

Evolution and the Origin of Biological Information

BY DENNIS VENEMA

“If your heart is right, then every creature is a mirror of life to you and a book of holy learning, for there is no creature - no matter how tiny or how lowly - that does not reveal God’s goodness.”

Thomas a Kempis - Of the Imitation of Christ (c.1420)

One prominent antievolutionary argument put forward by the Intelligent Design Movement (IDM) is that significant amounts of biological information cannot be created through evolutionary mechanisms – processes such as random mutation and natural selection. ID proponent and structural biologist Doug Axe frames the argument this way (his comments begin at approx. 15:19 in the video):

“Basically every gene, every new protein fold... there is nothing of significance that we can show [that] can be had in that gradualistic way. It’s all a mirage. None of it happens that way.”

The importance of this line of argumentation for the IDM can be seen clearly in Stephen Meyer’s book *Signature in the Cell* (published in 2009). In this book, Meyer claims that an intelligent agent is responsible for the information we observe in DNA because, in his words, natural mechanisms “will not suffice” to explain it:

Since the case for intelligent design as the best explanation for the origin of biological information necessary to build novel forms of life depends, in part, upon the claim that functional (information-rich) genes and proteins cannot be explained by random mutation and natural selection, this design hypothesis implies that selection and mutation will not suffice to produce genetic information ... (p. 495)

It’s hard to overstate the importance of this argument for Meyer in *Signature*, and for the IDM as a whole. In the conclusion to a pivotal chapter entitled “The Best Explanation” Meyer presents the following summary of his case:

Since the intelligent-design hypothesis meets both the causal-adequacy and causal-existence criteria of a best explanation, and since no other competing explanation meets these conditions as well –or at all–it follows that the design hypothesis provides the best, most causally adequate explanation of the origin of the information necessary to produce the first life on earth. Indeed, our uniform experience affirms that specified information ... always arises from an intelligent source, from a mind, and not a strictly material process. So the discovery of the specified digital information in the DNA molecule provides strong grounds for inferring that intelligence played a role in the origin of DNA. Indeed, whenever we find specified information and we know the causal story of how that information arose, we always find that it arose from an intelligent source. It follows that the best, most causally adequate explanation for the origin of the specified, digitally encoded information in DNA is that it too had an intelligent source. (p. 347)

Put more simply, Meyer claims that if we see specified information, we infer design, since we know of no mechanism that can produce specified information through an unintelligent, natural process. As a logical argument, Meyer's position only works if (and this is a big if) – his premises are correct.

The issue is that Meyer's case is open to refutation by counterexample, and even one counterexample would suffice. If any natural mechanism can be shown to produce "functional, information-rich genes and proteins", then intelligent design is no longer the best explanation for the origin of information we observe in DNA, by Meyer's own stated criteria. His entire (500+ page) argument would simply unravel.

The obvious problem for Meyer's case is that biologists are well aware of a natural mechanism that does add functional, specified information to DNA sequences (and in some cases, creates new genes *de novo*): natural selection acting on genetic variation produced through random mutation. Not only are biologists aware of some examples of natural selection adding functional information to DNA, this effect has been observed time and again, and in some cases it has been documented in exquisite detail. When I reviewed *Signature* for the American Scientific Affiliation journal *Perspectives on Science and Christian Faith* (PSCF) what struck me, repeatedly, was that Meyer made no mention of the evidence for natural selection as a mechanism to increase biological information. I fully expected him to dispute the evidence, certainly – but the surprise for me was that he simply denied it to be sufficient without addressing any evidence. The closest Meyer comes in addressing natural selection in *Signature* is in a section discussing evolutionary algorithms used to simulate evolution. As I said in my review:

Meyer's denial of random mutation and natural selection as an information generator notwithstanding, in a discussion about evolutionary computer simulations, Meyer makes the following claim:

If computer simulations demonstrate anything, they subtly demonstrate the need for an intelligent agent to elect some options and exclude others- that is, to create information.

Employing this argument, Meyer claims that any mechanism that prefers one variant over another creates information. As such, the ample experimental evidence for natural selection as a mechanism to favor certain variants over others certainly qualifies as such a generator. Meyer, however, makes no mention of the evidence for natural selection in the book.(pp. 278-279)

In the PSCF review I went on to point out a few examples of known instances in biology where random mutation and natural selection have indeed led to substantial increases in biological information, but the limitations of space in that format precluded me from exploring those examples in more detail, or from presenting that information at a level readily accessible to non-specialists. In this paper I will attempt to remedy that shortcoming by exploring several examples in depth. The question of how new specified information arises in DNA, far from being an "enigma", is one of great interest to biologists. While the IDM avoids this evidence to present a flawed argument for design, responding to this flawed argument provides an excellent opportunity to discuss some particularly elegant experiments in this area.

Of course, it should be noted that describing how specified information can arise through natural means does not in any way imply God's absence from the process. After all, natural processes are equally a

manifestation of God’s activity as what one would call supernatural events. So-called “natural” laws are what Christians understand to be a description of the ongoing, regular and repeatable activity of God. As such, the dichotomy presented in ID writings of “naturalism” versus theism is a false one: is not God the Author of nature, after all?

E. Coli vs. Intelligent Design

So far, we have explored the claim made by Stephen Meyer, a leader in the Intelligent Design Movement (IDM), that “specified, complex information” cannot arise through natural means. This is crucial to Meyer’s argument, since any natural mechanism that can be shown to produce information would render his argument that information only arises from intelligent sources null and void.

A second member of the IDM who frequently makes this argument is Douglas Axe, a researcher at the Biologic Institute. Axe’s specialty is in protein structure / function relationships, and he has published a few papers in this area in the mainstream scientific literature. Axe’s work also forms the basis for Meyer’s arguments in this area in his book *Signature in the Cell*. I met Axe a few years ago when I gave a presentation at Baylor, and again last year in Austin for the Vibrant Dance conference (for whatever reason, it seems we only cross paths in Texas). Axe was present in the audience for a discussion session I shared with Richard Sternberg, and we had a significant amount of back-and-forth. As such, I am familiar with his line of argument, and it matches what we saw previously in *Signature* (as one might expect, since Meyer bases his work on Axe).

Perhaps the best summary of Axe’s argument is his quote I highlighted previously (begins approx. 15:19):

“Basically every gene, every new protein fold... there is nothing of significance that we can show [that] can be had in that gradualistic way. It’s all a mirage. None of it happens that way.”

One of the interesting features of the IDM is that though it has not yet brought forward strong hypotheses with which to test ID, it frequently makes testable predictions about natural processes. Specifically, Axe’s hypothesis is that mutation and natural selection will be unable to produce anything significant in a gradual way.

Has natural selection been Axed?

The ideal way to test this hypothesis, of course, would be to follow a population of organisms over thousands of generations and track any genetic changes that occur to see if they result in any new functions. Even better would be the ability to determine the precise molecular mutations that brought about these changes, and compare the offspring side-by-side with their ancestors. An experiment with this level of detail might sound too good to be true, but one of exactly this sort has been going on since the late 1980s, studying the bacterium, *E. Coli*. It’s called the Long Term Evolution Experiment (LTEE), and it’s the brainchild of Dr Richard Lenski at Michigan State University.

