Complex Specification (CS)
— A New Proposal For Identifying Intelligence

Darel R. Finley

In *Three Issues With No Free Lunch* (iscid.org), I argued that William Dembski was exploiting a double definition of the term "information" on page 183 of his book *No Free Lunch* (Rowman & Littlefield 2002, ISBN 0-7425-1297-5). (See *Three Issues* for details.) In this paper I intend to refine what I meant by that particular charge, tweaking the definition of "information", separating it from the term "specification," and in so doing, show that Complex Specification (CS) is actually the true hallmark of intelligent communication.

This paper will draw on the example of a lone five-letter English word. Some sequences of five letters form words in the English language:

```
NIGHT
SIGNS
GLASS
```

Whereas other five-letter sequences do not:

```
VIOWO
RSSPN
BLAUY
```

What amount of information is communicated by a five-letter English word? Is it specified? *How* specified is it?

The number of different sequences of five letters from the English alphabet is 26^5, or about 2^{23}. But of those sequences, only about 10,000 of them are English words — that's about 2^{13}. So, the amount of *information* that can be conveyed by an English word is 13 bits. The remaining 10 bits are *specification.* Specification is a form of error-checking or overhead, that helps to ensure reliable communication.

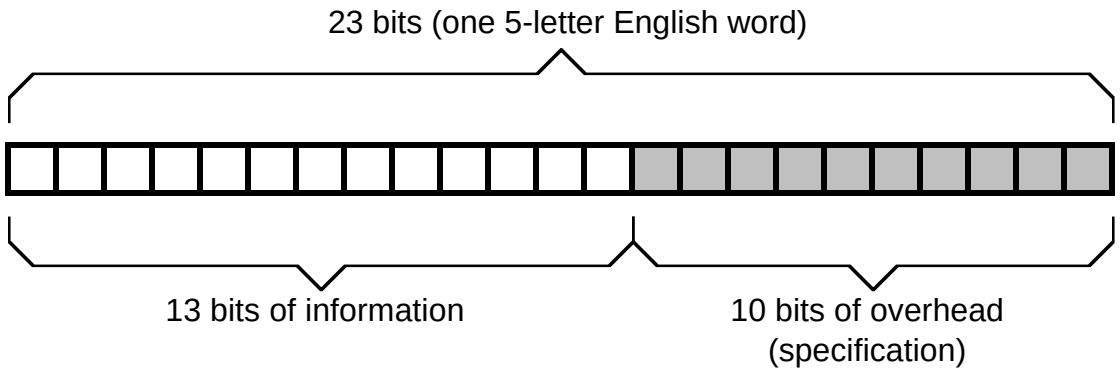

Figure 1. Data Content of a Five-Letter English Word

Attaching a significant amount of overhead to a piece of data, as illustrated in Figure 1, is a common practice in the field of communications technology. For example, on a music CD, although the majority of the data on the disc is audio information, a substantial percentage of the data is not — it is there specifically for error detection and correction. Likewise, the fact that only about 1 in 1,000 (1 in 2^{10}) five-letter sequences form English words can be helpful in detecting errors, and even correcting them. For example, most people would realize that the character sequence "Come to my house tomorrow nbght" was supposed to be "Come to my house tomorrow night." Even the sequence "nbght" by itself might be recognized as "night," because in this case, there is only one English word that can be achieved by a single-letter substitution.

The ten bits of specification data in Figure 1 above do not contain information, because their state is entirely determined by the state of the 13 information bits. What might the 10 specification bits contain? They can contain anything, so long as it is consistent for any given 13 information bits. For example, the 10 specification bits could contain:

```
0000000000
```

regardless of the content of the 13 information bits. But that would be a very poor choice of specification, because ten zeros might be generated by accident, or by natural processes. A better choice would be a random-looking string. So, we could set the 10 specification bits to:

```
0010111010
```

again regardless of the content of the 13 information bits. This is better, because it helps us to recognize a legitimate 23-bit packet — we simply notice that the last 10 bits contain 0010111010, and have a fairly high confidence that we are looking at a 23-bit English word packet (given that this particular specification is some sort of accepted standard).

But we can still do better. If the 10 specification bits can be made somehow dependent on the content of the 13 information bits, then the specification bits can serve as a form of error-checking for the information bits. Suppose, for example, that our 13 information bits contain:

```
1101100100110
```

and we arrange them into two groups of 10 and 3 bits, and perform an exclusive-or function:

```
1101100100
       110
----------
1101100010
```

Then we use the 10 resulting bits as our specification, so the full 23-bit packet is:

```
11011001001101101100010
```

It is then a simple matter to test any 23-bit packet using the above exclusive-or technique to see if meets the specification we are using. This method is better than using a fixed 10-bit value, because it has a good chance of revealing when the 13 information bits have been corrupted (altered in transit).

