

Genome truncation vs mutational opportunity: can new genes arise via gene duplication?—Part 1

Royal Truman and Peter Borger

Gene duplication and lateral gene transfer are observed biological phenomena. Their purpose is still a matter of deliberation among creationist and Intelligent Design researchers, but both may serve functions in a process leading to rapid acquisition of adaptive phenotypes in novel environments. Evolutionists claim that copies of duplicate genes are free to mutate and that natural selection subsequently favours useful new sequences. In this manner countless novel genes, distributed among thousands of gene families, are claimed to have evolved. However, very small organisms with redundant, expressed, duplicate genes would face significant selective disadvantages. We calculate here how many distinct mutations could accumulate before natural selection would eliminate strains from a gene duplication event, using all available 'mutational time slices' (MTSs) during four billion years. For this purpose we use Hoyle's mathematical treatment for asexual reproduction in a fixed population size, and binomial probability distributions of the number of mutations produced per generation. Here, we explore a variety of parameters, such as population size, proportion of the population initially lacking a duplicate gene (x_0), selectivity factor (s), generations (t) and maximum time available. Many mutations which differ very little from the original duplicated sequence can indeed be generated. But in four billion years not even a single prokaryote with 22 or more differences from the original duplicate would be produced. This is a startling and unexpected conclusion given that 90% and higher identity between proteins is generally assumed to imply the same function and identical three dimensional folded structure. It should be obvious that without new genes, novel complex biological structures cannot arise.

Novel metabolic networks, signal cascades, bone joints, wings, sonar, brains, immune systems, nervous systems, and so on, are coded for by collections of genes.¹ According to evolutionary theory, millions of new and very different genes had to arise, starting from simpler genomes. New genetic material might become available by duplication of genes² or imported.³⁻⁶ Sometimes natural selection is assumed to be involved, in other models it would not play a major role.⁷

Would something useful occasionally be produced? This depends on how many mutational attempts would be necessary. In our analysis we shall take into account a key fact which has been neglected in the evolutionary literature: small genomes carrying expressed genes not immediately needed will be subject to natural selection. Many different duplicate genes could be generated over time but serve little evolutionary purpose if each lineage is weeded out before enough mutations could accumulate to permit novel genes to be produced.

Candidate organisms

Gene sequence comparisons^{8,9} suggest that little gene transfer from bacteria or archaea to multicellular eukaryotes has occurred. Bacteria and archaea strains may benefit from genes provided from other single-celled organisms, via several processes collectively called Lateral or Horizontal Gene Transfer (LGT or HGT). The mechanisms are clearly designed.¹⁰⁻¹³ Evolutionary phylogenetic assumptions are the major reason for invoking HGT, but the true extent

of this activity, based on various statistical indices,¹⁴ is controversial.

There are several reasons why new genes have the best chance of arising among single-celled prokaryotes as opposed to more complex organisms:

- Huge populations permit more gene alternatives to be generated.
- Generation times are short.
- Non-sexual, fission reproduction does not dilute inherited change.
- They have supposedly existed for about 4 billion years.¹⁵

Two competing factors determine the chances of producing a novel gene: (a) genome streamlining, and (b) the number of mutational alternatives produced.

We will examine the net outcome of these two effects over many generations.

Proportion of a population carrying superfluous genes

The proportion over time of a fixed sized population of bacteria or archaea (i.e. with fission type reproduction) having an advantageous feature providing a selectivity coefficient, s , can be calculated¹⁶ using equation (1):

$$f = \frac{x_0 e^{st}}{1 + x_0 (e^{st} - 1)} \quad (1)$$

where f is the fraction of a population which possesses a particular property; s is the selectivity coefficient favouring propagation of the property; t is time, and refers here to number of generations; $x_0 = f$ at $t = 0$, and refers here to organisms shedding unneeded duplicate genes.

Genome expansion vs compression

The specific details of the selectively useful feature in equation (1) are irrelevant. As long as the biological feature is passed on according to fission-type reproduction and it offers on average a selective advantage, the mathematical description describes correctly what will happen to the population over time. In our analysis the selectivity coefficient, s , refers to the loss of unnecessary genetic material, whether having originated via LGT or chromosomal gene duplication during replication. Genes can also be lost when DNA polymerase skips a region of DNA during genome replication, producing a truncated daughter chromosome. We shall neglect this major contribution to genome streamlining.

Natural selection will disfavour lineages with larger genomes *ceteris paribus*: (i) there is a significant metabolic cost, and (ii) the generation times will be longer.

Metabolic costs for most individual genes of the eukaryotic microbe *Saccharomyces cerevisiae*, and an estimated total amount of energy per second generated, were calculated recently.¹⁷ In Part 2¹⁸ we show that for prokaryotes, s for an unnecessary gene is about the reciprocal of the number of chromosomal genes.

The effect of longer chromosome replication time is also examined in Part 2.¹⁸ Both factors favour genome truncation, suggesting we should use a value of s , favouring strains lacking a superfluous duplicated gene, of between 10^{-4} and 10^{-3} . The smaller the genome, the stronger the streamlining effect.

Probability distribution of mutations

For typical bacteria, each nucleotide (nt) replicated has a failure rate of about 10^{-10} .¹⁹ To favour the evolutionary model we shall use a mutational rate of 10^{-9} /nt. The average number of mutations on the extra gene can be calculated over many generations using the binomial distribution⁸ (discussed more fully in Part 2):

$$p(m) = \frac{n!}{m!(n-m)!} p^m q^{n-m} \tag{2}$$

where p = probability of a success per trial (the mutation rate); $q = 1-p$; n = number of trials (adding a new nucleotide during DNA replication); m = number of successes after n trials (mutations).

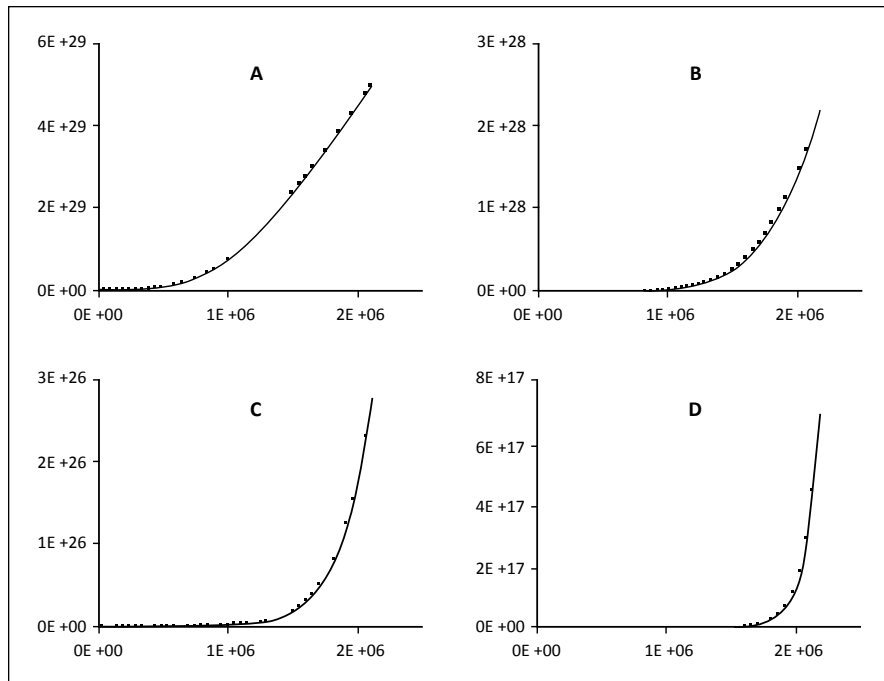


Figure 1. Number of prokaryotes with m mutations generated over time when selectively neutral. Initial population size= 10^{31} , proportion with duplicate gene, $x_0=0.5$, $s=0.42$. Y axis: Number of prokaryotes with m mutations: A: $m=4$; B: $m=7$; C: $m=10$; D: $m=20$. X axis: Generations.

