably specified to have been produced by natural causes. N.A.S.A.'s search for extra-terrestrial

intelligence (S.E.T.I.)[xv] presupposed that specified information imbedded in electromagnetic signals from space would indicate an intelligent source [104].[xvi] As yet, however, radio-astronomers have not found such information-bearing signals coming from space. But closer to home, molecular biologists have identified specified informational sequences and systems in the cell, suggesting, by the same logic, an intelligent cause for these effects.

4.2 An Argument from Ignorance?

Of course, many would object that any such argument to design constitutes an argument from ignorance. Since, say objectors, we don't yet know how specified biological information could have arisen we invoke the mysterious notion of intelligent design. On this view, intelligent design functions, not as an explanation, but as a kind of placeholder for ignorance.

While admittedly the design inference does not provide a deductively certain proof (nothing based upon empirical observation can), it does not qualify as a fallacious argument from ignorance. Instead, the design inference from biological information constitutes an "inference to the best explanation" [105, pp. 32-88]. Recent work on the method of "inference to the best explanation" suggests that determining which among a set of competing of possible explanations constitutes the best depends upon knowledge of the causal powers of competing explanatory entities [105; 106; 107; 108; 14, p. 77-140]. Causes that have the capability to produce the evidence in question constitute better explanations of that evidence than those that do not. This essay has evaluated and compared the causal efficacy of four broad categories of explanation—chance, necessity, (and the combination) and design—with respect to their ability to produce large amounts of specified complexity or information. As we have seen, neither scenarios based upon chance nor those based upon necessity (nor those that combine the two) can explain the origin of specified biological information in a prebiotic context. This result comports with our ordinary uniform human experience. Matter—whether acting randomly or by necessity—does not have the capability to generate novel specified information.

Yet it is not correct to say that we do not know how specified information arises. We know from experience that conscious intelligent agents can create specified informational sequences and systems. To quote Quastler again, the "creation of new information is habitually associated with conscious activity" [47, p. 16]. Furthermore, experience teaches that whenever large amounts of specified information are present in an artifact or entity whose causal story is known, invariably creative intelligence—intelligent design—played a causal role in the origin of that entity. Thus, when we encounter such information in the bio-macromolecules necessary to life, we may infer based upon our present *knowledge* of established cause-effect relationships that an intelligent cause operated in the past to produce the specified information necessary to the origin of life.

Further, as noted above, we often infer the causal activity of intelligent agents as the best explanation for certain kinds of events and phenomena. Dembski's examples of design inferences—from archeology and cryptography to fraud detection and criminal

forensics—show that we make design inferences frequently and we do so, apparently, without worrying about committing fallacious arguments from ignorance. Moreover, we do so for good reason. Intelligent agents have unique causal powers that matter (especially non-living matter) does not. When we observe features or effects that, from experience, we know only agents produce, we rightly infer the prior activity of intelligence.

Thus, the inference to design does not depend upon our ignorance, but instead upon present knowledge of the demonstrated causal powers of natural entities and intelligent agency, respectively. Inferences to design, therefore, depend upon the standard uniformitarian methods of reasoning used in all historical sciences. These inferences do not constitute arguments from ignorance any more than other well-grounded inferences in geology, archeology or paleontology—where provisional knowledge of cause-effect relationships (derived from past or present experience) guides inferences about the causal past. Recent developments in the information sciences merely help define and formalize knowledge of these relationships, allowing us to make inferences about the causal histories of various artifacts, entities or events based upon the complexity and information-theoretic signatures they exhibit [33, pp. 36-66, esp. p. 37]. In any case, present knowledge of established cause-effect relationships, not ignorance, justifies the design inference as the best explanation for the origin of specified biological information in a prebiotic context.

Objectors complain, of course, that future inquiry may uncover other natural entities possessing as yet unknown causal powers. They object that the design inference presented here depends upon a negative generalization—purely physical and chemical causes cannot generate large amounts of specified information—that future discoveries may well later falsify. We should 'never say never,' they say. Yet science often says never, even if it can't say so for sure. Indeed, negative or proscriptive generalizations play an important role in science. As many scientists and philosophers of science have pointed out, scientific laws often tell us not only what does happen, but also what does not happen [13, p. 28; 109, pp. 65-92; 110, pp. 35-37]. The conservation laws in thermodynamics, for example, proscribe certain outcomes. The first law tells us that energy is never created or destroyed. The second tells us that the entropy of a closed system will never decrease over time. Those who claim that such 'proscriptive laws' do not constitute knowledge simply because they are based upon past, but not future, experience, will not get very far if they want to use their skepticism to justify funding for, say, research on perpetual motion machines.