The error-checking method used on a music CD is quite a bit more complex than the one shown above, and much more robust in detecting errors and facilitating their correction. But the principle is the same. A percentage of the CD's data capacity is sacrificed for the sake of specification; the rest is used for actual music data.

In the case of five-letter English words, the 13 data bits and the 10 specification bits are not merely linked causally, they are mathematically mixed together into one 23-bit mass, in a complex and arbitrary way. Separating the 13 information bits out of the 23 data bits is not as simple as just lopping off the last 10 bits. In fact, to do it

requires a large amount of arbitrary knowledge of what is an English word and what isn't.

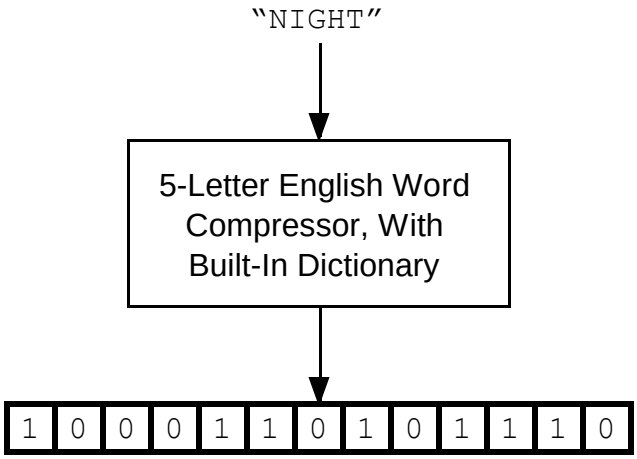

Figure 2. Dictionary Algorithm To Separate Out Information Bits

This is because the English language requires memorization of a very arbitrary list five-letter words. The hypothetical compressor depicted in Figure 2 has a full, alphabetized list of all five-letter English words, with a 13-bit index value assigned to each. When passed a word, it looks the word up in its list, then returns the 13-bit index value.

Extreme Cases

To see what role information bits and specification bits serve when it comes to identifying intelligence, we must consider extreme cases. First, consider the case of all information and no specification. Suppose we communicated in a language called "AnyFive", which is similar to English (same alphabet, punctuation, grammar, etc.), except that in AnyFive, *every* five-letter sequence is a meaningful word; each with a unique meaning not exactly duplicated by any other five-letter word. In that case, the data diagram of a five-letter word would look like this:

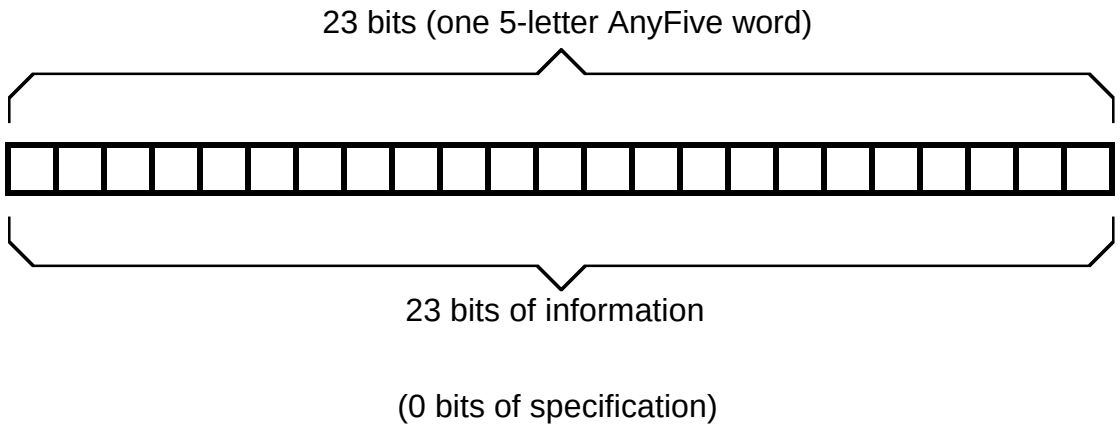

Figure 3. Data Content of a Five-Letter AnyFive Word

Since pure information is indistinguishable from randomly-generated data, it would be difficult or impossible to know whether these 23 bits of data were generated by an intelligence, or by natural processes. If we received a radio signal from deep space, containing 23 (or more) bits of apparently random data, we could not reliably attribute it to extraterrestrial intelligence. Or, if you noticed five Scrabble tiles laid out in a loose row on the table, there would be no way to know if their order communicates something or is just random, because in AnyFive, the five tiles will always spell a valid word.