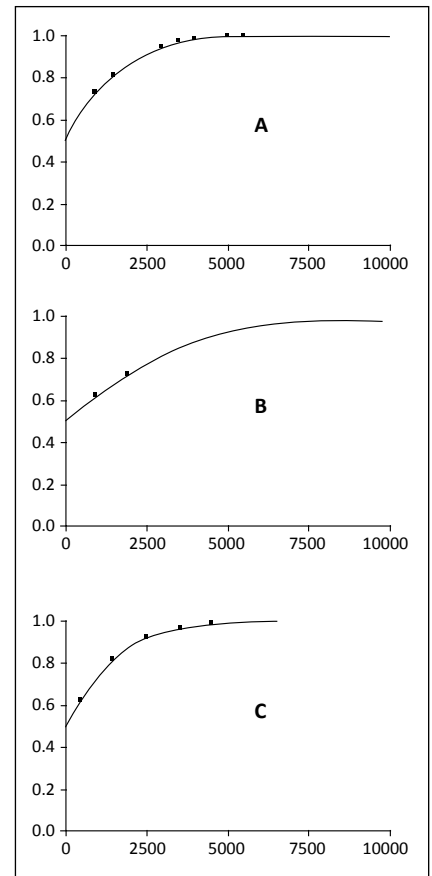


Figure 2. Population proportion lacking a duplicate gene as a function of generations and selectivity factor favouring genome truncation. Initial population size= 10^{31} , proportion lacking duplicate gene, $x_0=0.5$. Selectivity factors, s , favouring genome truncation: A: $s=0.001$;⁴³ B: $s=0.0005$;⁴⁴ C: $s=0.0001$.²³ Y axis: Proportion of a population. X axis: Generations.

Population size

Let us assume that in 4 billion years there would be on average 10^{31} prokaryotes/generation^{20,21} and estimate the maximum number of mutants eventually produced per duplicate gene.

Calculations. Mutations on the duplicate gene would rarely occur. Each of about 1,000 nt positions has a chance of 10^{-9} of mutating per generation. Then a single random nt mutation on a duplicate would require about one million generations, assuming the size of the duplicate being 1 kb (the average size of a prokaryotic gene). For each generation since a duplicate gene arose, the number of members which still carry the duplicate was calculated using the population size and equation (1), taking the negative selectivity into account. This value was multiplied by the probabilities of accumulating 0, 1, 2, 3 ... mutations up to that number of generations, calculated with equation (2), to give the number of organisms with a certain number of mutations during that generation (see also eq. (3) in Part 2¹⁸).

All calculations and curves were performed using Microsoft Excel. Most of the spreadsheets are available online.²²

Methods and results

To establish an upper limit which favours the evolutionary viewpoint, we assumed half the world's prokaryote population²¹ would initially have a duplicated gene of any kind (0.5×10^{31} members). The maximum number of mutational differences from the original gene which would be generated before natural selection eliminates such lineages was calculated.

If duplicate genes were selectively neutral, many mutants would indeed result (figure 1). Over time the duplicate gene

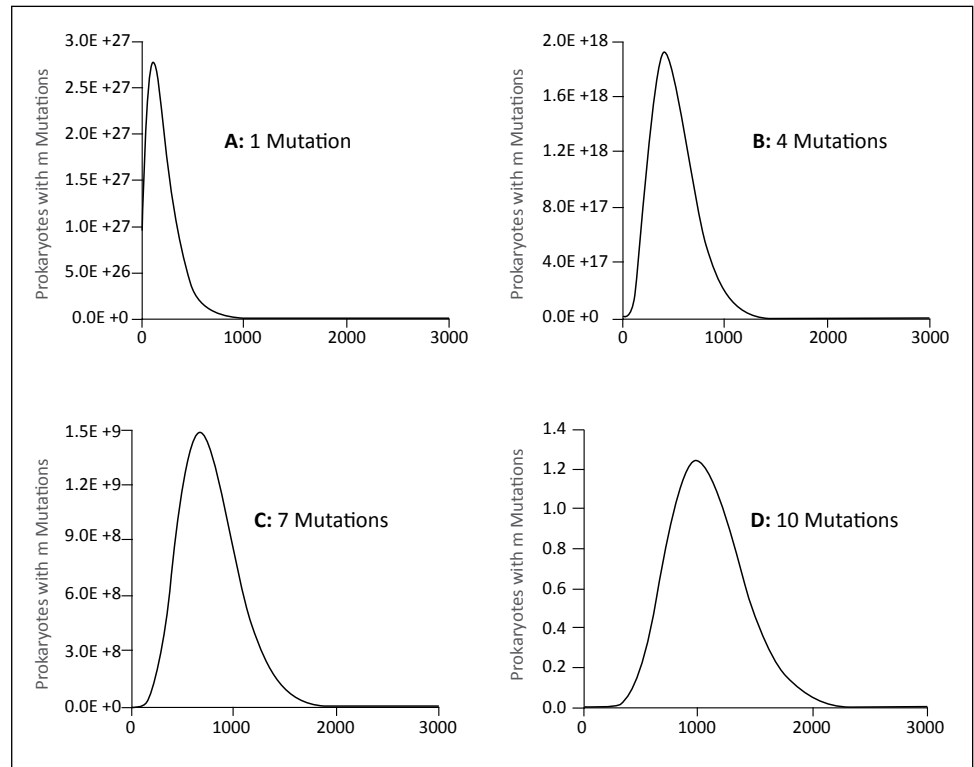


Figure 3. Prokaryotes with m mutations, selectivity factor favouring truncation $s = 0.001$. Initial population size = 10^{31} , proportion with duplicate gene, $x_0 = 0.5$.⁴³ Y axis: Prokaryotes with m mutations: A: $m=1$; B: $m=4$; C: $m=7$; D: $m=10$.

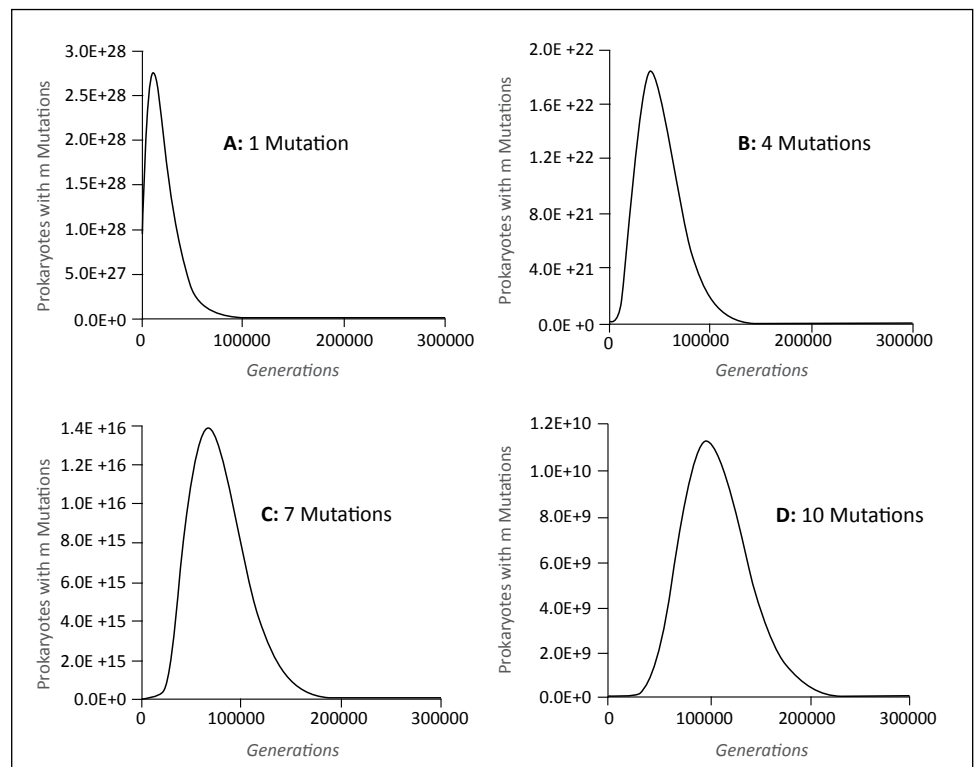


Figure 4. Prokaryotes with m mutations, selectivity factor favouring truncation $s = 0.0001$. Initial population size = 10^{31} , proportion with duplicate gene, $x_0 = 0.5$.²³ Y axis: Prokaryotes with m mutations: A: $m=1$; B: $m=4$; C: $m=7$; D: $m=10$. X axis: Generations.

would increasingly differ from the original sequence. Within two million generations a considerable number of prokaryotes having up to 20 mutations would exist (figure 1D). However, the number of mutations which could actually be generated from a duplicate gene is counteracted by the genome streamlining effect, as we shall demonstrate.