Further, without proscriptive generalizations, without knowledge about what possible causes cannot or do not produce, historical scientists could not make determinations about the past. As work on the method of the historical sciences has shown, reconstructing the past requires making (abductive) inferences from present effects back to past causal events [14, pp. 77-140; 101, pp. 4-5; 67, pp. 249-50]. Historical scientists judge the plausibility of such inferences against experiential knowledge of the efficay of competing possible causes. Making inferences about the best historical explanation

requires a progressive elimination of competing causal hypotheses. Deciding which causes can be eliminated from consideration requires knowing what effects a given cause can—and cannot—produce. If historical scientists can never say that particular entities do not have particular causal powers, then they could never eliminate them—even provisionally—from consideration. Thus, they could never make historical inferences. Yet they do so all the time for good reason. To determine the best explanation scientists do not need to say 'never, for sure.' They only need to say that a postulated cause is best given what we know at present about the demonstrated causal powers of competing entities or agencies. That cause C can produce effect E, makes it a better explanation of E than some cause D that has never produced E (especially if D seems incapable of doing so on theoretical grounds), even if D may later demonstrate causal powers of which we are presently ignorant [cf: 111].

Thus, the objection that the design inference constitutes an argument from ignorance reduces in essence to a restatement of the problem of induction. Yet one can make this objection against any scientific law or explanation, or any historical inference that takes knowledge of natural laws and causal powers into account. As Barrow and Tipler have noted, to criticize design arguments, as Hume did, simply because they assume the uniformity and (normative character) of natural law cuts just as deeply against "the rational basis of any form of scientific inquiry" [112, p. 69]. Our knowledge of what can and cannot produce large amounts of specified information may later have to be revised, but so might the laws of thermodynamics. Inferences to design may also later prove incorrect, but so may inferences implicating various natural causes. Such a possibility does not stop scientists from making generalizations about the causal powers of various entities or using these generalizations to identify probable or most plausible causes in particular cases. Inferences based upon past and present experience constitute knowledge (albeit provisional), not ignorance. Those who object to such inferences object to *science* as much as they object to a particular science-based hypothesis of design.

4.3 But is it Science?

Of course, many simply refuse to consider the design hypothesis on the grounds that it does not qualify as "scientific." Such critics affirm an extra-evidential principle known as "methodological naturalism." [113; 106; 107]. Methodological naturalism asserts that, as a matter of definition, for a hypothesis, theory or explanation to qualify as "scientific" it must invoke only naturalistic or materialistic entities. Clearly, on this definition, the intelligent design hypothesis does not qualify as "scientific." Yet, even if one grants this definition, it does not follow that some non-scientific (as defined by methodological naturalism) or metaphysical hypothesis may not constitute a better, more causally, adequate explanation. Indeed, this essay has argued that, whatever its classification, the design hypothesis, does constitute a better explanation than its naturalistic rivals for the origin of specified biological information. Surely, simply classifying this argument as metaphysical does not refute it.

In any case, methodological naturalism now lacks justification as a normative definition of science. First, attempts to justify methodological naturalism by reference to metaphysically neutral (i.e., non-question begging) demarcation criteria have failed [106; 107; 114-117]. Second, asserting methodological naturalism as a normative principle for all of science has a negative affect on the practice of certain scientific (especially historical scientific) disciplines. In origin-of-life research, for example, methodological naturalism artificially restricts inquiry and prevents scientists from seeking some hypotheses that might provide the most likely, best, or causally adequate, explanations. For origin-of-life to be truth-seeking (or truth-tropic), the question that it must address is not 'which materialistic scenario seems most adequate?' but rather 'what actually caused life it to arise on earth?' Clearly, one of the possible answers to this latter question is 'Life was designed by an intelligent agent that existed before the advent of humans.' Yet if one accepts methodological naturalism as normative, scientists may never consider this possibly true causal hypothesis. Such an exclusionary logic diminishes the significance of any claim of theoretical superiority for any remaining hypothesis and raises the possibility that the best 'scientific' explanation (as defined by MN) may not be the best in fact.

As many historians and philosophers of science now recognize, scientific theory evaluation is an inherently comparative enterprise. Theories that gain acceptance in artificially constrained competitions can claim to be neither 'most probably true' nor 'most empirically adequate.' Instead, such theories can at best be considered the 'most probably true or adequate among an artificially limited set of options.' Openness to design would seem necessary, therefore, to any fully rational historical biology—to one that seeks the truth "no holds barred" [118, p. 535]. Further, given this more open definition of science—i.e., one where scientists use only metaphysically neutral criteria such as causal adequacy to evaluate competing explanations—the theory of intelligent design would now seem to provide the best, most causally adequate, explanation for the origin of the specified information necessary to the first living organism.