The LTEE started in 1988 with twelve populations of *E. Coli* all derived from one ancestral cell. The design of the experiment is straightforward: each day, each of the twelve cultures grow in 10ml of liquid medium with glucose as the limiting resource. In this medium, the bacteria compete to replicate for about

seven generations and then stop dividing once the food runs out. After 24 hours, 1/10th of a ml of each culture is transferred to 9.9 ml of fresh food, and the cycle repeats itself. Every so often, the remaining 9.9 ml of leftover bacterial culture is frozen down to preserve a sample of the population at that point in time – with the proper treatment, bacteria can survive for decades in suspended animation. Early in the experiment this was done every 100 generations, and later this was shifted to every 500 generations. A significant feature of the LTEE is that these frozen ancestors can be brought to life again for comparison with their evolved descendants: in essence, the freezers in the Lenski lab are a nearly perfect “living fossil record” of the experiment.

It is important to note several things about the LTEE. First, there is no artificial selection taking place. The environment for the bacteria is kept constant: the same food, the same temperature and the same dilution routine are maintained each day. Second, the bacteria in the experiment are asexual: this means that genetic recombination, a hugely important source of genetic variation in sexual organisms, is absent. New genetic combinations in the LTEE must arise solely by mutation. Third, the bacterial populations that started the experiment are unlike any natural population, since they are all identical clones of each other. (In other words, genetic variation in the original 12 cultures was essentially zero). While natural populations have genetic variation to draw on, these twelve cultures started from scratch.

Since its inception, the twelve cultures have gone their separate ways for over 50,000 generations. Early on, the cultures quickly adapted to their new environment, with variants in each population arising and outcompeting others. In order to confirm that the new variants indeed represented increases in function (and thus, an increase in “information”) the evolved variants were tested head-to-head against their revived ancestors. Numerous papers from the Lenski group have documented these changes in great detail. What was remarkable about the early work from the Lenski group was that tracking the 12 cultures showed that evolution in the different populations was both contingent and convergent: similar, but not identical, mutations appeared in many of the lines, and the different populations had similar, but not identical, increases in fitness relative to the ancestral populations. In the details, evolution was contingent, but overall, the pattern was convergent. As Lenski puts it:

To my surprise, evolution was pretty repeatable. All 12 populations improved quickly early on, then more slowly as the generations ticked by. Despite substantial fitness gains compared to the common ancestor, the performance of the evolved lines relative to each other hardly diverged. As we looked for other changes—and the “we” grew as outstanding students and collaborators put their brains and hands to work on this experiment—the generations flew by. We observed changes in the size and shape of the bacterial cells, in their food preferences, and in their genes. Although the lineages certainly diverged in many details, I was struck by the parallel trajectories of their evolution, with similar changes in so many phenotypic traits and even gene sequences that we examined.

In other words, there were many possible genetic states of higher fitness available to the original strain, and random mutation and natural selection had explored several paths, all leading to a higher amount of “specified information” – information that specifies increased reproduction and survival in the original environment. All this was by demonstrably natural mechanisms, with a complete history of the relevant mutations, the relative advantages they conferred, and the dynamics of how those variants spread through

a population. The LTEE is at once a very simple experiment, and an incredibly detailed window into the inner workings of evolution.

And so the work continued, day in and day out, for years – until one day, a completely new biological function showed up in one of the cultures.

One of the defining features of *E. Coli* is that it is unable to use citrate as a food source. The food used to culture the strains, however, has a large amount of citrate in it – a potential food source that remained beyond the reach of the evolving strains. For tens of thousands of generations, no variants arose that could make use of this potential resource – even though every possible single DNA letter mutation (and every possible double mutation combination) had been “tested” at some point along the way. There seemed no way to for the populations to generate “specified information” to use citrate as a food source – they couldn’t “get there from here.” Then one day, the fateful change occurred in one of the 12 populations. Lenski puts it this way:

Although glucose is the only sugar in their environment, another source of energy, a compound called citrate, was also there all along as part of an old microbiological recipe. One of the defining features of *E. coli* as a species is that it can’t grow on citrate because it’s unable to transport citrate into the cell. For 15 years, billions of mutations were tested in every population, but none produced a cell that could exploit this opening. It was as though the bacteria ate dinner and went straight to bed, without realizing a dessert was there waiting for them.

But in 2003, a mutant tasted the forbidden fruit. And it was good, very good.

Details, details

Tracking down the nature of this dramatic change led to some interesting findings. The ability to use citrate as a food source did not arise in a single step, but rather as a series of steps, some of which are separated by thousands of generations:

1. The first step is a mutation that arose at around generation 20,000. This mutation on its own does not allow the bacteria to use citrate, but without this mutation in place, later generations cannot evolve the ability to use citrate. Lenski and colleagues were careful to determine that this mutation is not simply a mutation that increases the background mutation rate. In other words, a portion of what later becomes “specified information for using citrate” arises thousands of generations before citrate is ever used.
2. The earliest mutants that can use citrate as a food source do so very, very poorly – once they use up the available glucose, they take a long time to switch over to using citrate. These “early adopters” are a tiny fraction of the overall population. The “specified information for using citrate” at this stage is pretty poor.
3. Once the (poor) ability to use citrate shows up, other mutations arise that greatly improve this new ability. Soon, bacteria that use citrate dominate the population. The “specified information for using citrate” has now been honed by further mutation and natural selection.
4. Despite the “takeover”, a fraction of the population unable to use citrate persists as a minority. These cells eke out a living by being “glucose specialists” – they are better at using up glucose rapidly and then going into stasis before the slightly slower citrate-eaters catch up. So, new “specified information to get

the glucose quickly before those pesky citrate-eaters do” allows these bacteria to survive. As such, the two lineages in this population have partitioned the available resources and now occupy two different ecological niches in the same environment. As such, they are well on their way to becoming different bacterial species.

Don't tell the bacteria

The significance of these experiments for the Intelligent Design Movement is clear. Complex, specified information can indeed arise through natural mechanisms; it does not need to arise all at once, but rather accrue over thousands of generations; independent mutations that do not confer a specific advantage can later combine with other mutations to produce new functions; new functions can be quite inefficient when they arise and then be honed through further mutations and selection; and the entire process can occur without ever reducing the fitness of a specific lineage within a population. Moreover, these findings have been demonstrated with a full historical record of the genetic changes involved for the entire population they occurred in, as well as full knowledge of their fitness at every step along the way.

In other words, what the IDM claims is impossible, these “tiny and lowly” organisms have simply been doing – and it only took 15 years in a single lab in Michigan. Imagine what could happen over 3,500,000,000 years over millions of square miles of the earth’s surface.

CSI on Steroids

Next, we will look at an example of new information and function arising during vertebrate evolution: the elegant work of the Thornton lab on steroid hormones and their protein receptors.

To briefly recap, so far we have noted that:

5. CSI does not need to arise all at once, but can arise piecemeal through independent mutation events.
6. Separate mutations that later combine to form CSI do not need to confer a specific advantage on their own. In other words, mutations that are “neutral” with respect to the survival of the organism can later be co-opted into CSI that does have a distinct survival advantage.
7. Neutral mutations may open up new future paths. In the LTEE, the brand-new ability of one bacterial population to use citrate as a food source required that a neutral mutation appear several thousand generations before it combined with other mutations to provide the CSI for using citrate.
8. When CSI arises, it can be pretty poor at the beginning. Nascent CSI, though poor, provides a survival advantage because it is the “best game in town” at that time. Further mutation in, and natural selection on, the offspring of the original CSI-holder quickly refine the nascent information into ever-more “specified” CSI.