The single duplicate gene case

Natural selection would rapidly eliminate most individuals carrying duplicate genes before even a single mutation would occur. We assumed 0.5×10^{31} members (i.e. half the population) would initially have a duplicate to see if these daunting odds could be overcome. Based on figure 2A, for $s=0.001$ virtually none would have this gene after about six thousand generations and for $s=0.0001$ within about sixty thousand generations (figure 2C). Figure 3D shows (for $s=0.001$) that not even 2 members out of the 10^{31} prokaryotes would manage to accumulate $m=10$ or more mutations in any generation before those strains have been eliminated by natural selection.²³

Relaxing the penalty to a less realistic $s=0.0001$ (figures 4 and 5) shows that not even one member in any generation would accumulate 15 or more mutations.

Positive selection considered

We showed above that the proportion of organisms possessing highly mutated duplicate genes in any generation is miniscule. A few generations after the gene duplication event no mutations would have occurred on the duplicate yet. Natural selection steadily decreases the number of members carrying a duplicate gene, including the later ones which eventually do accumulate some mutations. Most highly mutated variants quickly go extinct. Only if a particular combination of mutations confers an immedi-

ate and dramatic selective advantage might such a variant occasionally escape rapid extinction. Although less than one individual having fifteen or more mutations is expected for *any given* generation ($x_0=0.5$ and $s=0.0001$, figure 5), we need to take *all* the generations into account before this collection of prokaryotes has gone extinct.

The logic behind these new calculations is explained more fully in Part 2.¹⁸ In brief, all mutant lineages are added up by integration of the mutant number over generations. Of the 0.5×10^{31} organisms initially having a duplicate gene ($x_0=0.5$ and $s=0.001$), less than ten individuals having eleven mutations, and none with twelve mutations or more would be produced (table 1, figure 6). Clearly, if less than half the world's prokaryote population ($x_0=0.5$) was initially endowed with a duplicate gene free to mutate, fewer (if any) individuals with eleven or more mutations would be generated before all were removed by natural selection (figure 7, table 2).

Using $s=0.0001$ predicts considerably more individuals with m mutations, but none with 18 or more mutations (table 3, figure 8). Although some members with 16 or 17 mutations could have arisen from the original duplicate, in the absence of an immediate positive selection these soon die out. That is why the proportion of highly mutated duplicate genes during *every generation* is miniscule. These are not likely to survive for thousands of generations, each generation facing a negative selection, waiting for addition mutations, one of which may turn out to be advantageous. The highly mutated individuals actually alive at any point always represent an insignificant proportion of the total population, and face daunting probabilities of ever fixing.

The multiple duplicate gene case

If several duplicate genes are present on a genome, more random mutations could occur on one or more of them. Perhaps this would improve the chances of creating a new gene by chance mutations. But countering this effect is the overall increased negative selectivity for such individuals.

An illustrative example is given in table 5, based on fifteen mutations and one vs five duplicates. Using $s=0.0001$ per gene implies an overall $s=0.0005$ for that organisms, *ceteris paribus*, when five duplicates are present. We find that five times more prokaryotes with $m=15$ are generated if only a single gene was initially duplicated (figure 9C vs figure 9D). Attrition by natural selection of strains having five duplicates is more significant than the five times greater number of mutational possibilities (figure 9A vs figure 9B). Thus, taking natural selection into account demonstrates that starting off with larger numbers of duplicate genes in prokaryotes does *not* improve the odds of producing new genes via random mutations!

The prokaryotes having five duplicates would also compete against subsequent lineages having only one to four duplicates due to gene losses during chromosome replication. Therefore, attrition of the multiple duplicate variants would occur even more rapidly than if only two

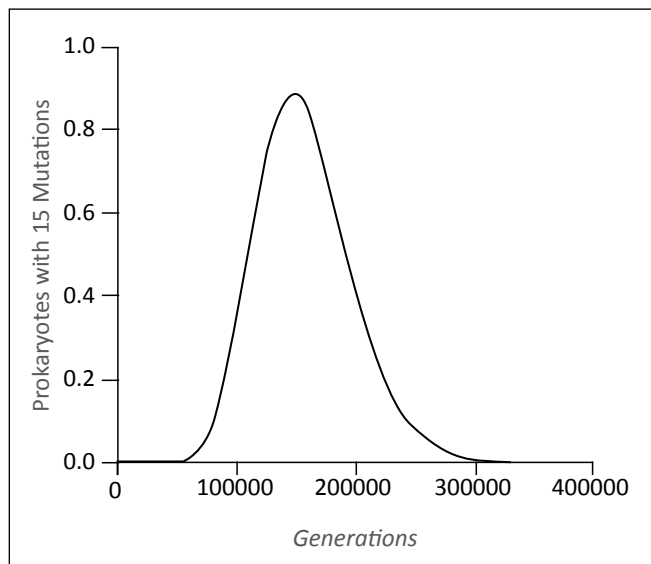


Figure 5. Number of prokaryotes having $m=15$ mutations. Selectivity factor favouring genome truncation, $s=0.0001$, initial population size = 10^{31} , proportion initially lacking duplicate gene, $x_0=0.5$.²³

