A hypothetic mechanism for evolution of enzymatic pathways

Jorma Jormakka
Contact: jorma.o.jormakka@gmail.co

Abstract: This is basically a white paper that suggests for looking for a new mechanism that could explain the evolution of enzymatic pathways. Natural selection does not seem to be that mechanism as it is not likely that there is a long chain of mutations to an enzyme that on every step are advantageous. Random mutations to pseudogenes may be able to create the first enzyme in an enzymatic pathway, but cannot account for cases when several enzymes are needed for some task. The problem is getting all mutations to the same individual without some selection process that makes a gene sweep in the population. A learning mechanism, similar to the way the variable protein part of an antibody is found in the immune system is proposed, but as a cell-based mechanism in all cells, including germ cells, which is necessary in order to get inherited mutations. The last section describes the little evidence that I can mention to support the hypothesis.

1. The problem of the evolution of enzymes

Evolution of enzymes has always been one of the problems in neo-Darwinism, and it is also not explained by horizontal transfer of genes or genes introduced by retroviruses. Richard Dawkins addresses this essential problem only very briefly in one of his books (page 92 in [1] in the translation). Apparently Fred Hoyle had sent a letter to Dawkins pointing out how improbable it is to get an enzyme through random mutations. Dawkins explains in the book that Hoyle, a mere physicist, does not understand anything about natural selection. Of course, Dawkins states, these mutations are not made all on one step but in many small steps. Yet, as it is, Hoyle had a good point: neo-Darwinists should give a reason why such a chain of small steps should exist.

It is not at all likely that such a chain of mutations that are always improving survivability exists. We should expect that after few mutations an enzyme usually does not work in its original task as a catalyst of a chemical reaction. Very seldom the mutated enzyme can catalyze another useful reaction and fill some positive function in a cell. This seems to be the case with digestive enzymes in primates. The study [2] mentions several cases where a gene for a digestive enzyme has become duplicated, a fairly common error in copying DNA. In most cases in [2] the duplicate only produces more of the same enzyme, but occasionally it may mutate to a different enzyme and still be active. One such case in [2] is pancreating ribonuclease RNASE1 in African colobines (a less commonly known group of primates). However, the RNASE1 gene has mutated only few SNPs (single nucleotide polymorphism) from the original form.

This is exactly what we would expect: a gene must become duplicated since the original gene is needed for producing the original form of the enzyme. The duplicate is initially active and produces more of the same enzyme, which can be advantageous, but any one of the duplicates can mutate without endangering the production of the original enzyme. A mutated gene may still catalyze the original reaction and if so, the mutation is largely neutral. If the enzyme does not catalyze the original reaction after a mutation, there can be a rare case where it catalyzes some other reaction, but most probably it does not catalyze any useful reaction and the gene becomes a pseudogene. Usually pseudogenes are turned off by a mutation in the control part of some gene because producing unneeded and potentially harmful enzymes is not favored by natural selection. A pseudogene can mutate further, but as it is not active, it is not under any selection pressure, and it mutates randomly. Because of this,

we can make a rough estimate how much a pseudogene can mutate in a given time. Such an estimate is not precise: there are changes in mutation rates of base pairs in inactive DNA. These differences are related to other mechanism than natural selection. They may be caused e.g., by errors in copying DNA and in the behavior of DNA repair.

An obvious question of random mutations in a pseudogene is if this mechanism can in a given time create any DNA segment that encodes a protein that can act as a catalyst of any chemical reaction. We can first estimate how difficult it may be to find a protein catalyzing a pre-selected chemical reaction. An enzyme is a quite long amino acid chain, certainly over 30 amino acids. The number $20^{30}=10^{39}$ of combinations that can be made from 20 amino acids in this very short 30 amino acid chain, but this huge figure is not the number of trials needed to find a protein that can act as an enzyme for a pre-selected chemical reaction. The real number of trials must be much smaller, as the immunosystem can find an antibody to any alien protein in a short time. Comparing to the immunosystem we get a better estimate for the number of trials that are needed.