REFERENCES

1. Kamminga, H., Protoplasm and the Gene. In *Clay Minerals and The Origin of Life,* edited by A.G. Cairns-Smith and H. Hartman. Cambridge: Cambridge UP, 1986.
2. Oparin, A. *Genesis and Evolutionary Development of Life.* New York: Academic P, 1968.
3. Watson, J., and Crick, F. A Structure for Deoxyribose Nucleic Acid. *Nature* 171:737-38, 1953.
4. Watson, J., and Crick, F. Genetical Implications of the Structure of Deoxyribose Nucleic Acid. *Nature* 171:964-67, esp. 964, 1953.
5. Schneider, T.D. Information Content of Individual Genetic Sequences. *Journal of Theoretical Biology* 189:427-41, 1997.
6. Loewenstein, W.R., *The Touchstone of Life: Molecular Information, Cell Communication, and the Foundations of Life.* New York: Oxford UP, 1999.
7. Kuppers, B. *Information and the Origin of Life.* Cambridge, MA: MIT P, 1990.
8. Kay, L.E., Who Wrote the Book of Life?: Information and the Transformation of Molecular Biology. *Science in Context* 8, 4:601-634.

9. Kay, L.E. Cybernetics, Information, Life: The Emergence of Scriptural Representations of Heredity. *Configurations* 5:23-91.
10. Kay, L.E., *Who Wrote the Book of Life?*. Stanford, CA: Stanford UP, 2000, pp. xv-xix.
11. Haeckel, E. *The Wonders of Life,* translated by J. McCabe. London: Watts, 1905.
12. Huxley, T.H. On the Physical Basis of Life. *The Fortnightly Review* 5:129-45, 1869.
13. Oparin, A.I. *The Origin of Life,* translated by S. Morgulis. New York: Macmillan, 1938.
14. Meyer, S.C. *Of Clues and Causes: A Methodological Interpretation of Origin of Life Studies*. Ph.D. thesis, Dept. of History and Philosophy of Science, Cambridge University, 1991.
15. Asturby, W.T., and Street, A. X-Ray Studies of the Structure of Hair, Wool and Related Fibres. *Philosophical Transactions of the Royal Society of London* A 230:75-101, 1932.
16. Judson, H. *Eighth Day of Creation.* New York: Simon and Schuster, 1979.
17. Olby, R. *The Path to the Double Helix*. London: Macmillian, 1974.
18. Sanger, F., and Thompson, E.O.P. The Amino Acid Sequence in the Glycyl Chain of Insulin. (1 and 2). The Identification of Lower Peptides from Partial Hydrolysates. *Biochemical Journal* 53:353-66, 366-74, 1953.
19. Kendrew, J.C.; Bodo, G.; Dintzis, H.M.; Parrish, R.G.; and Wyckoff, H. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* 181:662-66, esp. 664, 1958.
20. Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; and Watson J.D. *Molecular Biology of the Cell.* New York: Garland, 1983.
21. Oswald, Avery T.; MacCleod, C.M.; and McCarty, M. Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III. *Journal of Experimental Medicine* 79:137-58, 1944.
22. Chargaff, E. *Essays on Nucleic Acids.* Amsterdam: Elsevier Publishing, 1963.
23. Matthei, J.H., and Nirenberg, M.W. Characteristics and Stabilization of DNAase-Sensitive Protein Synthesis in E. Coli Extracts. *Proceedings of the National Academy of Sciences, USA* 47:1580-88, 1961.
24. Nirenberg, M.W., and Matthei, J.H. The Dependence of Cell-Free Protein Synthesis in *E. coli* upon Naturally Occurring or Synthetic Polyribonucleotides. *Proceedings of the National Academy of Sciences, USA* 47:1588-1602, 1961.
25. Wolfe, S.L. *Molecular and Cellular Biology*. Belmont, CA: Wadsworth Publishing, 1993.
26. Shannon, C. A Mathematical Theory of Communication. *Bell System Technical Journal* 27:379-423, 623-56, 1948.
27. Dretske, F. *Knowledge and the Flow of Information.* Cambridge, MA: MIT P, 1981.
28. Kuppers, B. On the Prior Probability of the Existence of Life. In *The Probabilistic Revolution,* edited by Kruger, *et al.,* Cambridge, MA: MIT P, 1987. 355-69.
29. Shannon, C., and Weaver, W. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois P, 1949.
30. Yockey, H.P. *Information Theory and Molecular Biology.* Cambridge: Cambridge UP, 1992.
31. Sarkar, S. Biological Information: A Skeptical Look as Some Central Dogmas of Molecular Biology. In *The Philosophy and History of Molecular Biology: New Perspectives,* edited by S. Sarkar. Dorcrecht, Netherlands, Boston Studies in the Philosophy of Science, 1996. 196, 199-202.
32. Crick, F. On Protein Synthesis. *Symposium for the Society of Experimental Biology* 12:138-63, esp. 144, 153, 1958.
33. Dembski, W.A. *The Design Inference: Eliminating Chance Through Small Probabilities.* Cambridge: Cambridge UP, 1998.
34. Bowie, J., and Sauer, R. Identifying Determinants of Folding and Activity for a Protein of Unknown Sequences: Tolerance to Amino Acid Substitution. *Proceedings of the National Academy of Sciences, USA* 86:2152-56, 1989.
35. Bowie, J.; Reidhaar-Olson, J.; Lim, W.; and Sauer, R. Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitution. *Science* 247:1306-10, 1990.
36. Reidhaar-Olson, J., and Sauer, R. Functionally Acceptable Solutions in Two Alpha-Helical Regions of