And, as we noted previously, understanding how natural processes create information is in no way a threat to God’s ordaining and sustaining of creation. Rather, it is an opportunity to explore some of the mechanisms by which He does so.

With these important principles in mind, we are ready to examine a second fascinating case of a novel function arising through mutation and selection: the evolutionary history of steroid hormones and their protein receptors in vertebrates.

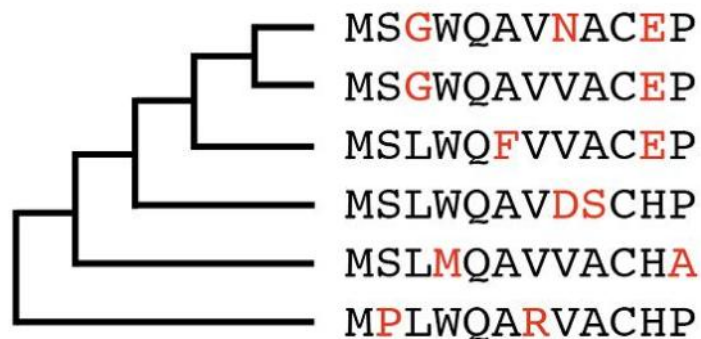
Experimental Evolution as Textual Criticism

The elegant work by the Lenski group has one very distinct advantage that other researchers surely envy: when new structures and functions arise in the experiment, a trip to the freezer is all that is needed to resurrect and examine the relevant ancestors. For researchers who study other organisms less amenable to laboratory experimentation, or for evolutionary transitions that happened deep in the past, other methods are needed. One approach to this type of problem is to “resurrect” ancient proteins in the lab in order to study their properties.

Bringing an ancient gene back to life starts with determining what its DNA sequence was, (and thereby determining the sequence of amino acids that made its functional protein product). While researchers don’t have direct access to ancient DNA, we have the next best thing: many modern examples of genes copied from the ancestral one.

For those who are familiar with textual criticism, the principles are very similar. Textual criticism is the process of recovering the words of an ancient manuscript by comparing several very similar, but still imperfect, copies. In general, as more copies agree on a certain wording, the more likely it is that the original had that wording. Also, the more widespread and older a certain wording is, the more likely it is original. Groups of manuscripts that have similar copying errors or other variations can be grouped together as more closely related, and so on. Given enough manuscripts, it is possible to recreate a copy of an ancient text with a very high degree of accuracy. As Christians, we benefit from this type of analysis daily when we read the Bible: though no two Greek manuscripts of the New Testament are exactly alike, scholars have used these methods to recover the original text with a very high degree of confidence.

And so too, for ancient gene sequences. Consider a hypothetical amino acid sequence in six modern organisms:

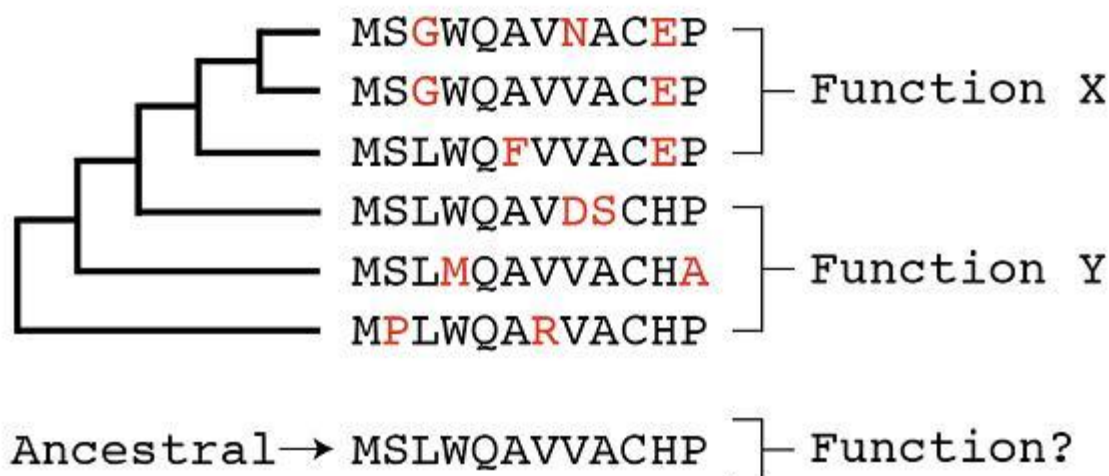


Ancestral → MSLWQAVVACHP

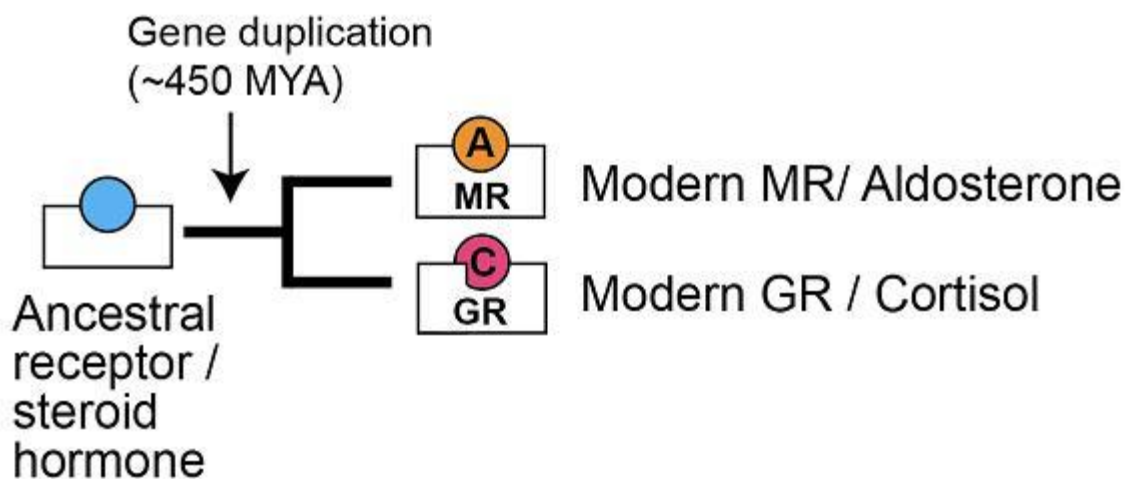
Though none of the modern sequences are identical, it is easy to see that there is a “consensus” at each of the 12 amino acid positions. This consensus sequence is very likely to be the ancestral sequence: explaining the pattern in any other way requires many more changes, with many changes occurring in parallel after species separate.

Once the researchers determine the correct ancestral amino acid sequence, it’s a relatively small matter to engineer a DNA sequence that encodes it and give it to cells to make into protein. This protein can then be tested to see how it functions compared to the modern sequence.