Humans have about 10^{10} B-cells that search for antibodies for alien proteins. An antibody has a variable protein that recognizes an alien protein in a way that is similar to the working of an enzyme: the variable protein of the antibody is like a key hole to the key of the alien proteine. B-cells form new variable proteins by encoding them from RNA that the B-cell combines from small RNA segments that are encoded from random small DNA segments of about 300 genes of the genome. Each ribosome codes proteins from RNA with the speed of one 200 amino acid chain in a minute, i.e., 10/3 animo acids in a seconds. Within two weeks, $1.2*10^6$ seconds, the body has created antibodies to the alien proteine. The immunosystem could try some 10^{17} combinations, i.e., $10^{10}*(10/3)*1.2*10^6=4*10^{16}$, but probably not all 10^{10} B-cells take part in the immuno reaction and many tried combinations must be the same. The real number of tried combinations may be 1/1000 of this number, some 10^{14} . This figure is then a sufficient number of trials for finding the variable protein of an antibody for any alien proteine. A sufficient number of trials for finding a protein that can catalyze a chosen chemical reaction may be of the same order, about 10^{14} .

The number 10^{14} is still rather large. Let us see if it is too large by looking at a protein that is not an enzyme. There is a mammalian specific protein superfamily SCGB, it contains mouse ABP and cat Fel d 1 secretoglobin proteines. These two proteins differ by 50% in their amino acids. Thus, the protein coding parts of the encoding gene differ by 50% in base pairs. Fel d 1 is 90 amino acids long which gives N=270 base pairs. The difference of 50% means the difference of 135 base pairs. This proteine family may have already already been in pre-mammals some 260 Myr ago, but mammals diversified 60 Myr ago and it is more likely that mouse ABP and cat Fel d 1 developed after this diversification. Let us use the the time T=260 Myr. A typical estimate for the average mutation rate is r=0.5*10-9 mutations per base pair per year. In T=260 Myr there should be p=rT=0.13 mutations per base pair. In average there should be pN=35 mutations in the protein encoding DNA segment in the time T. There are two lineages, mouse and cat. Each lineage should have 35 (different) mutations, thus the genes should differ by 70 mutations, but there is a 135 mutations difference and thus there must be 135/2=67.5 mutations in each lineage. That is 67.5-35=32.5 mutations over the average. The standard deviation of the binomial distribution of mutations is $\sqrt{p(1-p)N}$ =5.5, thus 32.5 mutations is 5.9 standard deviations. The probability of 5.9 standard deviations is on the range of 10^{-7} and it is highly unlikely to get in time T so many mutations in a single germ line where each generation has only one individual, and even harder to get two lineages deviate this much from the mean. If we set T=60 Myr, which is more realistic, then we should set $r=0.5T^{-1}=0.8*10^{-8}$ in order to get 50% base pairs changed in T. Basically, the mutation rate might be so high for some genes, but this evolution rate seems too slow for an average gene.

However, in a large population we also have the population size M. It is not a singe germ line with one individual in each generation. The population size can affect the calculation in two ways. Natural selection uses the population size so that each mutation is advantageous and causes a sweep in the population (or a large part of it). If the sweep is thorough the whole population, we can replace r by Mr, since it is irrelevant in what individual a mutation happens: positive mutations always spread to the whole population. Even a relatively small population size M allows natural selection to work very fast, as long as finding advantageous mutations does not become too difficult. In the beginning of a chain of mutations we may estimate that 1/100 or 1/1000 mutations are advantageous. Setting $M=10^5$, r=0.5*10-9, and the probability of a mutation being positive (and making a sweep) as q=1/1000, we need $T=p(qrM)^{-1}=10$ Myr for getting 50% of genes changed. This would be a good and fast way of evolution, but the problem is that q is unlikely to stay in a so high value. It most probably goes very fast to zero when there are more mutations.

The other way M can influence evolution is that we assume that mutations occur in pseudogenes. There is no sweep over the population, instead, there are several lineages. Assuming that the generation time is 3 years, which can be reasonable for small mammals, there are T/3 generations in the time T. In each generation, in each individual, we have in average $3Nr=3*270*0.5*10^{-9}=4*10^{-7}$ mutations. If $M=10^5$, then in 25 generations we have 1 mutation in the population. Mutations can be backwards mutations, but let us assume for simplicity that each 25th population differs from all earlier populations. In T=260 Myr we have 10^7 different populations, thus there are on the range of 10^7 different trials for the protein. It is basically possible to find one or two cases where the number of mutations deviates from the average number of mutations by 5.9 standard deviations.