Lambda Repressor. *Proteins, Structure, Function, and Genetics* 7:306-10, 1990.

37. Dawkins, R. *The Blind Watchmaker.* London: Longman, 1986.
38. Gates, B. *The Road Ahead.* Boulder, CO: Blue Penguin, 1996.
39. Orgel L.E. *The Origins of Life on Earth.* New York: John Wiley, 1973.
40. Schrödinger, E., *What is Life? & Mind and Matter*. Cambridge: Cambridge UP, 1967, p. 82.
41. Davies, P. *The Fifth Miracle.* New York: Simon & Schuster, 1998, p. 120.
42. Thaxton, C., and Bradley, W. Information and the Origin of Life. In *The Creation Hypothesis: Scientific Evidence for an Intelligent Designer*, edited by J.P. Moreland. Downers Grove, IL: InterVarsity P, 1994. 173-210.
43. Kauffman, S. *The Origins of Order.* Oxford: Oxford University Press, 1993.
44. Crick, F. *Life Itself.* New York: Simon and Schuster, 1981.
45. Monod, J. *Chance and Necessity*. New York: Vintage Books, 1971, pp. 97-98, 143.
46. Kenyon, D., and Steinman, G. *Biochemical Predestination.* New York: McGraw-Hill, 1969.
47. Quastler, H. *The Emergence of Biological Organization.* New Haven: Yale University P, 1964.
48. Wald, G. The Origin of Life. *Scientific American* 191 August:44-53, 1954.
49. Shapiro, R. *Origins: A Skeptic's Guide to the Creation of Life on Earth.* New York: Summit Books, 1986.
50. Crick, F. The Origin of the Genetic Code. *Journal of Molecular Biology* 38:367-79, 1968.
51. Kamminga, H., *Studies in the History of Ideas on the Origin of Life*. Ph.D. thesis, University of London, 1980, pp. 303-04.
52. De Duve, C. The Constraints of Chance. *Scientific American* January: 112, 1996.
53. Morowitz, H.J. *Energy Flow in Biology.* New York: Academic P, 1968.
54. Hoyle, F., and Wickramasinghe, C. *Evolution from Space.* London: J.M. Dent, 1981.
55. Cairns-Smith, A.G. *The Life Puzzle.* Edinburgh: Oliver and Boyd, 1971.
56. Prigogine, I.; Nicolis, G.; and Babloyantz, A. Thermodynamics of Evolution. *Physics Today* November:23, 1972.
57. Yockey, H.P. Self Organization, Origin of Life Scenarios and Information Theory. *Journal of Theoretical Biology* 91:13-31, 1981.
58. Axe D.D., Biological function places unexpectedly tight constraints on protein sequences. *Journal of Molecular Biology* 301(3):585-96.
59. Axe, D.; Foster, N.; and Fersrt, A. Active Barnase Variants with Completely Random Hydrophobic Cores. *Proceedings of the National Academy of Sciences, USA* 93:5590-94, 1996.
60. Behe, M. Experimental Support for Regarding Functional Classes of Proteins to be Highly Isolated from Each Other. In *Darwinism: Science or Philosophy?,* edited by J. Buell, and G. Hearns, 1994. 60-71.
61. Borel, E. *Probabilities and Life*, translated by M. Baudin. New York: Dover, 1962.
62. Pennisi, E. Seeking Life's Bare Genetic Necessities. *Science* 272:1098-99, 1996.
63. Mushegian, A., and Koonin, E. A Minimal Gene Set for Cellular Life Derived by Comparison of Complete Bacterial Genomes. *Proceedings of the National Academy of Sciences, USA* 93:10268-73, 1996.
64. Bult, C., *et. al*. Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus Jannaschi. Science* 273:1058-72, 1996.
65. Mora, P.T. Urge and Molecular Biology. *Nature* 199:212-19, 1963.
66. Hacking, I. *The Logic of Statistical Inference.* Cambridge: Cambridge UP, 1965.
67. De Duve, C. The Beginnings of Life on Earth. *American Scientist* 83:429-37, 1995.
68. De Duve, C. *Blueprint for a Cell: The Nature and Origin of Life.* Burlington, NC: Neil Patterson Publishers, 1991.
69. Joyce, G., and Orgel, L. Prospects for Understanding the Origin of the RNA World. In *RNA World,* edited by R.F. Gesteland, and J.F. Atkins. Colds Spring Harbor, NY: Colds Spring Harbor Laboratory