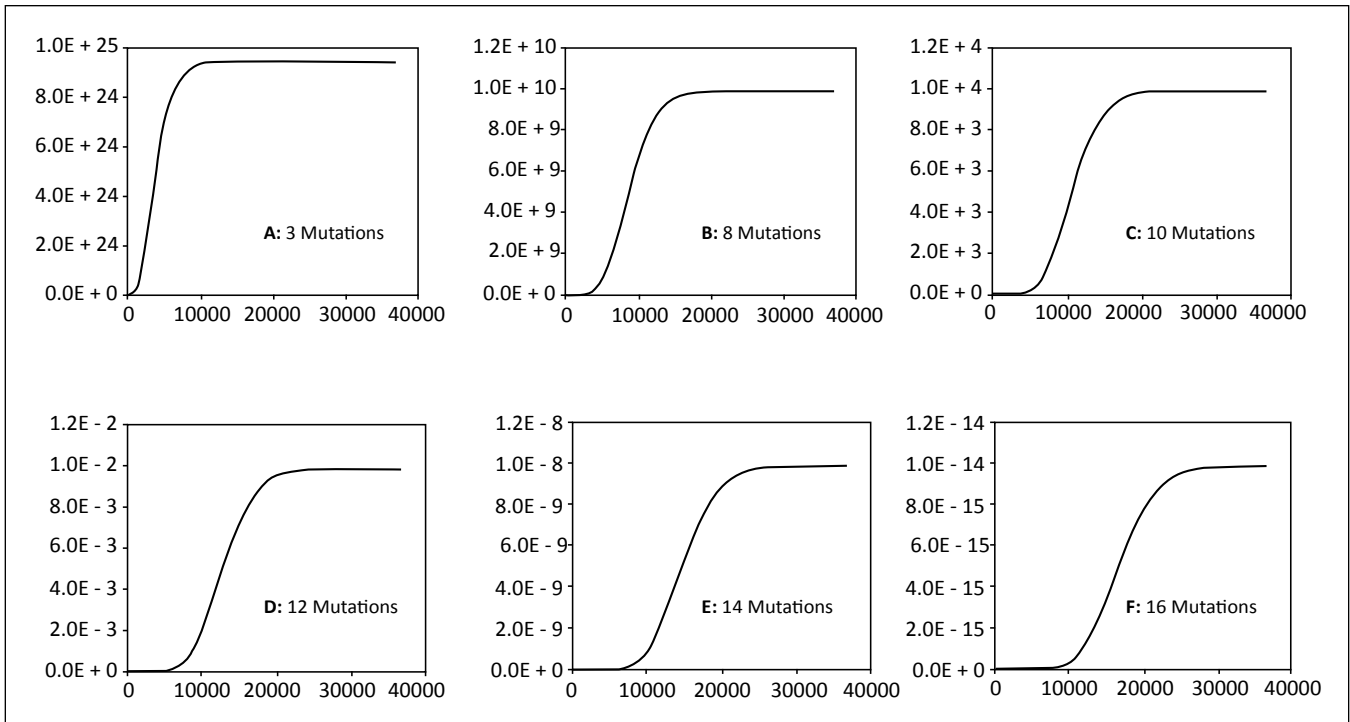


Figure 6. Maximum number of prokaryotes with m mutations from a single duplicate gene event. Selectivity favouring genome truncation, $s=0.001$; initial population size= 10^{31} ; proportion initially lacking a duplicate gene, $x_0=0.5$. Y axis: Number of prokaryotes generated during the MTS having m distinct mutations: A: $m=3$; B: $m=8$; C: $m=10$; D: $m=12$; E: $m=14$; F: $m=16$. X axis: Generations.⁴⁵

possibilities existed (e.g. a model which compares zero vs five extra genes).

The key issue is that the penalty for carrying extra genes, causing rapid elimination by natural selection, outweighs the extra mutational opportunities provided. Furthermore, the chances of interference with existing biochemical processes and of producing sub-optimal stoichiometric proportions of proteins would increase with numbers and kinds of duplicate genes.

Small vs large populations

We evaluate next a scenario whereby a large number of isolated smaller populations are present. A useful mutation would then be more likely to fix than within huge populations. Calculations were performed using 10^{11}

members, about a dense litre full of bacteria. For $s=0.0001$ no descendent from a single gene duplicate would attain eight or more mutations (figure 10B), and using $s=0.001$ none would accumulate six or more mutations (figure 10C).

Multiple attempts to generate new genes

We can estimate how many different variants of a duplicate gene could be generated over four billion years. We saw above that if half the world's population of prokaryotes had a duplicate gene initially, no descendents with 18 or more amino acid modifications would be produced, assuming a generous attrition selectivity of only $s=0.0001$. However, after these mutated lineages die out, additional evolutionary attempts could occur. Let us call

Table 1. Summary of results based on an initial population size of 10^{31} prokaryotes; $x_0=0.5$ (initially half possess a duplicate gene free to mutate). Selectivity factor favouring removal of a duplicate gene, $s=0.001$. See Fig. 6.⁴⁵

n mutations:	8	9	10	11	12	13	14	15	16
Maximum surviving mutants in any generation:	1.4E+6	1.3E+3	1.2E+0	1.2E-3	1.1E-6	1.1E-9	1.1E-12	1.0E-15	9.8E-19
Generation with maximum surviving mutants:	7 995	8 992	9 990	10 989	11 988	12 987	13 986	14 985	15 984
Total different mutants per MTS (a):	9.9E+9	9.9E+6	9.9E+3	9.9E+0	9.9E-3	9.9E-6	9.9E-9	9.8E-12	9.8E-15
Plateau for new mutants, generations (b), (c):	18000	20000	23000	24000	26000	28000	30000	32000	35000
Maximum mutants ever produced (d):	5.7E+19	5.1E+16	4.5E+13	4.3E+10	3.9E+7	3.7E+4	3.4E+1	0.032	2.9E-5

(a) MTS: "Mutational Time Slice", see main text.

(b) Approximate generation where virtually no new mutants form with a specific number of mutations. By visual inspection of Fig. 6.

(c) Due to round-off errors, calculations were carried out to only 36,700 generations.

(d) Based on 26,000 generation per year, 4 billion years evolutionary time, the resulting number of MTS available, and the number of total different mutants per MTS.

Table 2. Sensitivity analysis for x_0 , proportion of population assumed to lack a duplicate gene initially. Initial population size = 10^{31} prokaryotes. Example based on $m=11$ mutations. Selectivity factor favouring removal of a duplicate gene, $s=0.001$.⁴⁵

x_0 :	0.9	0.99	1
Maximum mutants in a generation:	1.3×10^{-4}	1.2×10^{-5}	1.2×10^{-6}
Total different mutants from 0.5×10^{31} batch:	1	0.1	0.01

the interval between duplicate appearance and extinction a ‘mutational time slice’ (MTS). We need to estimate an upper bound to how many MTSs could be produced during four billion years.

Now, instead of using time until extinction of all copies of duplicate genes, we note that the accumulated number of mutants generated during an MTS eventually levels off (figures 6 and 8). Let us record the number of generations at the point where hardly any additional individuals with some number m of mutations are being produced. During that interval some duplicates would have mutated the whole preceding time and some novel duplicates would have been just generated. Let us minimize waste of the limiting resource, time. As soon as this levelling off plateau is reached, we assume all these organisms graciously forfeit their role in history, freeing up their *Lebensraum* to permit a new MTS to be initiated. Half the world’s prokaryote population will again be instantaneously graced with a duplicate of any gene, to initiate a new MTS.

Let us assume all new variants represent a unique mutation, having never been generated in any earlier MTS, so the overall sum (over all possible MTSs) defines the maximum variability which could ever be produced. From the number of generations ‘used up’ during an MTS, a generation time of about twenty minutes, and a total of four billion years it is easy to estimate the number of MTS sequences which would be available. From table 1, using $s=0.001$ predicts that about 34 prokaryotes having 14 mutations could be generated in all assumed evolutionary history, and none with 15 or more mutations. A lower penalty of $s=0.0001$ predicts none would ever be produced with 22 or more mutations (table 3, figure 8F).

No prokaryotes ever produced with an estimated number of mutations on a duplicate genes ranging between 15 and 22 is a sobering insight, and this estimate is surely too generous. For example, initiating each MTS with 0.5×10^{31} members ($x_0=0.5$) is probably too high, especially since this means actively expressed genes from the beginning until the end of the MTS (otherwise natural selection would not be able to identify a useful combination of mutations). Sensitivity analysis with alternative x_0 values shows (table 2 and table 4) that a better estimate for number of variants ever generated may well be much lower.

Increasing the number of MTSs

Shorter prokaryote generation times would provide more MTSs but would simultaneously increase the rate of attrition per time period. As more fully discussed in Part 2 and intuitively obvious, shorter generation times leads to proportionally greater selective disadvantages of carrying superfluous genes.

Discussion

Our analysis presents a very different picture from what is commonly taught today. It is often claimed that given ‘enough evolutionary time everything is possible’. In a leading standard textbook on cell biology, whose main author is U.S. Academy of Science president and Harvard professor Bruce Alberts, one reads,

‘... only about one nucleotide pair in a thousand is randomly changed every 200,000 years. Even so, in a population of 10,000 individuals, every possible nucleotide substitution will have been “tried out” on about 50 occasions in the course of a million years, which is a short span of time in relation to the evolution of species.’²⁴

This claim is absurd. With four possible nucleotides at each DNA position, the fifth part of a single average size gene alone offers about 4^{200} possibilities, which is more than 10^{120} alternatives. In 4 billion years all organisms which ever lived could not have ‘tried out’ but an insignificant fraction of the alternatives of even one single gene, far less in but a million years as claimed.