What makes this type of analysis even more interesting is that sometimes related genes acquire new functions. In cases like these, bringing the ancestral gene back to life in the lab allows researchers to not only test its properties, but to test hypotheses about what the specific amino acid changes were that changed the protein's function over time:



The laboratory of Joseph Thornton at the University of Oregon has used this method (with great success) to determine how certain hormone / protein receptor complexes arose during vertebrate evolution. Hormones are small molecules that act as signals by binding to a protein target, called a receptor. The receptor / hormone pair then goes on to effect a change in the target cell by regulating other genes.



In vertebrates, two hormone – receptor pairs were of interest to the Thornton group: the mineralocorticoid receptor (MR), which binds a steroid hormone called aldosterone, and the glucocorticoid receptor (GR), which binds a second steroid hormone called cortisol (see diagram above). Cortisol can also activate MRs, but an enzyme that breaks down cortisol is present in tissues where MR is used so cortisol cannot accumulate. Aldosterone, on the other hand, cannot activate GR – it is specific to its binding partner MR. Even though these two hormone / receptor pairs regulate different processes in modern organisms, the

two receptors are the result of an ancient gene duplication that occurred early in vertebrate evolution, around 450 MYA (million years ago). As time has gone by, the derivatives of the original gene have picked up distinct binding partners and physiological roles. Thornton and colleagues wanted to tease out the details of these important changes.

They started out by determining the ancestral sequence of the original receptor gene, prior to the duplication, and recreating it in the lab. When they tested this lab-designed protein, they found that it, like modern MRs, (but not GR's) could bind either cortisol or aldosterone indicating that the ancestral protein must have been able to bind both. This result suggested that somewhere along the line the GR lost its ability to bind aldosterone and became specific to cortisol. This is interesting, because at the time the ancestral receptor was present, aldosterone didn't exist. Aldosterone is a relative newcomer on the scene: it is present only in four-limbed vertebrates (tetrapods), which arose around 390 MYA. So, the ancestral receptor present prior to 450 MYA already had the ability to bind a hormone that wouldn't evolve for tens of millions of years. Of course, the ancestral receptor "didn't mind" – it had its own binding partner - a steroid hormone closely related to cortisol and aldosterone. It wasn't sitting around doing nothing in the meantime.

This finding strongly suggested that the reason aldosterone binds only to MRs is because modern GRs, in contrast to the ancestral protein, have lost the ability to bind it. By comparing the amino acid differences between MRs and GRs, the Thornton group was able to test different combinations to see what the key changes likely were. They also did the (difficult) work of determining the precise new shape of the receptor for each of the changes that had an effect. All in all, it is an impressive body of scientific work.

Through these techniques, the Thornton group demonstrated that the loss of aldosterone sensitivity in GR occurred in a series of mutational steps that progressively remodeled the portion of the GR that binds the hormone molecule:

1. First, a mutation occurred that altered one of the amino acids near the hormone binding site. This change had no effect on its own (it was a neutral mutation).
2. Second, a change in an amino acid outside the binding pocket bent one side of the binding site into a new shape. Now the amino acid from the first neutral mutation in step #1 was thrust up against the hormone binding site. This amino acid can interact appropriately with cortisol, but not very well with aldosterone. The receptor was now strongly biased towards cortisol.
3. Later, several more mutations accrue that "tune" the receptor to its new specificity. Some of the mutations are neutral at first (like step #1) and then combine with later mutations to refine the receptor into its modern cortisol-specific role.

As the GR and MR lineages were becoming functionally distinct, other changes in other genes accumulated that refined their ability to regulate different processes (such as the enzyme that breaks down cortisol where MR is present, or the target genes that the two hormones regulate). While many of those details remain to be worked out, this work is an elegant demonstration of how a new function arose: gene duplication; sequence divergence with a neutral mutation that opened up a new possible trajectory; a second mutation that altered function in one of the gene copies; further mutations that refined this nascent difference; and the final result of new structures and functions that act as key regulators of important physiological processes in tetrapods, including humans.

In other words, new CSI.

Over and against these lines of evidence, however, the Intelligent Design Movement claims that such novelty is inaccessible to random mutation and natural selection. Rather, they claim that functional protein shapes are incredibly rare and therefore so isolated from each other that random mutation and natural selection cannot bridge the vast gulfs between them. Though Thornton's work (and the work of Lenski that we examined previously) refutes this claim with detailed, concrete examples, new comparative genomics tools have addressed this issue with greater power and breadth than ever before. Next in this paper, we'll explore the question: are these examples rare, isolated cases, or indicative of a wider pattern?

Lost in (Sequence) Space

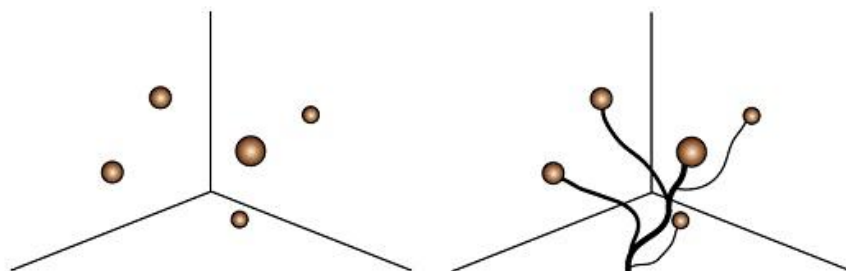
Previously, we explored two concrete examples of how new structures and functions arose through mutation and natural selection: the ability of E. Coli to utilize citrate that appeared during a controlled laboratory experiment, and the duplication and divergence of a steroid hormone receptor gene that acquired a new hormone binding partner and went on to regulate new processes distinct from its predecessor.

Both of these examples were notable for their intricate level of detail that carefully teased out the intermediates on the path to new functions. However, we also noted that:

Over and against these lines of evidence, however, the Intelligent Design Movement claims that such novelty is inaccessible to random mutation and natural selection. Rather, they claim that functional protein shapes are incredibly rare and therefore so isolated from each other that random mutation and natural selection cannot bridge the vast gulfs between them.

The issue here is that functional proteins seem to be a very small subset of possible proteins. Proteins are chains of repeated structures (amino acids) that are typically one hundred or more repeats in length. There are 20 amino acids found in proteins, so at every position in a protein chain, there are 20 different possible choices. So, for a protein with only two amino acids (not even a realistic scenario) there are 20² possible combinations. For a protein with 100 amino acids, there are 20¹⁰⁰ combinations – a vast “sequence space” of possible states, of which only a relative few will be functional.

As we have seen in Parts 2 and 3, proteins “explore” their sequence space through random mutation. Mutation may produce protein forms that reduce or remove function, changes that are neutral with respect to function, or changes that improve function (or add new functions). Over time, evolution predicts that proteins will “branch” through sequence space – with each modern form connected to a previous form of which it is a modified descendant. The Intelligent Design Movement (IDM), as we have noted, predicts a different pattern: isolated, separately designed (created), functional proteins that lack prior transitional forms.