P, 1993. 1-25.

70. Mora, P.T. The Folly of Probability. In *The Origins of Prebiological Systems and of their Molecular Matrices*, edited by S.W. Fox. New York: Academic P, 1965. 311-12.
71. Bertalanffy, L.V. *Robots, Men and Minds.* New York: George Braziller, 1967.
72. Dobzhansky, T. Discussion of G. Schramm's Paper. In *The Origins of Prebiological Systems and of their Molecular Matrices*, edited by S.W. Fox. New York: Academic P, 1965. 310.
73. Pattee, H.H. The Problem of Biological Hierarchy. In *Toward a Theoretical Biology, vol. 3,* edited by C.H. Waddington. Edinburgh: Edinburgh University P, 1970. 123.
74. Von Neumann, J. *Theory of Self-reproducing Automata*. Completed and edited by A. Berks. Urbana, IL: University of Illinois P, 1966.
75. Wigner, E. The Probability of the Existence of a Self-reproducing Unit. In *The Logic of Personal Knowledge,* edited by E. Shils. London: Kegan and Paul, 1961. 231-35.
76. Landsberg, P.T. Does Quantum Mechanics Exclude Life? *Nature* 203:928-30, 1964.
77. Morowitz, H.J. The Minimum Size of the Cell. In *Principles of Biomolecular Organization,* edited by O'Connor and Churchill. London: Churchill, 1966. 446-59.
78. Nelson, P. Anatomy of a Still-Born Analogy. *Origins and Design* 17 (3):12, 1996.
79. Steinman, G., and Cole, M.N. Synthesis of Biologically Pertinent Peptides Under Possible Primordial Conditions. *Proceedings of the National Academy of Sciences, USA* 58:735-41, 1967.
80. Steinman, G. Sequence Generation in Prebiological Peptide Synthesis. *Arch. Biochem. Biophys.* 121:533-39, 1967.
81. Kok, R.A.; Taylor, J.A.; and Bradley, W.L. A Statistical Examination of Self-Ordering of Amino Acids in Proteins. *Origins of Life and Evolution of the Biosphere* 18:135-42, 1988.
82. Prigogine, I., and Nicolis, G. *Self-Organization in Non-Equilibrium Systems.* New York: John Wiley, 1977.
83. De Duve, C. *Vital Dust: Life as a Cosmic Imperative.* New York: Basic Books, 1995.
84. Thaxton, C.; Bradley, W.; and Olsen, R. *The Mystery of Life's Origin: Reassessing current theories*. Dallas: Lewis & Stanley, 1992.
85. Kenyon, D., and Mills, G. The RNA World: A Critique. *Origins and Design* 17 (1): 9-16, 1996.
86. Kenyon, D., and Davis, P.W. *Of Pandas and People: The Central Question of Biological Origins.* Dallas: Haughton, 1993.
87. Meyer, S.C. A Scopes Trial for the '90's. *The Wall Street Journal* 6 Dec. 1993, A14.
88. Stalnaker, R. *Inquiry.* Cambridge, MA: MIT P, 1984.
89. Polanyi, M. Life's Irreducible Structure. *Science* 160:1308-12, esp. 1309, 1968.
90. Yockey, H.P. A Calculation of the Probability of Spontaneous Biogenesis by Information Theory. *Journal of Theoretical Biology* 67:377-98, esp. 380, 1977.
91. Eigen, M. *Steps Toward Life.* Oxford: Oxford UP, 1992.
92. Walton, J.C. Organization and the Origin of Life. *Origins* 4:16-35, 1977.
93. Smith, J.M. Hypercycles and the Origin of Life. *Nature* 280:445-46, 1979.
94. Dyson, F. *Origins of Life.* Cambridge: Cambridge UP, 1985.
95. Shapiro, R. Prebiotic cytosine synthesis: A critical analysis and implications for the origin of life. *Proc. Natl. Acad. Sci. USA* 96:4396-4401, 1999.
96. Waldrop, M.M. Did Life Really Start Out in an RNA World? *Science* 246:1248-49, 1989.
97. Shapiro, R. Prebiotic Ribose Synthesis: A Critical Analysis.*Origins of Life and Evolution of the Biosphere* 18:71-85, 1988.
98. Joyce, G.F. RNA evolution and the origins of life. *Nature* 338:217-24, 1989.
99. Hager, A.J., Polland J.D. Jr, & Szostak, J.W. Ribozymes: aiming at RNA replication and protein synthesis. *Chemistry & Biology* 3:717-25, 1996.
100. Dose, K. The Origin of Life: More Questions Than Answers. *Interdisciplinary Science Reviews* 13:348-56, 1988.
101. Sober, E. *Reconstructing the Past.* Cambridge, MA: MIT P, 1988.