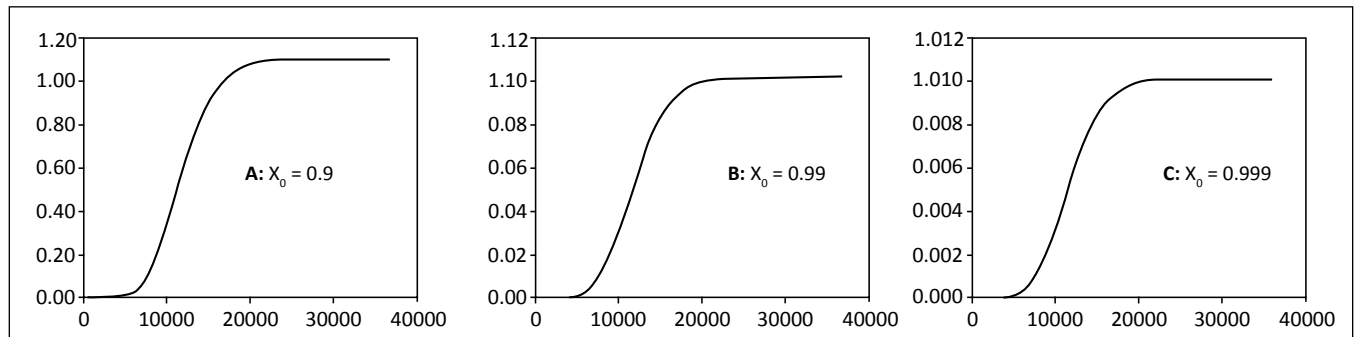
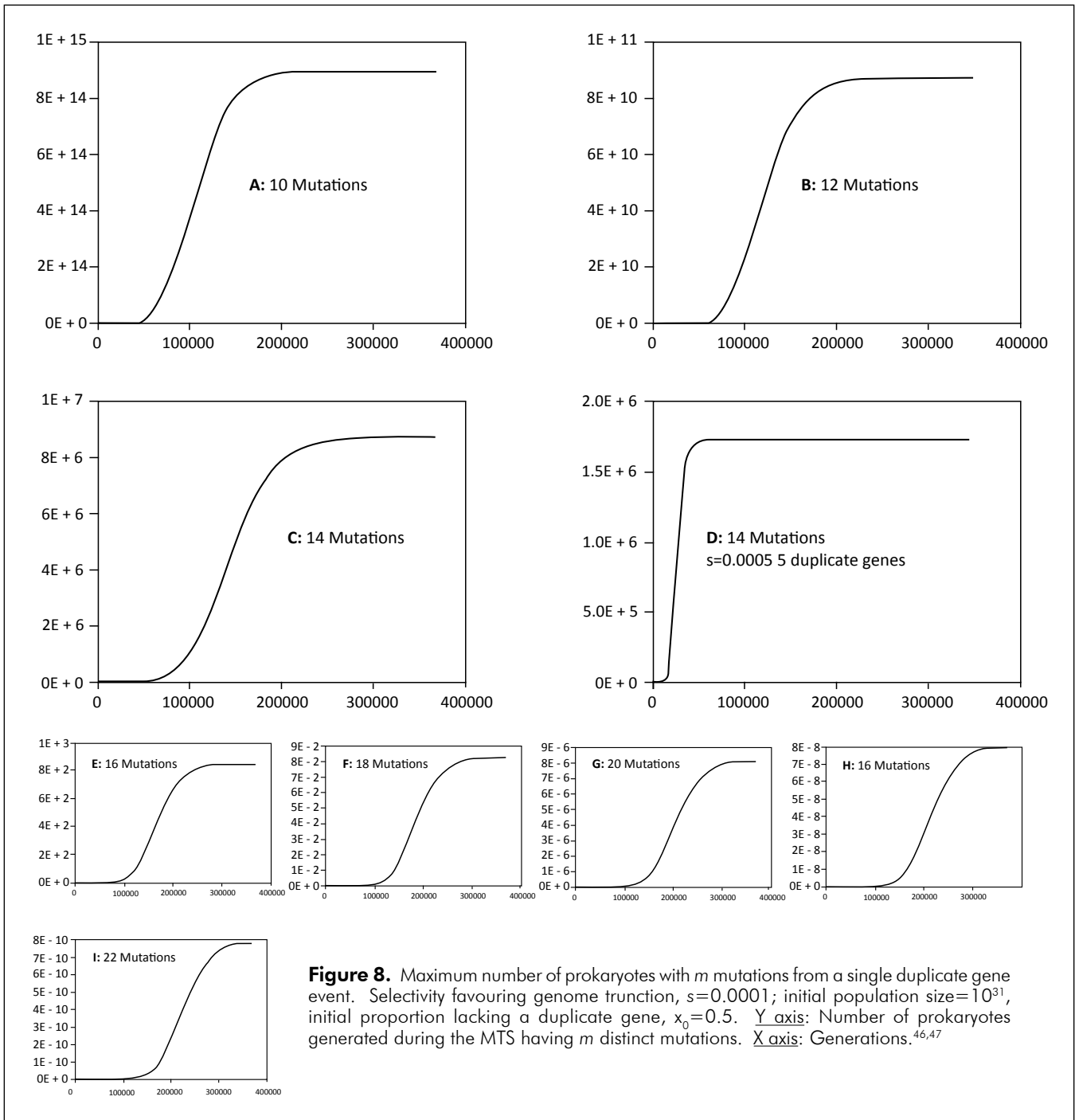


Figure 7. Maximum number of prokaryotes with $m=11$ mutations which could have arisen from a single duplicate gene event. Sensitivity analysis of alternative initial proportions lacking a duplicate gene, x_0 . A: $x_0=0.9$; B: $x_0=0.99$ C: $x_0=0.999$.⁴⁵ Selectivity favouring genome truncation, $s=0.001$; initial population size = 10^{31} . Y axis: Number of distinct mutations generated. X axis: Generations.



Without experimental data under natural conditions, we cannot know how often a duplicate gene would be generated and then fixed in a population. Current estimates are based on phylogenetic trees driven by evolutionary assumptions and not real data. Created genomes may have included both identical duplicates and similar genes for robustness and protein dosage reasons. Subsequent gene loss during chromosome replication can occur, and under the right circumstances be selectively advantageous. This process would be slow, leading to a mixture of strains with various numbers of gene copies today.

We suspect that distinct prokaryote strains for each taxon may have been created in relatively small numbers, for and in different environments. Subsequent distribution over the earth plus mutations would lead to the distributions observed today.

Tandem genes

Duplicate copies of genes do exist, and these are very often identical or nearly so.²⁵ They may have arisen in the very recent past and lack enough time to have diverged. There may also be genetic mechanisms designed to ensure

Table 3. Summary of results based on an initial population size = 10^{31} prokaryotes; $x_0=0.5$ (half initially lack a duplicate gene). Selectivity factor favouring removal of a duplicate gene, $s=0.0001$. See Fig. 8.⁴⁷

n mutations:	10	12	14	15	16	18	19	20	21	22
Maximum surviving mutants in any generation (a),(b):	1.1E+10	1.0E+6	9.2E+1	8.8E-1	8.5E-3	7.8E-7	7.5E-9	7.3E-11	7.0E-13	6.8E-15
Generation nr. with maximum surviving mutants:	99 015	118 813	138 614	148 515	158 416	178 214	188 106	197 984	207 891	217 652
Total different mutants per MTS (d),(f):	9.0E+14	8.8E+10	8.6E+6	8.5E+4	8.4E+2	8.3E-2	8.2E-4	8.1E-6	7.9E-8	7.8E-10
Plateau for new mutants, generations (g),(h):	240000	250000	280000	290000	310000	330000	350000	360000	380000	390000
Maximum mutants ever produced (c),(e),(i):	3.9E+23	3.7E+19	3.2E+15	3.1E+13	2.8E+11	2.6E+7	2.4E+5	2.3E+3	22	0.21

- (a) Average number of nucleotide mutations assumed / generation: 10^{-9} / nt. Drake estimated³ for prokaryotes about 10^{-10} / nt each generation.
- (b) Natural selection favours smaller genomes, ceteris paribus. Selectivity coefficient, s , to remove unnecessary duplicate genes is about inversely proportional to the number of genes present. Here $s=0.0001$ was assumed.
- (c) All available putative evolutionary time is about 4 billion years. Note that from the origin of life and dramatic increase in complexity far less time would have been available.
- (d) Out of a total prokaryote population of 10^{31} this is the maximum number of individuals calculated to possess m mutations during an MTS. Although organisms with m mutations will increase with generations, t (i.e. more mutations would have occurred), at the same time natural selection is decreasing the proportion which carry an extra duplicate gene. This is why a maximum is reached in the absence of positive selection.
- (e) Eqn. (3) in the main text was used, with an Excel spreadsheet.
- (f) MTS: 'Mutational Time Slice'. Eqn. (3) in the main text was numerically integrated over the number of generations in the MTS. Average total population size assumed: 10^{31} .
- (g) Approximate generation at which virtually no new mutants form with a specific number of mutations, by visual inspection. See Fig. 6 for an example.
- (h) Due to round-off errors, calculations were carried out to only 367,000 generations, which was sufficient, since at this point natural selection would have left but a negligible number of individuals still carrying the duplicate gene.
- (i) Based on 26,000 generations per year (c. 20 minutes average generation time), 4 billion years evolutionary time and the number of MTSs available (which depends on the selectivity coefficient s and number of mutations, m).

the sequences remain homogenous, the opposite of what the gene duplication followed by divergence model requires.