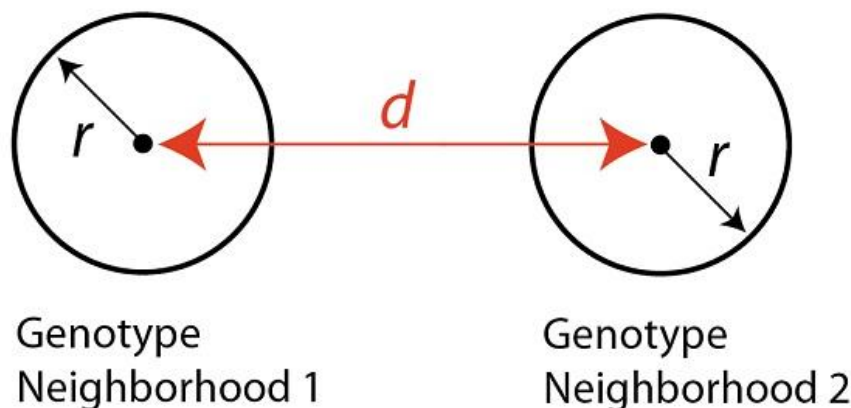


In other words, the IDM views protein sequence space to be like the diagram on the left. The brown spheres represent functional protein shapes (each of which allows for some small variation within the sphere). These are separated by large gaps of nonfunctional sequences. In contrast, an evolutionary model predicts that modern-day functional sequences (brown spheres) are connected in sequence space by functional intermediates across time (black lines).

The two examples we have already examined (citrate metabolism and novel hormone / receptor pairs; see sidebar for links) are strong support for the evolutionary model: in both cases new functions and structures were connected to prior forms (that had different functions) through a series of functional intermediates. The question remains, however: are all proteins so connected? Are these examples rare exceptions? Certainly if evolution has produced the diversity in protein form and function that we observe today this pattern should be common.

Welcome to the Neighborhood

That was the question that recently led two researchers to examine a large number of protein enzymes with known functions: 28,862 different proteins from a wide array of organisms, to be exact. Specifically, the researchers examined “genotype neighborhoods”: proteins that have similar amino acid sequences and group together in sequence space (such as those represented by the spheres in the diagram above). A two-dimensional cross-section of two such spheres can be represented as follows (redrawn from Figure 2 in Ferrada and Wagner, 2010):



Where each sphere has a radius (r), and the two are separated in sequence space by a distance (d). The radius and the distance are percent differences in amino acids. For example, we may consider all proteins that differ by at most 2% of their amino acids within the two neighborhoods ($r=1$ for both). The distance between the two neighborhoods (d) is also a percent difference in amino acids (for example, d could be 10%).

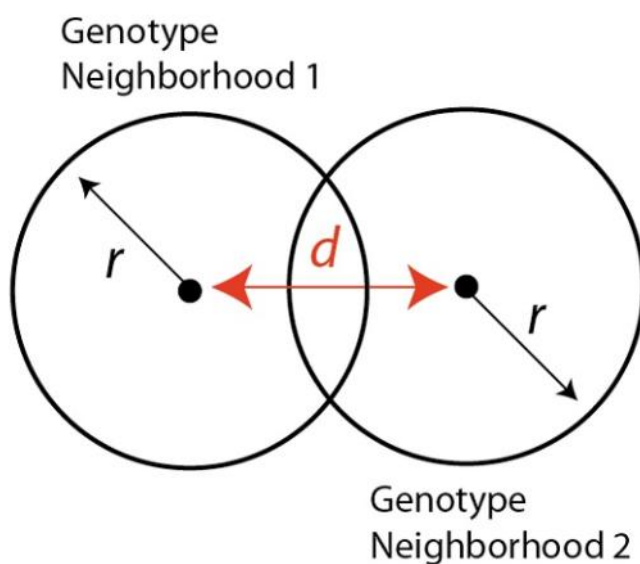
Since the data set used by the researchers was for enzymes with known functions, pairs of genotype neighborhoods were assessed to determine if they contained the same enzymatic functions, or distinct functions. For example, if neighborhood 1 contains enzyme functions A, B and C, and neighborhood 2 contains only enzyme functions A and B, then enzyme function C is unique to neighborhood 1. The fraction of unique functions for pairs of genotypic neighborhoods can thus be analyzed as functions of r and d .

In other words, how different do two genotype neighborhoods have to be before new functions are encountered in protein sequence space? Are existing protein families situated in protein space as isolated

islands of (independently designed) function in a sea of nonfunctionality, as the IDM predicts? Or can new functions be reached as enzymes explore sequence space through random mutation and natural selection?

Not surprisingly, the researchers found that as the percent amino acid differences (d) increased between two genotype neighborhoods, the fraction of unique functions increased. What was interesting (in terms of assessing the claims of the IDM) was that unique functions can be readily observed even for low values of d . For example, genotype neighborhoods with a 20% difference in amino acids ($d = 20$) had unique functions over 45% of the time when r was held constant at a 5% difference. Smaller differences, such as $d = 10$, did not eliminate unique functions (nearly 20% had unique functions; see figures 3A and 3B in Ferrada and Wagner for results for the data set as a whole).

A second interesting result was that even when genotype neighborhoods overlap (i.e. d is less than the sum of the two radii), they still may have unique functions:



This simultaneously underscores two observations: that highly similar sequences may have different functions (as is well known from other studies), as well as the contingent nature of proteins exploring sequence space (even closely related proteins cannot reach the same potential functions via a short search, depending on their position in their genotype neighborhood). This result is also consistent with what we have seen previously in parts 2 and 3: neutral mutations that move a sequence within its genotype neighborhood can bring it into reach of new potential functional states. Such neutral mutations were key in opening up future possibilities both for the evolution of citrate metabolism in *E. Coli* as well as in for steroid hormone receptors in vertebrates.

Does maintaining a specific protein structure prevent exploration?

Having obtained this result, the researchers went on to add a constraint to the analysis: they restricted their data set to protein sequences known to fold into a specific structure (the data for the TIM barrel domain can be seen in Figures 4A and 4B; compare with 3A and 3B). They chose a very common protein fold (called a TIM barrel) that many protein sequences can fold into (4,132 sequences in the data set), and that performs many different enzymatic functions (53 distinct chemical reactions currently known).

The amino acid sequences that form a TIM barrel can be 100% different (i.e. $d = 100$) or very similar ($d \sim 0$). As before, the researchers examined how functions are distributed in sequence space for pairs of genotype neighborhoods, but now restricted to this structure alone. Significantly, their results were the same as before. Genotypic neighborhoods close to each other still showed different functions, and overlapping neighborhoods contained unique functions. To be certain that this was not an effect specific to the TIM domain, the researchers repeated the analysis for 36 additional structures, all of which gave similar results.

Put another way, constraining a protein to a particular three-dimensional structure (i.e. protein fold) does not seem to hinder its ability to traverse sequence space and acquire new functions in the process.

Taken together, this paper demonstrates some key findings for how protein sequences, structures and functions are distributed in protein sequence space:

9. The distribution of protein sequences, structures and functions we observe is strongly consistent with the hypothesis that proteins traverse sequence space and acquire new functions over time through random mutation and selection.
10. Functional sequences in protein sequence space are distributed such that a significant subset of protein families are close to areas with new functions. In some cases, genotype neighborhoods can overlap where one neighborhood contains functions that the other does not.
11. Not all areas of a genotype neighborhood are equivalent: neutral mutations within a genotype neighborhood can move a sequence to regions where new functions can be reached, or into areas where those same functions are not accessible.
12. Constraint on protein structure is not a constraint on acquiring new functions. When the analysis was restricted to a common structure, the same results were obtained (consistent for 37 different structures).