The proteins in SCGB are not enzymes. We seem to have a problem finding 10^{14} trials in order to find a protein that can act as an enzyme for a pre-selected chemical process as we only have on the range of 10^7 different trials, but there are two issues that make it possible. Firstly, there are hundreds of potentially useful chemical reactions, this gives 10^2. Secondly, it is not necessary that a pre-selected gene segment becomes the enzyme. In the human genome there are about 20,000 protein coding genes and about as many pseudogenes [3]. Any one of these 10^4 pseudogenes can mutate into a new enzyme. In this way we get just about the correct number of trials, 10^{14}. It is possible to get one enzyme by random mutations to pseudogenes.

We see now a serious problem in obtaining a new enzymatic pathway through random mutations and natural selection. Such a pathway often includes several enzymes. The first enzyme catalyzes substrates in the cell to first intermediate products, the next enzyme takes one or more products as substrates and produces the second intermediate products, and so on, until we get to the final products. There are typically from one to few enzymes in a pathway. The classical comparison of enzyme and reaction as a key and keyhole, though misleading in some other cases, is quite helpful here: changing the enzyme by mutations means changing the key. Then the key does not fit to the old keyhole, thus we must find another keyhole where it fits, assuming that there is such a keyhole. Opening the door protected by the lock leads to a room where there is another unknown keyhole. We must find another key to this new keyhole, another enzyme to process the intermediate results to something. After having found the second key we get to the second room and there is again a keyhole. Every step of finding a new key is random mutations to a pseudogene, thus, there is no guidance by any selection pressure. What chance this lucky chain of guessing of the keys has to lead to any useful end products? Clearly, this cannot be the way new enzymatic pathways develop. Though one enzyme may be possible to obtain by random mutations to pseudogenes, surely this is not the way to get an enzymatic pathway with several enzymes, or two enzymatic pathways for the same metabolic task, as in lactose synthesis and lactase for digesting lactose.

Yet, there have appeared new enzymatic pathways, even in mammals: one enzymatic pathway produces lactose and a one-enzyme pathway breaks it to simple sugars for consumption. Let us first look at the second enzymatic pathway.

2. The case of lactose synthesis and lactase

The enzyme breaking lactose to simple sugars is a protein called lactase, encoded by a gene LCT (lactase-encoding gene). Birds and molluscs have a homologue to the mammalian LCT, as is shown in [4]. The authors of [4] compare precursors of the chicken homologue to the mammalian-specific lactase enzyme LPH (lactase-phlorizin hydrolase) in fig. 7 in [4]. Two of these precursors, LPHdII and LPHII, have 43 base pairs. Authors identify 14 base pairs as originating from the common ancestor of the human lactase precursor LHH and the homologous chicken and mussel proteins. The precursor LPH dIV has 66 base pairs. The authors mark a sequence of 19 bps in chicken and another sequence of 20 as deriving from the common ancestor, but the mussel homologue does not in fact have the same amino acids as the mammalian proteins.

We do not need more than figure 7 from [4]. Calculating 19/66=0.29 and 14/43=0.33, we notice that about one third of nucleotides in the protein coding part of the gene has remained, thus about two thirds have changed. The common ancestor of birds and mammals lived around 340 million years ago. Mammals developed the LCT gene about 200-150 million years ago. In at least one of the two lineages, one third of base pairs must have changed.

Can the difference between the protein coding part of chicken and mammalian lactase gene precursor be a result of random mutations under natural selection? It hardly can be. In mammals the protein coding part of the lactase encoding gene has not changed practically at all since it appeared some 130 million years ago, i.e., before the subclasses Marsupial and Placental diverged. This is different from what is the case with genes coding proteins that do not act as enzymes. For instance, proteins of hemoglobin (and the corresponding protein-coding DNA) have changed in mammals quite much after mammals diversified, but this is very seldom the case with enzymes. The stability of the protein coding part of LCT demonstrates the usual rule that enzymes cannot mutate many steps from the original form and in every step catalyze some reaction that is useful to the cell. A large number of mutations to the protein coding part can only happen when the gene is inactive, a pseudogene.