The need for new genes

Even the simplest metabolic networks require several unrelated enzymes. The individual chemical reactions are too slow without such enzymes to be of any value, and until all components are in place, properly regulated, a novel network can't work. A bacterium such as *E. coli* has about 1000 different chemical processes going on concurrently. Evolutionary theory must assume a very large number of unrelated gene families have arisen over time.

There is good evidence that considerable sequence distances separate functional proteins. Axe²⁶ showed that although alternative amino acids may be acceptable at individual positions, total functionality is rapidly lost when multiple residues are mutated. Another study²⁷ confirmed this conclusion. Several studies²⁸ showed that often less than one polypeptide chain out of 10^{50} leads to a functional protein. How realistic is it that a duplicate gene could mu-

Table 4. Sensitivity analysis for x_0 , proportion of population which lack a duplicate gene initially. Initial population size = 10^{31} prokaryotes. Example based on $m=16$ mutations. Selectivity factor favouring removal of a duplicate gene, $s=0.0001$.⁴⁶

X_0 :	0.5	0.9	0.99	0.999
Maximum mutants in a generation:	8.5×10^{-3}	9.4×10^{-4}	8.6×10^{-5}	8.5×10^{-6}
Different mutants from 0.5×10^{31} originally carrying a duplicate gene:	844	93.6	8.5	0.83

tate into something brand new, in light of the calculations presented here?

In the **Methods and results** section, we estimated how many mutations on average could build up before natural selection would eliminate lineages having a duplicate gene. One criterion is the proportion of individuals with highly mutated versions of the duplicate gene *in any generation* during a specific MTS world-wide. Since natural selection quickly eliminates lineages with duplicate genes, we assumed half the world's prokaryote population, 0.5×10^{31} individuals, were initially endowed with a duplicate to provide as many mutational opportunities as possible. For $s=0.001$ about one individual at most with ten mutations would ever exist in any generation (figure 3D), and none with more mutations. Throughout millions of different MTSS this value would not change significantly. Basic statistical principles ensure that repeating an experiment with the same parameters a very large number of times with huge samples is going to result in very similar mathematical outcomes. For example, define an experiment $E =$ 'toss a million fair coins a billion times and record the number of "heads" vs

Table 5. Comparison of initiating with one vs. five duplicate genes. Initial population size = 10^{31} prokaryotes; $x_0=0.5$ (half initially lack duplicate gene(s) free to mutate). Selectivity factor favouring removal of a duplicate gene, $s=0.0001$ / gene. Comparison based on $m=15$ mutations.

Number of duplicate genes	Distinct Nr. of mutants generated in a Time Slot	Generations until plateau in more distinct mutants	Max. Nr. Individuals with 14 mutations in any generation
1	85 278	c. 300 000	0.88
5	17 056	c. 70 000	0.88

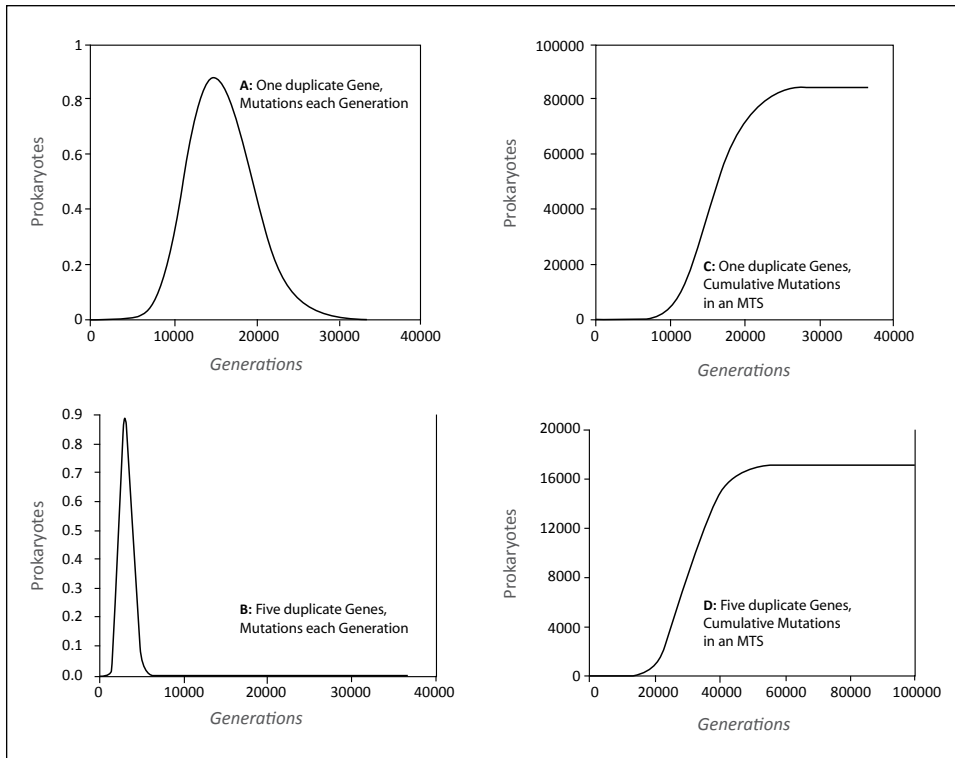


Figure 9. Prokaryotes with 15 mutations as a function of number of initial duplicate genes. Initial population size = 10^{31} , proportion lacking a duplicate gene $x_0 = 0.5$. A: 1 duplicate gene initially ($s = 0.0001$). B: 5 duplicate genes initially ($s = 0.0005$). Y axis: Prokaryotes having $m = 15$ mutations. X axis: Generations.

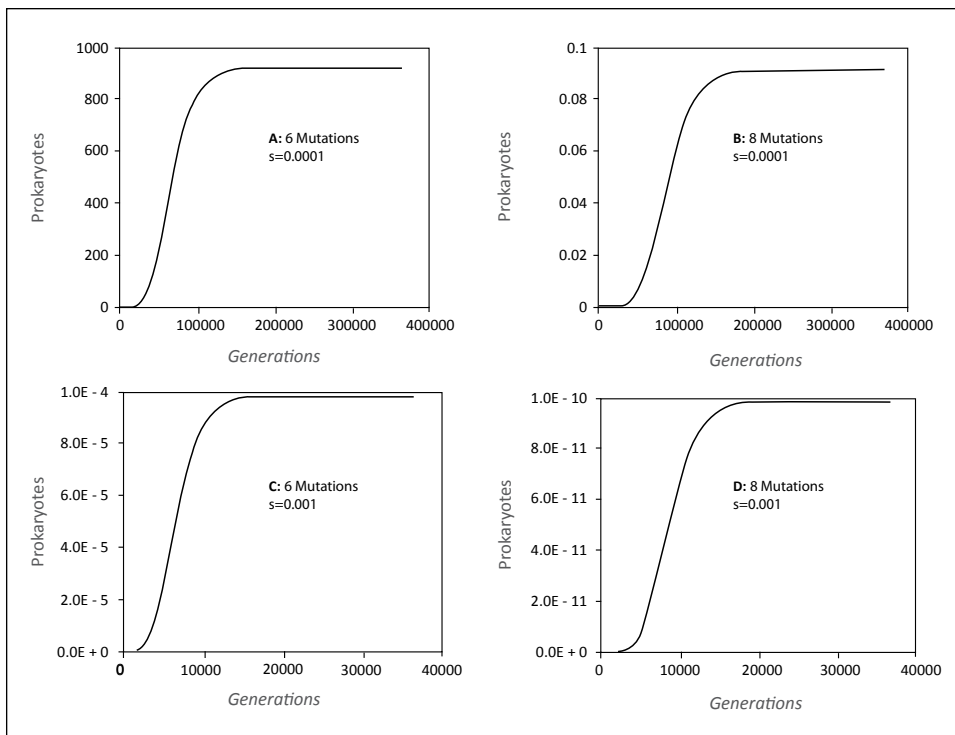


Figure 10. Maximum number of different mutations which could arise from a single duplicate gene event. Initial population size = 10^{11} , proportion lacking a duplicate gene, $x_0 = 0.5$. Y axis: Number of individuals generated during an MTS having m distinct mutations. X axis: Generations.⁴⁸