Moreover, this work is based on the largest sample size examined to date (over 28,000 proteins), and thus is much more likely to apply to protein sequence space as a whole than studies (such as those performed by members of the IDM) that attempt to extrapolate from studies of one protein (or a handful of related proteins) to protein sequence space in general. Despite the claims of the IDM, proteins do not appear to be “lost” in sequence space.

Paralogs, Synteny and WGD

Let’s review that which I wrote at the beginning of this paper:

One prominent antievolutionary argument put forward by the Intelligent Design Movement (IDM) is that significant amounts of biological information cannot be created through evolutionary mechanisms – processes such as random mutation and natural selection. ID proponent and structural biologist Doug Axe frames the argument this way (his comments begin at approx. 15:19 in the video):

“Basically every gene, every new protein fold... there is nothing of significance that we can show [that] can be had in that gradualistic way. It’s all a mirage. None of it happens that way.”

The importance of this line of argumentation for the IDM can be seen clearly in Stephen Meyer’s book, *Signature in the Cell* (published in 2009). In this book, Meyer claims that an intelligent agent is responsible for the information we observe in DNA because, in his words, natural mechanisms “will not suffice” to explain it:

Since the case for intelligent design as the best explanation for the origin of biological information necessary to build novel forms of life depends, in part, upon the claim that functional (information-rich) genes and proteins cannot be explained by random mutation and natural selection, this design hypothesis implies that selection and mutation will not suffice to produce genetic information ... (p. 495)

It's hard to overstate the importance of this argument for Meyer in *Signature*, and for the IDM as a whole. In the conclusion to a pivotal chapter entitled "The Best Explanation" Meyer presents the following summary of his case:

Since the intelligent-design hypothesis meets both the causal-adequacy and causal-existence criteria of a best explanation, and since no other competing explanation meets these conditions as well –or at all—it follows that the design hypothesis provides the best, most causally adequate explanation of the origin of the information necessary to produce the first life on earth. Indeed, our uniform experience affirms that specified information ... always arises from an intelligent source, from a mind, and not a strictly material process. So the discovery of the specified digital information in the DNA molecule provides strong grounds for inferring that intelligence played a role in the origin of DNA. Indeed, whenever we find specified information and we know the causal story of how that information arose, we always find that it arose from an intelligent source. It follows that the best, most causally adequate explanation for the origin of the specified, digitally encoded information in DNA is that it too had an intelligent source. (p. 347)

Put more simply, Meyer claims that if we see specified information, we infer design, since we know of no mechanism that can produce specified information through an unintelligent, natural process. As a logical argument, Meyer's position only works if (and this is a big if) – his premises are correct.

So that's how the paper began. With each step, I have shown why Meyer is wrong. Natural mechanisms can explain the origin of new information and they can do so on a grand scale. Step-by-step I have shown how new information has arisen in the history of life and I have shown how the mechanism occurs through natural means. I have also emphasized that this does not in any manner exclude God from the processes. The natural laws that bring about this increase in information are a reflection of God's ongoing natural activity. Without that activity all would disintegrate into nothingness. What science has not demonstrated, (in contrast to what Meyer has proposed) is that God's supernatural activity is necessary also. Furthermore there is nothing in Scripture which mandates that we should expect God's supernatural activity to be necessary to bring about an increase in information. This removes the matter from the realm of a religious question and causes us to evaluate Meyer's proposal on the basis of the quality of the science itself.

Now I will demonstrate the evidence that two huge increases (doublings) in information content likely occurred in the evolution of vertebrates (organisms with a backbone) about 450 million years ago. Indeed it is likely that these doublings served as a key prelude to the huge array of vertebrate organisms (including ourselves) that subsequently arose in the history of creation.

Biologists have known for a long time that genes exist in families where each gene within a family is clearly related to other members of the same family. Individual genes for building hemoglobin, for example are members of a single family. It is known that each member of the family arose through a series of

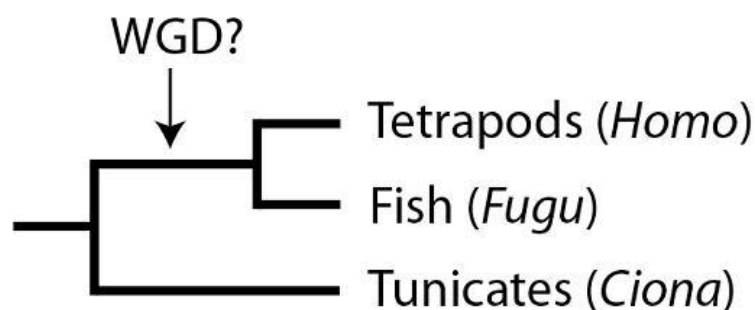
duplication events all of which trace back to a single ancestral gene. After each duplication event different mutations accumulate in opposing members of the duplicate pair and lead to divergence from identity. As gene family members change, so also their function becomes altered. Genes like this, (those which are clearly related to each other historically through a duplication event) are called paralogs. Every time a gene duplicates it provides an opportunity for new information (complex specified information, to use Meyer's term). This has happened routinely in the history of life and it occurs through natural mechanisms that are well understood.

Make mine a double - double

In addition to the widespread occurrence of duplication of individual genes and subsequent divergence, there has long been speculation that early in vertebrate evolution there was a time when the entire genome (the complete collection of genes) was doubled in a vertebrate ancestor. This is called a whole-genome duplication (WGD) event. Moreover, some evidence suggested that perhaps there was not just one, but rather two sequential WGD events in the early vertebrate lineage. WGD events provide a wealth of raw material for evolutionary innovation, since genes dedicated to one function now have a copy that is not constrained by natural selection to perform that role any longer (since the other copy can do that function). In many cases, the new gene copies are lost before they neofunctionalize (i.e. become different enough to gain a new function) – but in some instances, the copies are maintained and acquire new functions to become paralogs.

While there was some evidence to suggest that vertebrates had undergone one, or perhaps two, rounds of WGD early in their evolution, it was not, until recently, possible to distinguish the signature of a true vertebrate WGD event from smaller gene duplication events that accrued over time. Two key pieces of information were needed: first, the complete genome sequence of an organism closely related to vertebrates, and secondly, knowledge of the precise arrangement of the genes thought to be involved in the WGD events.

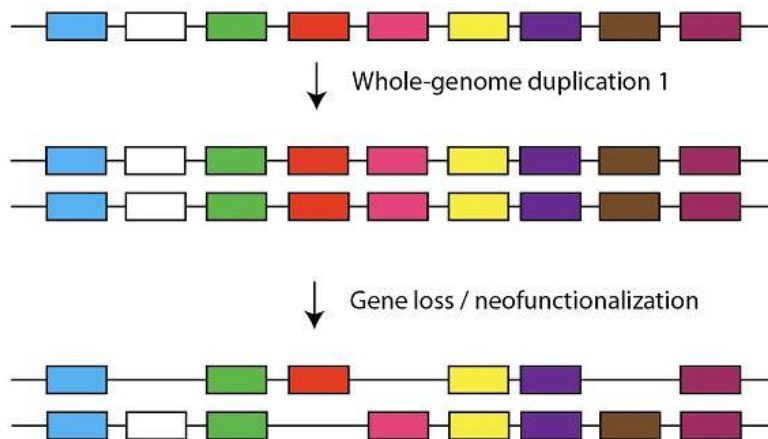
The first issue, that of a complete genome sequence of a close relative to vertebrates, was needed in order to determine which genes would have been present at the time of the proposed WGD event(s). Genes present in a modern close relative of vertebrates (the tunicate *Ciona intestinalis*) and in modern vertebrates (such as humans and fish) indicate which subset of modern vertebrate genes were present in the single ancestral species from which both tunicates like *Ciona* and vertebrates like us were derived. . With the *Ciona* genome in hand, and a vast array of vertebrate genomes, researchers were able to determine this set of genes, and distinguish them from genes that cropped up later in vertebrate evolution, or independently in either group.