102. Scriven, M. Causes, Connections, and Conditions in History. In *Philosophical Analysis and History,* edited by W. Dray. New York: Harper & Row, 1966. 238-64.

103. *McNeil-Lehrer News Hour*. Transcript 19 May 1992.

104. McDonough, T.R. *The Search for Extraterrestrial Intelligence: Listening for Life in the Cosmos.* New York: Wiley, 1987.

105. Lipton, P. *Inference to the Best Explanation.* New York: Routledge, 1991.

106. Meyer, S.C. The Scientific Status of Intelligent Design: The Methodological Equivalence of Naturalistic and Non-Naturalistic Origins Theories. In *Science and Evidence for Design in the Universe,* The Proceedings of the Wethersfield Institute, vol. 9. San Francisco: Ignatius Press, 2000, pp.151-212.

107. Meyer, S.C. The Demarcation of Science and Religion. In *The History of Science and Religion in the Western Tradition: An Encyclopedia*, edited by Gary B. Ferngren. New York: Garland Publishing, 2000. 17-23.

108. Sober, E. *The Philosophy of Biology.* San Francisco: Westview P, 1993.

109. Rothman, M. *The Science Gap*. Buffalo, NY: Prometheus, 1992.

110. Popper, K. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge & Kegan Paul, 1962.

111. Harre, R., and Madden, E.H. *Causal Powers*. London: Basil Blackwell, 1975.

112. Barrow, J., and Tipler, F. *The Anthropic Cosmological Principle.* Oxford: Oxford UP, 1986.

113. Ruse, M. McClean v. Arkansas: Witness Testimony Sheet. In *But Is It Science?* edited by Michael Ruse. Amherst, NY: Prometheus Books, 1988. 103.

114. Laudan, L. The Demise of the Demarcation Problem. In *But Is It Science?* edited by Michael Ruse. Amherst, NY: Prometheus Books, 1988. 337-50.

115. Laudan, L. Science at the Bar—Causes for Concern. In *But Is It Science?* edited by Michael Ruse. Amherst, NY: Prometheus Books, 1988. 351-55.

116. Plantinga, A. Methodological Naturalism? *Origins and Design* 18, 1:18-26.

117. Plantinga, A. Methodological Naturalism? *Origins and Design* 18, 2:22-34.

118. Bridgman, P. *Reflections of a Physicist*. Second edition. New York: Philosophical Library, 1955.

---

[i] Darwin's only speculation on the origin of life is found in an unpublished 1871 letter to Joseph Hooker. In it he sketched the outlines of the chemical evolutionary idea, namely, that life could have first evolved from a series of chemical reactions. As he envisioned it, ". . .if (and oh! what a big if!) we could conceive in some warm little pond, with all sorts of ammonia and phosphoric salts, light, heat, electricity, etc., that a proteine compound was chemically formed ready to undergo still more complex changes. . ." Cambridge University Library, Manuscripts Room, Darwin Archives. Courtesy Peter Gautrey.