“tails” obtained’. Repeating E millions of times is unlikely to generate any experiments for which the proportion deviates differently from the statistically expected value of 0.5. The point here is that in the absence of positive selection, duplicate genes with large numbers of mutations will represent but a very small portion of the prokaryote population.

Repeating the analysis with a more generous $s = 0.0001$ shows that at most one organism with a single gene duplicate having 15 or more mutations would be found in any generation of an MTS (figure 5).

A second criterion discussed above involved the *total number* of highly mutated individuals ever generated during an individual MTS. In our calculations we shall assume 26,000 generations per year would be possible for prokaryotes, to maximize the number of MTSs produced. From table 1 ($s = 0.001$), the number of generations consumed by an MTS depends on the number of mutations. In four billion years, taking all available MTS into account, only 34 individuals with 14 mutations would ever be produced and 0.03 prokaryotes with 15 mutations. Far fewer with yet more mutations on the duplicate gene (table 1) would be produced.

The calculations using $s = 0.0001$ (table 3) show that in all evolutionary history twenty one prokaryotes could ever be generated with 21 mutations, and none with 22 or more mutations. We are assuming here that all these mutants, distributed across all MTS, differ from each other, in favour of evolutionary theory. It is important not to overlook, that these few organisms with several mutations on the duplicate gene would not be present at the same time, but

would be distributed among huge populations throughout billions of years.

Once again, the key insight here is that on average a prokaryote carrying a duplicate gene faces a negative selectivity for hundreds of thousands of generations between each random mutation on that duplicate. This fact invalidates evolutionary computer models^{29,30} which attempt to show that a higher development from simple single-cell organisms is inevitable. These have *all* neglected the unavoidable attrition especially during the critical initial two to three billion years when genomes would have been especially small. All extant life forms share similar features (such as DNA replication and gene translation, without which life cannot exist) involving hundreds of unrelated proteins. Therefore, evolutionary theory must claim that all these novel genes arose somehow and fixed across all organisms in their populations.

Although 21 *judiciously* placed gene modifications might be enough to produce something useful, the chances of twenty two individuals (table 3, using $s=0.0001$) stumbling on such a happy coincidence within a search space of 10^{72} alternatives³¹ via *random* mutations is essentially zero. Many of the mutations produced would be ‘wasted’ in the sense of providing chances of discovering a new function. Furthermore, any useful mutations generated must over-compensate for the selective disadvantages of possessing a duplicate, discussed in Part 2, just to break even.

A less favourable selectivity factor, $s=0.001$, which is more realistic (discussed in Part 2), worsens the evolutionary scenario. At most 34 individuals with only 14 mutations on the duplicate gene would be generated, and these would have to cover a search space of 2×10^{49} mutational alternatives³² to find a new gene function.

Searching for needles in a haystack

A key question is how many random mutations (base pair changes, insertions and deletions) would be needed to create a new biological function. What is meant by a new protein based function? Suppose a biochemical step could process very similar molecular isomers. The design of the enzyme variants would be very similar but not 100% identical. To accommodate the slightly different geometries and electronic environment of the transition state, a few amino acids must be modified. Furthermore, the amount of enzyme variants present and when they are expressed may have to be slightly different. These are but trivial differences in function, predicted *a priori* by creationists to be present on the genomes if needed, and do not present evidence for evolutionary origin of truly new and unrelated protein functions. To illustrate, different identification numbers painted on two otherwise identical Boeing 747s render them non-identical, but clearly these models can be considered for all practical purposes the same and qualitatively different from the design of helicopters. The insignificant differences between these two airplanes cannot be extrapolated to say anything about helicopters.

Different ecological niches can require some proteins to remain folded more or less stably, for example in a hot temperature environment. This would require different amino acids at some positions. Optimal design, or built-in processes to permit such adjustments over time, is predicted by Design theorists and the existence of such variants is not evidence for macroevolutionary improvements.

How many mutations must occur to generate a truly new protein? In one study,³³ domain functions were classified according to the FLY + ENZYME scheme. For both enzymes and non-enzymes, 50% or higher sequence identity between pairs compared almost always resulted in the same or very similar functions.³⁴

In a second study,³⁵ ORFs (open reading frames) function classifications of yeast and *E. coli*, based on the MIPS and GenProtEC schemes, were compared. Over 90% of the pairs having sequences at least 50% identical provided an identical or very similar function, and for 90% and higher sequence identity essentially all the pairs were considered to have identical or very similar function.

A considerable amount of sequence divergence must occur for a new function to arise. The authors conclude that ORFs with 30% sequence identity and a reported e-value of 0.001 structure match in the PDB database have a two-thirds chance of having the same exact function. They point out that random mutations of 70% of a protein’s amino acids will rarely generate a stable new protein fold.

In a third study,³⁶ a set of 904 single domain *E. coli* proteins were collected from the NCBI site.³⁷ Their EC classification was extracted from the SwissProt database.³⁸ The authors showed that for 40% or higher sequence identity over 60% of those proteins display identical EC classifications. Unfortunately they do not report the percent sequence identity above which no exceptions for classification identity was found.

The authors do point out³⁹ that there is often ambiguity in how researchers annotate their data. Enzymes with the same cofactors or substrates may often have completely different functional classifications, and very similar enzyme subunits may occur in different complexes since they are parts of different molecular machinery. Although three *E. coli* DNA polymerases have an identical E.C. classification 6.4.1.2 they are sequentially very different and each provides unique functions.

In a fourth key study⁴⁰ Devos and Valencia prepared a dataset from the FSSP database⁴¹ using all proteins for which between 75 and 100% of the length of both sequences were aligned. Members with > 95% sequence identity were considered to be exactly the same proteins and not even included in their 2338 member dataset. At 80% and higher sequence identity all the EC classifications were completely identical. Note, however, that this 80% figure may well need to also include up to 25% of the residues ignored in the optimal alignment.

Evolutionary processes would also need to create brand new protein folding patterns. As a rule, a few mutations will not simply lead to a new, stable folded pattern. The

same dataset revealed that essentially all pairs studied have identical FSSP structural classification with up to 65% non-identical residues.

The data in these four studies provide valuable insight as to the number of amino acid differences needed to convert one protein into a totally new one. They and the comments above referring to great sequence differences separating functional proteins²⁸ lead us to offer the following claim:

‘If up to 10% of the residues of a protein are randomly mutated, over 99.9% of the time a new, non-trivially different function will not be generated.’

(Insertion or deletion of a residue is also permitted in this claim). Implied here is a new function which does not also require additional genes.

Modifying 10% or more of the amino acids of an average sized protein (approximately 300 amino acids), or about 30 residues, via mutations on a single duplicate gene is beyond the maximum of 21 which could be generated since evolutionists claim life began on earth. The parameters to establish this upper limit of 21 reflected extreme values with some semblance of realism to provide maximum variability and number of prokaryote mutants.