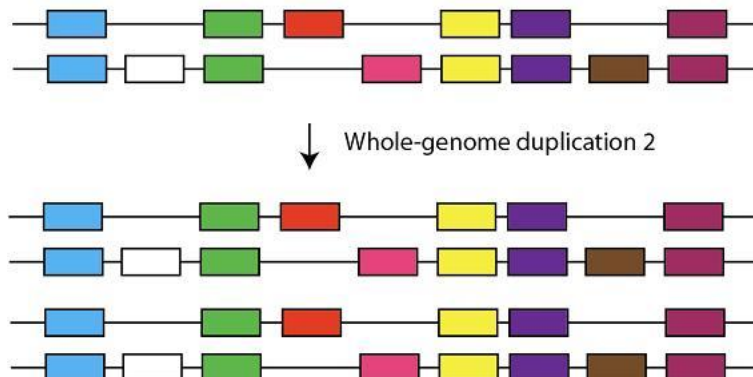
The next question was a relatively simple one: now that the subset of genes present at the proposed time for WGD was identified, how could the hypothesis of WGD be tested? The answer is in what is known as synteny: the physical arrangement of genes in a specific order on chromosomes. Darrel Falk and I have previously discussed synteny when examining human – chimpanzee common ancestry, and readers who have not read that post might find it helpful. In this case, when testing the hypothesis of WGD events, the researchers knew that this process would produce a specific pattern of gene arrangements in modern genomes – a pattern that accounts for both duplication and loss of genes. Additionally, it would resolve the 1x WGD versus 2x WGD debate, since these two possibilities would produce different genomic patterns.

Signature in the synteny, redux

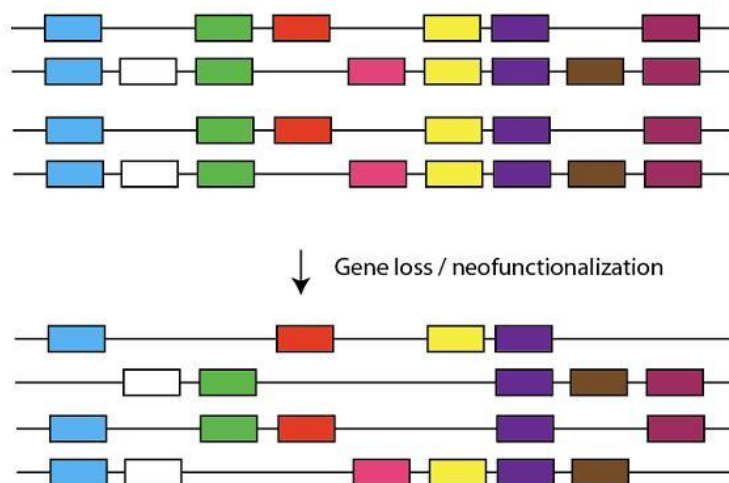
Consider a hypothetical genome that has only nine genes (represented by the colored boxes, redrawn from Figure 1 in Dehal and Boore, 2005) in a specific order on one chromosome pair. For simplicity's sake, we'll only show one copy of the chromosome. This chromosome is first duplicated in the WGD event, followed by large-scale loss of many redundant genes. Some genes, however, persist as paralogs (gene copies that pick up new functions and thus are not eliminated).



The point is that this event would retain the spatial pattern of the original gene set prior to duplication, even as some copies are lost. The new genome of this organism would now have two chromosome pairs, each copied from an original single chromosome, with some genes lost, and some genes now present as paralogs. Now imagine if a second WGD event occurred:



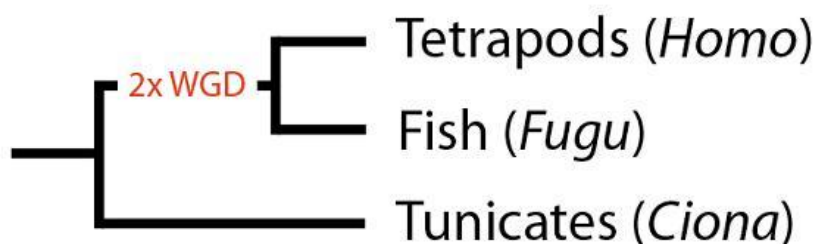
This event would produce an organism now with four chromosomes. As before, rapid gene loss (of redundant copies) and some neofunctionalization (to produce paralogs) would be expected:



The final pattern would have groupings within the same genome (groups of synteny) where paralogs are arranged in the same spatial pattern four times over. The four groupings of synteny would not be expected to be identical, since due to gene loss there might be as few as one copy remaining, or two, three or four paralogs persisting. While the hypothesis of WGD could not be resolved by looking at single paralog families, the overall pattern of four-fold synteny would be distinctive and unmistakable when investigators examined whole genomes. Tunicates are the control since even though they are closely related to vertebrates, the two hypothetical duplication events arose in vertebrate ancestral species which are not part of the tunicate “family tree.”

Testing the WGD hypothesis, and implications for ID

The research group first identified 3,753 genes that are present as single copies in *Ciona* but present as multiple paralogs in fish and humans. These are the genes that may still show a WGD or 2x WGD synteny signature in modern vertebrates (although chromosomal rearrangements may erase the synteny signature over time). Significantly, a large percentage of these modern paralogs are still present in four-fold synteny groups that span about 25% of the human genome (and thus have persisted as blocks of synteny for approximately 450 million years). This evidence is a strong indication that the modern vertebrate genome went through two rounds of WGD early in its evolution, and that these events provided substantial “raw material” for the acquisition of new information through gene divergence and neofunctionalization:



In other words, gene duplication and divergence to produce new CSI appears to be commonplace in evolution, including the evolution of our own species. Far from being rare exceptions, multiple lines of genomics evidence point to new structures, functions and information being produced through natural means. If the Intelligent Design Movement wishes to contest that natural mechanisms cannot produce new information, they need to address this widespread and compelling pattern.

A Long Look in the Mirror

This paper has been exploring the question of how new structures and functions arise through evolutionary mechanisms. This topic is one that is of considerable interest for Christians, since the Intelligent Design Movement claims that the generation of such features (what it terms Complex Specified Information, or “CSI”) is not possible for natural processes to produce in any significant measure. As such, it holds up examples of CSI in nature as evidence for a supernatural designer. Unfortunately, this approach has the effect that new scientific evidence that explains how CSI arises naturally diminishes the perceived evidence for God.