Now let's consider the other extreme. In the hypothetical language "OneFive," there is only *one* five-letter word: "NIGHT." A data diagram of that word looks like this:

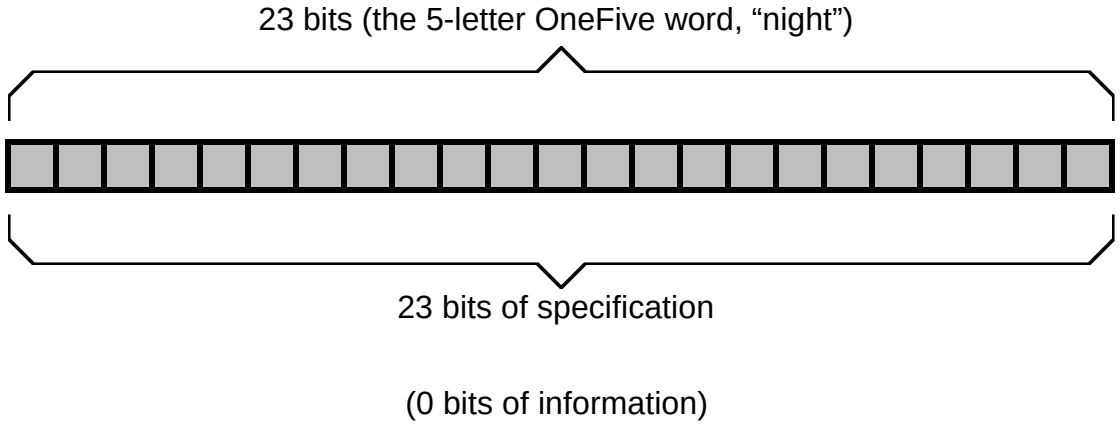

Figure 4. Data Content of the OneFive Word "Night"

Assuming that for some reason our communications are restricted to a single five-letter word, what can be communicated by the OneFive word "night?" No information can be transmitted, because there are no information bits to send. The only message that this word can transmit is "I am here," or "I exist," or "Communication has occurred." In effect, this message is similar to a computer networking "handshake" — it carries no information, but serves the purpose of reliably indicating to entity A that entity B exists.

Assuming you lived in a OneFive-speaking world, if you walked by the Scrabble table and saw five tiles out of the box in a loose row, and they spelled "NIGHT," you would have a fairly high confidence that it was not a random arrangement. In fact, the probability of five tiles randomly spelling a OneFive word is 1 in 2^{23}, or 1 in 8.4 million. So even though the word allows no communication of *information* at all (within the five-tile scenario), it still is a reliable indicator of intelligent *action*. In fact, it is much *more* reliable than an English five-letter word, since the English word has only 10 bits of specification, not 23.

Likewise, the sequence of prime numbers detected from the alien transmission in the movie *Contact* (as discussed in *No Free Lunch*), carried no information at all. If the aliens wanted to tell us anything about fusion reactors, or warp drives, or their own biology, they could not do so with a sequence of prime numbers, because it is entirely specified by mathematics, and hence carries only one message: "I am here, I know about prime numbers, and I can send radio signals." (Note: This ties in loosely with Walter Remine's "Message Theory" as described in his book *The Biotic Message.*)

The irony here is that any information-carrying standard which consists of pure information is completely unreliable, whereas a standard which consists of pure specification can be very reliable, but can carry no informational message. It seems like a catch-22, but the good news is that we can use a system that has a strong quantity of *both* information and specification. And in *Contact,* that's exactly what happens — the prime-number *(specification)* barrage was followed by a large quantity of *information* bits that told the humans how to build a warp drive.

Biology

How does this apply to biology? Again, let's consider the extreme cases. First, suppose that we lived in a world where *all* sequences of DNA made functional organisms (albeit each adapted for a specific subset of the Earth's environment).