The lactase gene LCT has indeed mutated, and even very fast, but only in the control part. One LCT mutation, often given as a schoolbook case of natural selection, is a mutation causing lactase tolerance in humans. There are three (or more) single nucleotide polymorphisms (SNP, a point mutation in one nucleotide) that each make a human adult able to digest lactose. However, this is not a case of evolution where new genes emerge through mutations. It is simply adaptation of an existing gene pool to the environment. Those simple mutations must have happened every thousand years or so in every mammalian species since early mammals, only the mutation was never favored by natural selection in non-human mammals. These apparently common mutations were finally favored in humans, who decided to drink milk of other animals as adults, i.e., consume milk intended for infants of another species. No change has been made to the (extended) gene pool: simple mutations must be counted as belonging to the gene pool.

Random mutations could have changed two-thirds of base pairs in the protein coding part of LCT in the given time: the average mutation rate is on the range of 10^{-8}-10^{-10} per base pair and per year, and the time is on the range or 10^{8} years. Yet, this is not what has happened with the lactase-encoding gene precursors in [4] figure 7. Random mutations would not keep 19 base pairs unchanged for 150 million years. We would see mutations spread fairly evenly over the gene. There must be some different mechanism.

Let us now look at the lactose enzymatic pathway. Females of the class Mammalia produce milk and for that reason mammals have a protein coding gene that makes lactalbumin. Lactabumin is a protein family: it countains different variants of α-lactalbumin, β-lactoglobulin, and so on. In different mammalian species these proteins differ to a degree. The gene encoding α-lactalbumin is called LALBA and the LGB gene encodes a precursor to β-lactoglobulin.

There is good evidence that lactalbumin has evolved from c-lysozyme [5][6]. First c-lysozyme become duplicated, probably 300 Myr (million years) ago and then the duplicated gene mutated [6]. As it was a duplicate, the original gene segment still made the work the gene was expected to do and the animal did not suffer from the copy. Much later, around 200 Myr ago, the copy started working and produced milk for the new class Mammalia. At least this is a logical way how it could have happened.

If we assume that the duplicate gene worked during the period 300-200 Myr when it mutated from c-lysozyme to lactalbumin, then the intermediate stages had to be useful or at least not harmful to the animal. If they were useful or neutral, some intermediate stages of the gene should have been preserved in some animal species. If it were so, there would be genes that are closer to lactalbumin than c-lysozyme, but there are no such genes. Charles Dawrin explained such lack of intermediate forms by extinction of intermediates, but I find the explanation and the claim of the existence of beneficial intermediate forms not convincing. Thus, the mutating copy was not working before it emerged as lactalbumin. This implies that after the copy had mutated too far, it stopped working and become a pseudogene. During the time the copy gene was not working, it could not be acted on my natural selection or any other mechanism affecting phenotypes. Therefore, the mutations on it were purely random and their probability can be calculated in a simple way from the binomial distribution.

Reference [6] gives information that can be used for an estimate of the difference between hen's lysozyme and cow lactalbumin. The proteins lactalbumin and lysozyme had 123 residues. Of them 40 residues were identical and 27 residues were chemically similar. 15 positions could not be located in the study, the reason for this must be that they were different but the researchers could not say what they corrrespond to. Let us subtract this 15 because it can also be that the proteins were of different sizes. Three base pairs are needed to code one aminoacid and the residues mentioned in [6] apparently correspond to amino acids. That means that the gene is 3*(123-15)=324 base pairs long.

The gene for c-lysozyme has been active all the time and could mutate so that natural selection acted on it. According to [2] c-lysozime was created 600 Myr ago. There are many lysozyme proteins. Reference [6] gives the figure 30-55% similarity in lysozime proteins. Thus, about 45-70% of amino acids are different in these lysozyme proteins. If 45-70% of base pairs change in 600 Myr, then the mutation rate is $0.75*10^{-9}$-$1.2*10^{-9}$ mutations per base pair per year. This mutation rate agrees well with estimations of the average mutation rate in animals. A natural explanation for the good match is that the different lysozyme proteins have evolved by the lysozyme protein being duplicated and the duplicate mutating randomly. The original c-lysozyme cannot have mutated much, or at all, because it would have lost the original function, but the mutated gene may have become activate and the original gene may have disappeared.