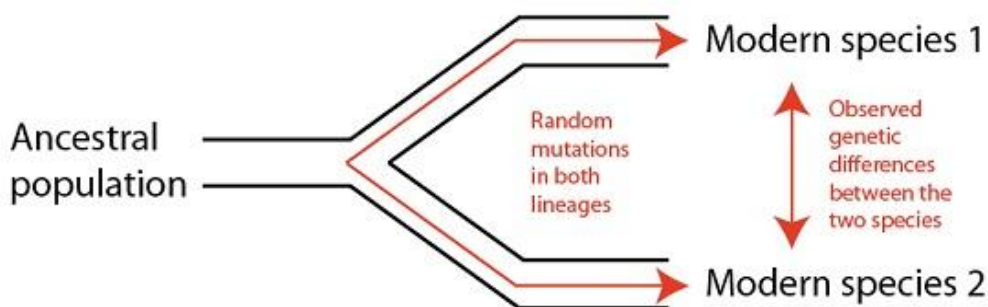
As we were careful to point out at the very beginning of our discussion, understanding how natural processes create information is in no way a threat to God’s ordaining and sustaining of creation. If it were so, the obvious conclusion would be that “natural mechanisms” and “God’s actions” are effectively a zero-sum game where every scientific discovery diminishes God’s activity. Indeed, the ID argument strongly tends in this direction. This is certainly not a historic Christian view of science, and one we would do well to steer the church away from.

With these theological considerations in mind, we have explored several examples of new CSI arising through evolutionary processes. In summary, we have seen that:

13. CSI does not need to arise all at once, but can arise piecemeal through independent mutation events.
14. Separate mutations that later combine to form CSI do not need to confer a specific advantage on their own. In other words, mutations that are “neutral” with respect to the survival of the organism can later be co-opted into CSI that does have a distinct survival advantage.
15. Neutral mutations may open up new future paths. For example, the brand-new ability of one bacterial population to use citrate as a food source required that a neutral mutation appear several thousand generations before it combined with other mutations to provide the CSI for using citrate. A second example we observed is how neutral mutations opened up new future possibilities during the evolution of hormone/hormone receptor complexes in vertebrates.
16. When CSI arises, it can be pretty poor at the beginning. Nascent CSI, though poor, provides a survival advantage because it is the “best game in town” at that time. Further mutation in, and natural selection on, the offspring of the original CSI-holder quickly refine the nascent information into ever-more “specified” CSI.
17. The detailed examples of new CSI arising through changes to existing proteins appear to apply generally to many, many protein families across multiple organisms. There is nothing about protein structure that prevents proteins from acquiring new functions through evolutionary means.
18. Comparative genomics evidence, especially evidence from synteny, strongly supports the hypothesis that large swaths of modern vertebrate genomes are the result of ancient whole-genome duplications, where some of the duplicated genes go on to acquire new functions through mutation and selection.

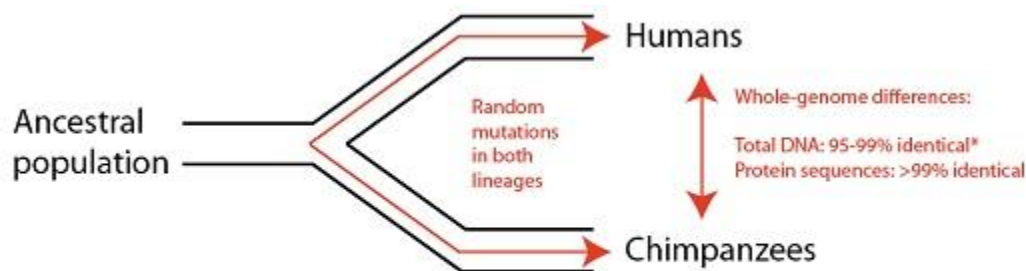
Comparative genomics and new CSI: details, details

One last way to assess the ability of natural processes to generate CSI that we will explore is based on comparing the genomes of two closely-related species that nonetheless have significant biological differences. The most detailed approach, of course, is to examine and compare the entire genomes of the species in question. Such a comparison shows us the total genetic differences that have arisen between the species since they parted ways:



It needs to be emphasized that only a subset of the observed differences will be meaningful. Put another way, many of the mutations that have occurred in the two lineages are neutral, having no discernable effect on the organisms in question. Indeed, the subset of truly meaningful differences is likely to be relatively small. Still, the subset of meaningful differences cannot exceed that of the total genetic differences. So, even if we do not, as of yet, understand all the details of how the species in question came to be biologically different, we can be sure that we know what the upper maximum is for the necessary mutations needed to bring about the differences we observe. So, while the total genetic differences between two species is an overestimation of the genetic changes needed to cause the differences, it is still a useful measure because we know that all of the meaningful changes must be accounted for within it.

Applying this test to humans and our closest (living) evolutionary relative, the chimpanzee, reveals that at a whole-genome level, we are over 95% identical. This value is even an underestimate, since it “counts” mutations that duplicate or delete sections of DNA as if they were separate mutations affecting individual DNA “letters” even though it was created by only one genetic change. Indeed if we use the same criteria to compare the diversity which exists within our own species, we humans are only 98% identical to each other. By whatever measure used, we are but a hand-breadth away from our evolutionary cousins at the DNA level (for those interested in a full treatment of how the human and chimpanzee genomes compare, please see this [recent article](#)).



* the 95% value is based on a very conservative measure that disproportionately counts duplications and deletions as *individual* mutations

Of interest for our purposes here is the simple realization that a relatively small number of subtle genetic changes undergird the large biological differences we observe between humans and chimpanzees. The increase in CSI associated with building the complex human brain and other distinctively human features in contrast to the body of our cousin, the chimpanzee does not appear to require huge changes at the genetic level. The differences we see, when examining these two genomes, are consistent with small changes of the sort easily accessible to evolutionary mechanisms. While this observation does not rule out the possibility of God directing this stage of human evolution in a more supernatural way, the genomics evidence suggests that this stage was accomplished in gradual, incremental steps. This observation also matches what we see in the fossil record, with gradual increases in brain capacity, tool making, and other features that mark us out as distinctly human.

Evolution and new CSI: cause for fear or celebration?

So, for evangelical Christians, what is perhaps the most challenging evidence for new CSI arising through evolutionary means comes from within our own genomes. Here we see that, at the genetic level, we are but a stone's throw from other primates such as chimpanzees. This realization leads to what may be for some an uncomfortable choice: either evolution is capable of generating significant novelty through mutation, genetic drift, and natural selection, or the differences between humans and other forms of life must be seen as insignificant. The only other option is to reject these lines of evidence altogether.

Of course, this response, for many, is one driven by fear: fear of having to re-consider the range of methods by which God creates, or perhaps how one interprets the opening chapters of Genesis. My hope is that this paper, while challenging for some, would not ultimately be cause for fear. Indeed, this response plays into the false "natural versus God" dichotomy discussed above. Rather, my hope is that understanding some of the natural means God uses to bring about biodiversity on earth, including for our own species, will provide an occasion to offer thanks and praise to our Creator.

Further reading:

Bridgham, J.T., Carrol, S.M., and Thornton, J.W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312: 97-100.

Dehal, P., and Boore, J. (2005). Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol* 3(10): e314

Ferrada, E., and Wagner, A. (2010). Evolutionary innovations and the organization of protein functions in genotype space. *PLoS ONE* 5(11); e14172.

Harms, M.J. and Thornton, J.W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology* 20: 360-366.

Thornton, J.W. (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Reviews Genetics* 5: 366-375.