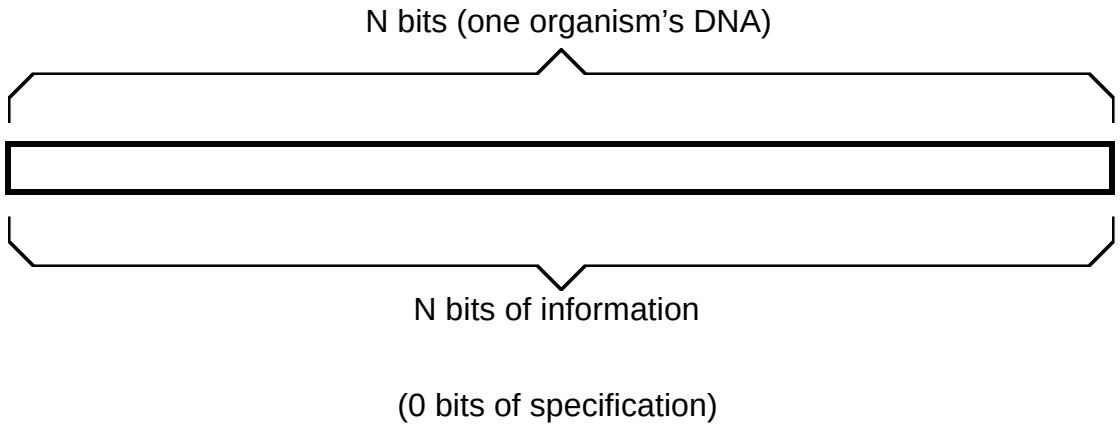

Figure 5. DNA Data Content If All DNA Sequences Made Functional Earth Organisms

In that case, even though the DNA might contain far more than 500 bits of data (Dembski's Universal Complexity Bound), there would be no reliable indicator of intelligent action, because any DNA sequence would do fine. Even if a designer *had* written the DNA, there would be nothing but the designer's pure whim represented there, and pure whim is not distinguishable from naturally-generated randomness.

This would also be true of a world in which, say, one out of every eight DNA sequences made an Earth-adapted organism.

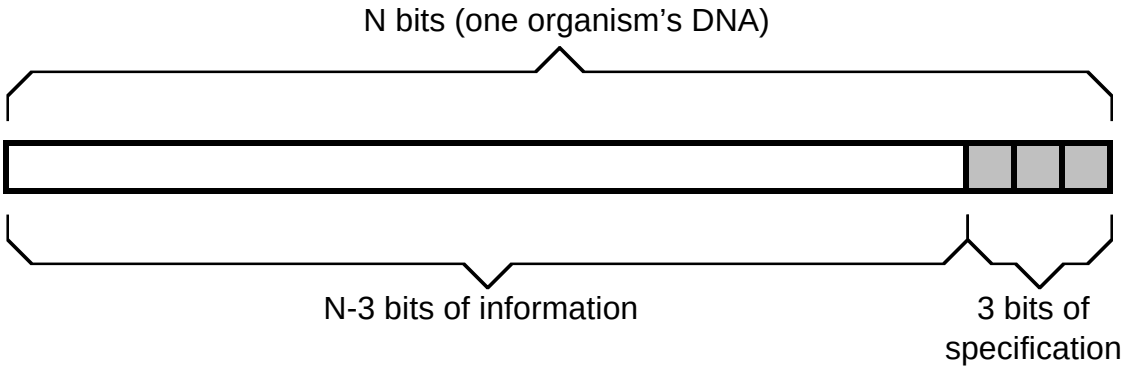

Figure 6. DNA Data Content If 1/8 of All DNA Sequences Made Functional Earth Organisms

Even though the total data content vastly exceeds 500 bits, the specification data is only 3 bits, and so there is no reliable indicator of intelligent action, since a random, evolutionary process might easily be able to jump 3-bit hurdles, using only naturally occurring chance mutations.

Now suppose that the world allowed only one DNA sequence — that of humans. Any change to that sequence at all (intelligent or otherwise) caused death. Humans were the only organism alive.