According to [5] there is considerable variation in amounts of change along diverse lactalbumin lineages in mammals. These differences can also be explained by lactalbumin gene having been duplicated and mutated in the duplicate. The mutated gene has become activated and the original gene has disappeared. Thus, originally the protein coding part of the lactalbumin gene was simply one of the mutated versions of the protein coding part of the lysozyme gene. If this is how these genes have evolved, we would expect that the difference between hen lysozyme and cow lactalbumin is also 45-70%.

We can compare this estimate to the one in [6]. The 40 identical residues mean 120 identical base pairs in the gene. The 27 residues that were chemically similar in [6] were not identical. Thus, something was changed in these residues. If we assume that only one of the three base pairs defining an aminoacid was changed, there were 27 base pair changes and 2*27=54 base pairs were not changed. This means: of the 324 base pairs 120+54=174 bp were not changed and 27 were certainly changed. Of the remaining 123 base pairs we can say that they are not identical in the two genes, thus they were also changed. We get 174/324=54% not changed and 46% changed. This agrees with the estimate that 45-70% of base pairs were changed.

The conclusion is that lactalbumin seems to fit well to the scenario where a gene gets duplicated, the duplicate mutates and becomes a pseudogene. As a pseudogene the protein coding part mutates with random mutations. Finally the mutated gene becomes activated. It may replace the original gene. Here this simple mechanism seems to fit the data, but in the case of the lactase-encoding gene there must be a different mechanims.

These observations can be combined in the following hypotheses of the evolution mechanism of enzymatic pathways.

Hypothesis 1: The first enzyme in a new enzymatic pathway evolves by random mutations as a pseudogene. The pseudogene may have developed in many ways. It may be a duplicate of an active gene, or it may e.g. be an endogamous retrovirus. In any case, it is not active and it can evolve through random mutations. In a long enough time and with a large enough population there arises such inherited mutations in pseudogenes that can act as catalists to some chemical reactions in a cell. Then in some part of the population all cells of an individual contain the mutated pseudogene.

Next we should get other enzymes of an enzymatic pathway, or enzymes of another pathway that is connected with the first pathway, as lactase is connected with lactose synthesis. We notice that though the pseudogene is inactive, it still catalizes tiny amounts of some substrate in the cell into some product. The product is usually waste for the cell and the cell has mechanisms to remove waste. Such mechanisms involve the cell finding a proteine that can catalize the waste product to another product. In the case of waste which is not a proteine (like the intermediate products of the lactose synthesis enzymatic pathway), the cell mechanism for removing waste is in lysosymes. We will make the following hypothesis.

Hypothesis 2: Lysosymes act in a similar way as the immunosystem and try combinations of small pieces of nuclear DNA and produce RNA segments. If any of the RNA segments codes a protein that can catalyze waste products into something useful, or some other waste, the waste removal system notices it (like the immunosystem notices when it finds an antibody, it starts to produce the antibody in large numbers). The RNA segment may get reverse transcribed to the genome DNA. As this can happen in all cells, it can also happen in germ cells and therefore the new DNA may be inherited. In this system we do not have 10^9 B-cells that work in parallel. The system is much slower than the immunosystem, but if the population size is sufficient, like 10^5, the mechanism may find the next stage of an enzymatic pathway in some tens of thousands of years (i.e., 10^9*10^{-5}).

More generally the mechanism for creating enzymatic pathways could be that the first enzyme is created by random mutations to a pseudogene. If the pseudogene mutates to a DNA sequence that encodes a protein that can act as a catalyst to some chemical reaction, this protein is produced in tiny amounts even if the gene is not turned on. It catalyses a reaction that takes some substrates and outputs products that can be proteins or some other molecules. They are waste for the cell and are treated by cell's waste treatment mechanism. They find a new protein that can turn these waste products to some other products. The new products can again be waste, and a new protein is found for turning the new waste to something else. Finally the products are useful for something, or able to leave the cell (as lactose is not, but

glucose is). New RNA segments that code these new proteins gets inverse translated to DNA in many, or all, cells, including germ cells. In this way the mechanims becomes inherited.