[ii] We now know, of course, that in addition to the process of gene expression, specific enzymes must often modify amino acid chains after translation in order to achieve the precise sequencing necessary to allow correct folding into a functional protein. The amino acid chains produced by gene expression may also undergo further modification in sequencing at the endoplasmic reticulum. Finally, even well-modified amino acid chains may require pre-existing protein "chaperons" to help them fold into a functional three-dimensional configuration. All these factors make it impossible to predict a protein's final sequencing from its corresponding gene sequence alone [31, pp. 199-202]. Nevertheless, this unpredictability in no way undermines the claim that DNA exhibits the property of "sequence specificity," or the isomorphic claim that it contains "specified information" as argued below in 2.5. Sarkar argues, for example, that the absence of such predictability renders the concept of information theoretically superfluous for molecular biology. Instead, this unpredictability shows that the sequence specificity of DNA base sequences constitutes a necessary, though not sufficient, condition of attaining proper protein folding—that is, DNA does contain specified information (see 2.5 below), but not enough to determine protein folding by itself. Instead, the presence of both post-translation processes of modification and pre-transcriptional genomic editing (through exonucleases, endonucleases, spliceosomes and other editing enzymes) only underscores the need for other pre-existing, information-rich biomolecules in order to process genomic information in

the cell. The presence of a complex and functionally integrated information processing system *does* suggest that the information on the DNA molecule is insufficient to produce proteins. It does not show that such information is *unnecessary* to produce proteins, nor does it invalidate the claim that DNA, therefore, stores and transmits specified genetic information.

[iii]See [30, pp. 246-58] for important refinements in the method of calculating the information carrying capacity of proteins and DNA.

[iv]Recall that the determination of the genetic code depended, for example, on observed correlations between changes in nucleotide base sequences and amino acid production in "cell free systems." [16, pp. 470-87].

[v]Indeed, of the two sequences, only the second meets an independent set of functional requirements. To convey meaning in English one must employ pre-existing (or independent) conventions of vocabulary (associations of symbol sequences with particular objects, concepts or ideas) and existing conventions of syntax and grammar (such as 'every sentence requires a subject and a verb.') When arrangements of symbols "match" or utilize these vocabulary and grammatical conventions (that is, functional requirements) meaningful communication can occur in English. The second sequence ("Time and tide wait for no man.") clearly exhibits such a match between itself and pre-existing requirements of vocabulary and grammar. The second sequence has employed these conventions to express a meaningful idea. It also, therefore, falls within the smaller (and conditionally independent) pattern delimiting the domain of all meaningful sentences in English and thus, again, exhibits a "specification."

[vi]Actually, Sauer counted sequences that folded into stable three-dimensional configurations as functional, though many sequences that fold are not functional. Thus, his results actually underestimate the probabilistic difficulty.

[vii]Dembski's universal probability bound actually reflects the "specificational" resources not the probabilistic resources in the universe. Dembski's calculation determines the number of specifications possible in finite time. It nevertheless has the effect of limiting the "probabilistic resources" available to explain the origin of any *specified* event of small probability. Since living systems are precisely specified systems of small probability the universal probability bound effectively limits the probabilistic resources available to explain the origin of specified biological information.

[viii]Cassette mutagenesis experiments have usually been performed on proteins of about 100 amino acids in length. Yet extrapolations from these results can generate reasonable estimates for the improbability of longer protein molecules. For example, Sauer's results on the proteins lambda repressor and arc repressor suggest that, on average, the probability at each site of finding an amino acid that will maintain functional sequencing (or, more accurately, that will produce folding) is less than 1 in 4 (1 in 4.4). Multiplying 1/4 by itself 150 times (for a protein 150 amino acids in length) yields a probability of roughly 1 chance in 10^{91}. For a protein of that length the probability of attaining both exclusive peptide bonding and homochirality is also about 1 chance in 10^{91}. Thus, the probability of achieving all the necessary conditions of function for a protein 150 amino acids in length exceeds 1 chance in 10^{180}.

[ix]Note that the "RNA World" scenario was not devised to explain the origin of the sequence specificity of biomacromolecules. Rather it was proposed as an explanation for the origin of the interdependence of nucleic acids and proteins in the cellular information processing system. In extant cells, building proteins requires instructions from DNA, but information on DNA cannot be processed without many specific proteins and proteins complexes. This poses a "chicken-or-egg" dilemma. The discovery that RNA (a nucleic acid) possesses limited catalytic properties (as modern proteins do) suggested a way to split the horns of this dilemma. By proposing an early earth environment in which RNA performed both the enzymatic functions of modern proteins and the information storage function of modern DNA, "RNA first" advocates sought to formulate a scenario making the functional interdependence of DNA and proteins unnecessary to the first living cell. In so doing, they sought to make the origin of life a more tractable problem from a chemical evolutionary point of view. In recent years, however, many problems have emerged with the RNA world (See section 3.6).