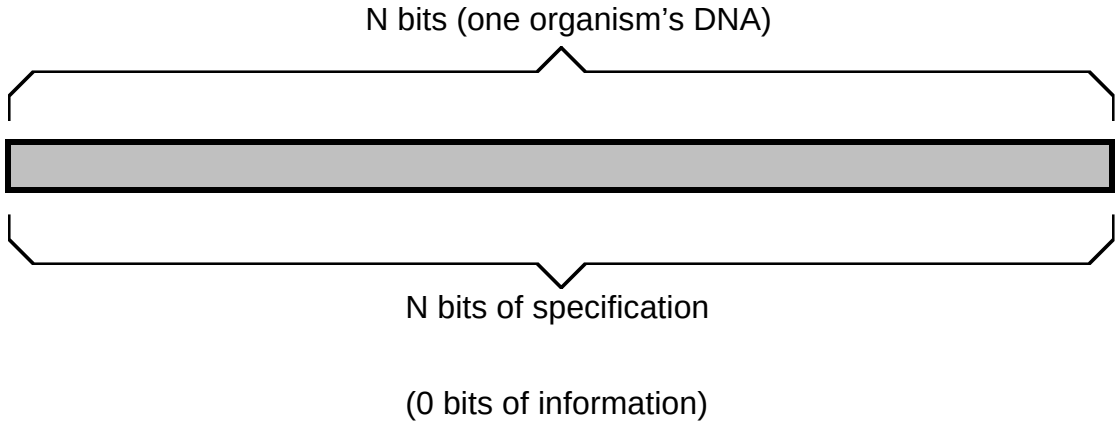

Figure 7. DNA Data Content If Only One DNA Sequence Made A Functional Earth Organism

In this case there is no contingent choice of one organism over another — the world allows only one. But nevertheless, our confidence that humans were designed

would be very high — higher, in fact, than they are under the actual set of circumstances, because there would be no plausible way for a natural process — with or without the help of natural selection — to produce the human genome all at once.

And one more example — the case of a world in which only eight DNA sequences form functional organisms, so there are at most eight species:

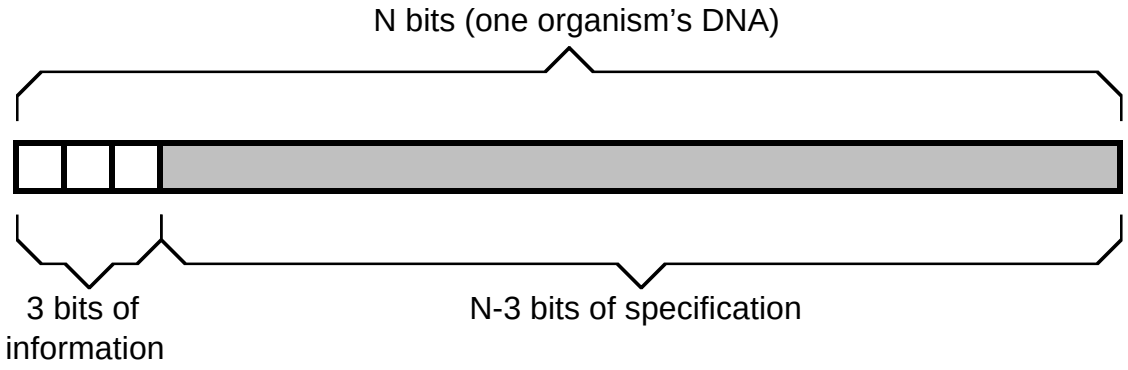

Figure 8. DNA Data Content If Eight DNA Sequences Made Functional Earth Organisms

In such case, the designer can choose from only eight different organisms, and thus communicates only three bits of information through the creative process — but still our confidence that those organisms were designed would be very high indeed.

Complex Specification

In conclusion, I propose that 500 or more bits of *specification,* not information, be considered the universally reliable indicator of design. The issue of real-vs-apparent-specified-complexity can now be answered as follows: A piece of data with high specification, but little or no information, can still be reliably attributed to design. An intelligent entity can make contingent choices, but those choices are expressed in the *information* content of the data, not in the specification content, and those choices are *not* useful for determining whether intelligence is responsible for the data in question. Intelligence is a measure of the ability to create data that matches a complex specification, whereas choice is simply the ability to prefer A

over B when either would have worked, and does not require intelligence — nor implies it.

On page 11 of *Three Issues*, I proposed the following scenario:

> The laws of physics are pre-loaded with all the CSI that is currently found in the DNA of living organisms, but do not provide an evolutionary path to realizing that data in actual organisms.

I must now qualify that proposal by saying that the laws of physics are pre-loaded with the Complex Specification of living organisms, but that the designer made contingent choices to realize only a subset of possible types. As human technology advances, we may choose to realize many more biological designs that our designer did not.

Summary

My thesis of Complex Specification is summarized as follows:

* I communicate with you by *whimsically choosing* a particular *information* content, wrapping it in a *specification* wrapper (which increases the data size somewhat), then sending it to you.

* You recognize that intelligent communication has occurred by the amount and type of *specification*, and you receive my whimsical message through the *information* content which was wrapped in the specification wrapper.

* Specification alone is only a handshake — it conveys no informational message — but it reliably reveals the presence of intelligent action.

* Information alone is not reliably recognizable as intelligent communication — it can represent only random whim — but it allows a desired message to be carried.

* Information and specification *combined* form reliable communication between intelligent entities.