Naturally this mechanism does not explain off all problems that neo-Darwinism has. Even if the evolution of lactose synthesis and lactase might be explained in this way, milk production is a very complicated process with five main pathways where lactose syntesis is just a small part of the first pathway, see [8]. Richard Dawkins, in his many books, gives the false impression that the evolution theory is more scientific than it is and that natural selection can explain much more than it actually can, including the claim that natural selection can explain even the most compicated problems of evolution. Fortunately, lately there have been other books directed to the general audience, such as David Quammen's book [9]. Though also that book is far too optimistic of hown much actually has been understood of evolution, it is still a step to a correct direction. Personally, I am not all that optimistic, as can be seen in [10].

3. Evidence for the proposed mechanism

Not being an expert on this topic, or any topic any more for that matter, I give here only some random observations which may or may not have any relevance to the hypothesis.

Richard Dawkins is undoubtedly the best known neo-Darwinist. In [11] (page 113 of 435 pages in the Finnsih translation) he describes Lenski's experiments of creting mutations in bacteria. As expected, in all cultures improvement speed decreases after a relatively small number of generations. Darwin's assumption that there always are advantageous small changes does not seem justified. Dawkins tells that in one of Lenski's cultures bacteria learned through two mutations to metabolize citrate. This is very unlikely if the mutations were SNPs. In bacteria much of evolution happens through horizontal gene transfer and one explanation is that the culture got some plasmid from some other bacteria. This would mean that the cuture was contaminated. A different explanation is that the genome already had the gene for citrate, but it was turned off if there is oxygen. Dawkins hints to this possibility by saying that the bacteria could not originally metabolize citrate at least in the presence of oxygen.

As for mutations to protein coding parts, exons, such mutations seem quite rare. As an example from the work of David Haussler, humans have a gene family NOTCH2NL, which may explains out larger brains. Monkeys and even orangutans do not have it: they only have the ancestral gene NOTCH2. Gorillas and chimpanzees have an ineffective form of NOTCH2NL and only humans have this gene in a working version, but Neanderthals also had it. It seems to have developed from the ancestral NOTCH2 3-4 million years ago in the same way new protein-coding genes usually arise: first the NOTCH2 become multiplied, but the copy was only partial and did not work. In gorillas and chimpanzees it is in an ineffective form, but in humans it got corrected by a mutation (that is, when it was a pseudogene) and now it works well. But it is not a new protein-coding exon. This new gene seems to work by a mutated control part turning off a function. The new NOTCH2NL gene did not start separation of one type of cells as it did in the ancestral version and therefore the gene continued producing neurons and we got bigger brains. But what we have here is turning existing exons on and off.

Then we have the interesting question of what the pseudogenes are for. Some years ago inactive DNA was known as junk DNA. Torrents et all [3] count 20,000 pseudogenes in human genome and estimate that there may be more pseudogenes than genes. They also say that most pseudogenes in human and mouse genome have been created after the common ancestor of human and mouse lived, and that three fourths of pseudogenes have retrotranspositional origin and the rest come from gene duplication. Retrotranspositional origin means that these genes were reverse translated from RNA to DNA and inserted, usually

randomly, to the genome. A part of retrotranspositional pseudogenes, in human 5-8%, arise from retroviruses. It is not stated in [3] where the rest of retrotranspositional pseudogenes come from. Possibly they could come from RNA created by human cells, which would imply that there is reverse translation of RNA that originally was not in the genome, like RNA combined from small RNA segments encoded from small segments of genome DNA, as the proposed mechanism suggests.

Humans are quite complex, so let us look at a worm [12]. The worm C. *elegans* has 22,227 protein-coding genes (humans have c. 26,000). It has at least 561 annotated pseudogenes, though the number can be larger. Exons are coding parts in a gene. Their average size in the worm is 123 base pairs (humans 386). There are 6.4 exons in a gene (humans 8.8). Introns are between exons. They are non-coding parts. The average size of introns in the worm is 47. The size of the worm gene size is about 3000 base pairs (10,000-15,000 in humans). Exons are 26% of a gene (similar in a human). We notice that there are fewer pseudogenes. It is possible that pseudogenes have some role in evolution, as the worm probably is less evolved than a human. There is a very interesting question of the genome size: lungfish, the origin of land vertebrates, and some salamanders, also close to early forms of land vertebrates, have very large genomes. Probably the large size is non-coding DNA. Amoeba, a eukaryotic single cell organism, has very large genome, even though it is not as large as originally appeared. It does not look like this DNA is quite junk DNA. In may have a role, or be a side product form something that has a role. retroviruses and horizontal gene transfer by parasites can explain part of this, but maybe not all. Ribosomal RNA increases by accretion of new parts to a frozen common core. Thus, it is not by small changes by simple mutations as Charles Darwin would maybe have preferred. What is the mechanism how new units become added to the core? There is Lamarckian inheritance in the sense of horizontal gene transfer. There may even be macro-mutations with evolutional importance, or what should one say of fusing of two chromosomes. The primate chromosome number was 48, but in humans two chromosomes fused together to chromosome 2 giving us 46 chromosomes. Figure 2 in [13] dates these changes.