[x]This, in fact, happens where adenine and thymine do interact chemically in the complementary base pairing *across* the message bearing axis of the DNA molecule.

[xi]As noted in 2.4, the information carrying capacity of any symbol in a sequence is inversely related to the

probability of its occurrence. The informational capacity of a sequence as a whole is inversely proportional to the product of the individual probabilities of each member in the sequence. Since chemical affinities between constituents ("symbols") increase the probability of the occurrence of one given another (i.e., necessity increases probability), such affinities decrease the information carrying capacity of a system in proportion to the strength and relative frequency of such affinities within the system.

[xii] A possible exception to this generalization might occur in biological evolution. If the Darwinian mechanism of natural selection acting on random variation can account for the emergence of all complex life, then a mechanism does exist that can produce large amounts of information—assuming, of course, a large amount of *pre-existing* biological information in a self-replicating living system. Thus, even if one assumes that the selection/variation mechanism can produce all the information required for the macro-evolution of complex life from simpler life, that mechanism will not suffice to account for the origin of the information necessary to produce life from non-living chemicals. As we have seen, appeals to *pre-biotic* natural selection only beg the question of the origin of specified information. Thus, based on our experience we can affirm the following generalization: 'for all non-biological systems, large amounts (see endnote **xiii** below) of specified complexity or information only originate from mental agency, conscious activity, or intelligent design.' Strictly speaking, our *experience* may even affirm this generalization without the qualification, since the claim that natural selection can produce large amounts of novel genetic information depends upon (somewhat controversial) theoretical arguments and extrapolation from observations of small micro-evolutionary changes, rather than direct observation of the macro-evolutionary changes that would establish large gains in biological information. In any case, the more qualified empirical generalization (stated just above) is sufficient to support the argument presented here, since this essay seeks only to establish intelligent design as the best explanation for origin of the specified information necessary to the origin of the *first* life.

[xiii] Of course, the phrase "large amounts of specified information" again begs a quantitative question, namely, "how much specified information or complexity would the minimally complex cell have to have before it implied design?" Recall that Dembski has calculated a universal probability bound of $1/10^{150}$ corresponding to the probabilistic/specificational resources of the known universe. Recall, further, that probability is inversely related to information by a logarithmic function. Thus, the universal small probability bound of $1/10^{150}$ translates into roughly 500 bits of information. Thus, chance alone does not constitute a sufficient explanation for the *de novo* origin of any specified sequence or system containing more than 500 bits of (specified) information. Further, since systems characterized by complexity (a lack of redundant order) defy explanation by self-organizational laws, and since appeals to pre-biotic natural selection presuppose but do not explain the origin of the specified information necessary to a minimally complex self-replicating system, intelligent design best explains the origin of the more than 500 bits of specified information required to produce the first minimally complex living system. Thus, assuming a non-biological starting point (see endnote **xii** above), the *de novo* emergence of 500 or more bits of specified information will reliably indicate design.

[xiv] Again, this claim applies at least in cases where the competing causal entities or conditions are non-biological—or where the mechanism of natural selection can be safely eliminated as inadequate means of producing requisite specified information.

[xv] Less exotic (and more successful) design detection occurs routinely in both science and industry. Fraud detection, forensic science and cryptography all depend upon the application of probabilistic or information theoretic criteria of intelligent design [33, pp. 1-35].

[xvi] Many would admit that we *may* justifiably infer a past human intelligence operating (within the scope of human history) from an information-rich artifact or event, but only because we already know that human minds exists. But, they argue, since we do not know whether an intelligent agent(s) existed prior to humans, inferring the action of a designing agent antedating humans cannot be justified, even if we observe an information-rich effect. Note, however, that S.E.T.I. scientists do not already know whether an extra-terrestrial intelligence exists. Yet they assume that the presence of a large amount of specified information (such as the first 100 prime numbers in sequence) would definitively establish the existence of one. Indeed, S.E.T.I. seeks precisely to establish the existence of other intelligences in an unknown domain. Similarly, anthropologists have often revised their estimates for the beginning of human history or civilization because they discovered information-rich artifacts dating from times that antedate their previous estimates.

Most inferences to design establish the existence or activity of a mental agent operating in a time or place where the presence of such agency was previously unknown. Thus, inferring the activity of a designing intelligence from a time prior to the advent of humans on earth does not have a qualitatively different epistemological status than other design inferences that critics already accept as legitimate.