Lastly, let me add some support to the claim of a cell-based immune system type mechanism. One hint is that the cell marks proteins that it uses so that they are not treated as waste. This marking is probably like marking of alien proteins in the immune system. As the genome can obtain new genes, this system must be able to learn new proteins and this learned identification probably becomes inherited. The second hint is CRISPR (clustered regularly interspaced short palindromic repeats) associated genes, CRISPR-*cas*. Quammen [9] makes some comments on Mojica's work, apparently these jumping genes have some immune system type behavior: it is proposed that they protect the cell from invading DNA. I cannot discuss this difficult and new topic, but cells seem to have mechanisms that resemble in a sense those of the immune system.

References:

[1] Richard Dawkins, *Climbing Mount Improbable*, 1996. (Translation in Polish in Prószyński i S-ka, 1998.)

[2] Mareike C. Janiak, "Digestive Enzymes of Human and Nonhuman Primates," *Evolutionary Anthropology (*216), 25:253–266.

[3] David Torrents et al, "A Genome-Wide Survey of Human Pseudogenes," Genome Res., Dec; 13(12), (2003), 2559-2567.

[4] JN Freund et al, Identification of homologues of the mammalian intestinal lactase gene in non-mammals (birds and molluscs). Biochem J. (1997) 332, 491-498.

[5] EM Prager and AC Wilson, "Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences," *J Mol Evol.* 1988;27(4):326-35.
https://link.springer.com/article/10.1007%2FBF02101195

[5] Keith Brew, Thomas C. Vanaman and Robert L. Hill, "Comparison of the Amino Acid Sequence of Bovine α-Lactalbumin and Hens Egg White Lysozyme," *Journal of Biological Chemistry* 242(16):3747-9 (1967)
http://www.jbc.org/content/242/16/3747

[6] David M Irwin, Jason M Biegel, and Caro-Beth Stewart, "Evolution of the mammalian lysozyme gene family," *BMC Evol Biol*. 2011; 11: 166.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3141428/

[7] M. G. Grütter, L. H. Weaver & B. W. Matthews, "Goose lysozyme structure: an evolutionary link between hen and bacteriophage lysozymes?", *Nature* Vol. 303, pages 828–831 (1983), https://www.nature.com/articles/303828a0

[8] James L Mcmanam and Margaret C Neville, "Mammary physiology and milk secretion," *Advanced Drug Delivery Reviews* (2003), 55(5):629-41.
https://www.researchgate.net/publication/7457351_Mammary_physiology_and_milk_secretion

[9] David Quammen, *The Tangled Tree. A Radical New History of Life*, 2018. (Translation in Polish in Zysk i S-ka, 2020.)

[10] Jorma Jormakka, "Evolution theory and intelligent selection," a simple story of my frustration after reading five books by Dawkins, in ResearchGate (it is not Intelligent Design), 2021.

[11] Richard Dawkins, *The Greatest Show on Earth, The Evidence for Evolution* (2009). (Translation in Finnish in Terra Cognita, 2009.)

[12] Liran Avda, Anat Nitzan, Ronen Zaidel-Bar, "Worm Cool Kit: An Online CRISPR Planner of Point Mutations to Facilitate Modeling of Human Genetic Variations in C. elegans Orthologs," *WBG*, (2020)
http://www.wormbook.org/chapters/www_overviewgenestructure/overviewgenestructure.htm

[13] Y. Fan et al, "Genomic Structure and Evolution of the Ancestral Chromosome Fusion Site in 2q13–2q14.1 and Paralogous Regions on Other Human Chromosomes," *Genome Res*. (2002): 1651-1662.