After extensive genome comparison of gamma proteobacteria, by far the largest grouping of bacteria, researchers concluded that ‘Gene duplication has contributed relatively little to the contents of these genomes; instead, LGT, over time, provides most of the diversity in genomic repertoires.’⁹ This is certainly consistent with the finding of the present study. But evolutionary theory requires a vast number of new protein-coding genes, with no discernable sequence similarities, which must come from somewhere. Mere transfer of an already functional gene from other organisms fails to explain its ultimate origin.

Other scenarios

Careful thought was given to other scenarios an evolutionist may propose, and none showed any promise. It makes no sense to argue that much smaller genes, or only a portion of one, needs to be modified to generate a new function, since the number of potential mutational targets shrinks proportionally (1000 nt were generally used in the scenarios discussed here). For example, obtaining just the right six mutations in a limited portion of a duplicate gene (such as the part coding for an active site of an enzyme) means we can no longer assume any of one thousand base pairs on the gene are candidates for mutation.

Another line of reasoning consists of step-wise improvements. Perhaps a small number of mutations would offer a selective advantage, that strain would multiply, and then another set of mutations may occur precisely on the modified new gene. The fact is, however, that most prokaryote genes are singletons, meaning no others of similar sequence are present on the same genome. But in addition, we showed above what happens when far less than 10^{31} candidate organisms are the target of random mutations: far fewer mutations are able to accumulate before these mutated

lineages go extinct. Therefore, arriving in such a bootstrap manner at more than 30 strategically placed mutations to produce a truly new function by multiple discrete steps is unrealistic.

Conclusions

Various evolutionary scenarios were examined by varying parameters such as prokaryote population size, mutational rate, generation times, proportion of population with additional genes, number of duplicate genes and selectivity coefficient favouring genome truncation. Assuming mutations on a duplicate are harmless would permit these to accumulate, but in reality natural selection would systematically remove the descendants of duplication events, drastically limiting both the total number and variety of mutants. Duplicate genes would be created, accumulate at most a very small number of mutations, and then go extinct, again and again. The number of distinct mutational variants generated would be far too small to explain the origin of novel cellular functions.

All scenarios using prokaryote populations failed to generate enough mutation to produce novel genes. The most promising approach assumes huge populations would be involved, although subsequently surviving and fixing would now become exceedingly unlikely.

Preventing novel gene families from developing denies nature the necessary infrastructure to produce complex new features. This finding contradicts what is being claimed by evolutionary biologists, which therefore invites other explanations as to the source of genetic complexity to be considered.

References

1. Some genes code only for useful RNA which is not translated into proteins.
2. Ohno, S., *Evolution by Gene Duplication*, Springer-Verlag, Berlin, 1970.
3. Gogarten, J.P. and Townsend, J.P., Horizontal gene transfer, genome innovation and evolution, *Nature Reviews Microbiology* **3**:679–687, 2005.
4. Jianzhi Zhang, J., Evolution by gene duplication: an update, *Trends in Ecology and Evolution* **18**(6):292–298, 2003.
5. Hooper, S.D. and Berg, O.G., On the nature of gene innovation: duplication patterns in microbial genomes, *Mol. Biol. Evol.* **20**(6):945–954, 2003.
6. McAdams, H.H., Srinivasan, B. and Arkin, A.P., The evolution of genetic regulatory systems in bacteria, *Nature Reviews Genetics* **5**:169–178, 2004.
7. Kimura, M., Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics, *Proc. Natl. Acad. Sci. USA* **88**:5969–5973, 1991.
8. Ochman, H.O., Lawrence, J.G. and Groisman, E.A., Lateral gene transfer and the nature of bacterial innovation, *Nature* **405**:299–304, 2000.
9. Lerat, E., Daubin, V., Ochman, H. and Moran N.A., Evolutionary origins of genomic repertoires in bacteria, *PLoS Biology* **3**(5):807–814, 2005.
10. Thomas, C.M. and Nielsen, K.M., mechanisms of, and barriers to, horizontal gene transfer between bacteria, *Nature Reviews Microbiology* **3**:711–721, 2005.

11. McAdams, H.H., Srinivasan, B. and Arkin, A.P., The evolution of genetic regulatory systems in bacteria, *Nature Reviews Genetics* **5**:169–178, 2004.
12. Hamoen, L.W., Venema, G. and Kuipers, O.P., Controlling competence in *Bacillus subtilis*: shared use of regulators, *Microbiology* **149**:9–17, 2003.
13. Discussions have been held in Europe and the USA (Drs. T. Wood, R. Truman, P. Borger, K. Andersen, J. Sanford and many others who wish to remain anonymous) to exchange viewpoints on this matter. It is possible that gene duplication may have been a designed feature properly regulated initially to respond to environmental signals. The current processes may reflect primarily degrading genomes.
14. Rocha, E.P.C., The replication-related organization of bacterial genomes, *Microbiology* **150**:1609–1627, 2004.
15. Hedges, S.B., Blair, J.E., Venturi, M.L. and Shoe, J.L., A molecular timescale of eukaryote evolution and the rise of complex multicellular life, *BMC Evolutionary Biology* **4**:1–9 2004.
16. Hoyle, F., *Mathematics of Evolution*, Acorn Enterprises LLC, Memphis, 1999. See equation 1.6 on p. 11.
17. Wagner, A., Energy constraints on the evolution of gene expression, *Mol. Biol. Evol.* **22**(6):1365–1374, 2005.
18. Truman, R. and Borger, P., Genome truncation vs mutational opportunity: can new genes arise via gene duplication? Part 1, *Journal of Creation* **22**(1):111–119, 2008.
19. Drake, J.W., Charlesworth, B., Charlesworth, D. and Crow, J.F., Rates of spontaneous mutation, *Genetics*, **148**:1667–1686, 1998. Estimated base pair errors per replication: for *Escherichia coli* 5.4×10^{-10} ; for *Neurospora crassa* 7.2×10^{-11} . See p. 1670.
20. Most prokaryotes live in the oceans, which contain about 1.4×10^{21} litres of water ('Earth', <seas.lpl.arizona.edu/nineplanets/nineplanets/earth.html>. Evolutionists believe the oceans were originally much smaller which would imply smaller prokaryote population sizes). About 10^{11} prokaryotes would be present per litre, which is about a tenth of the laboratory densities under optimal conditions.
21. The oceans are believed to have been considerably smaller in the past, meaning the average population sized assumed, 10^{31} prokaryotes, is probably exaggerated in favour of an evolutionary scenario.
22. Several Excel spreadsheets are available on the Internet, see references 42–48.
23. Generated with spreadsheet: <www.creationontheweb.com/gen trunc1_ref23>. Figs_S=0.0001.xls. The seven spreadsheets used for the calculations referred to in this paper have been made available online at this location: <www.creationontheweb.com/gen trunc1_supp>.
24. Albert, B. et al., *Molecular Biology of the Cell*, 3rd Edition, Garland Publishing, USA, p. 386, 1994.
25. Truman, R., The ubiquitin protein: chance or design? *Journal of Creation* **19**:116–127, 2005. Often many copies of this protein were present in tandem on the same genome, with no or insignificant variability. Furthermore, it is possible that the few non-identical sequences are not actually functional.
26. Axe, D.D., Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds, *J. Mol. Biol.* **341**:1295–1315, 2004.
27. Truman, R., Protein mutational context dependence: a challenge to neo-Darwinian theory: part 1, *Journal of Creation* **17**:117–127, 2003.
28. Truman, R., Searching for needles in a haystack, *Journal of Creation* **20**(2):90–99, 2006.
29. Lenski, R.E., Ofria, C., Pennock, R.T. and Adami, C., The evolutionary origin of complex features, *Nature* **423**:139–144, 2003; <myxo.css.msu.edu/papers/nature2003/>.
30. Truman, R., Evaluation of neo-Darwinian theory using the Avida Platform, two parts, *Progress in Complexity, Intelligence and Design* **3**(1), <www.iscid.org/pcid/2004/3/1/truman_avida_evaluation.php>.
31. $19^{19} \times 333 \times 332 \times \dots \times 315 = 9.9 \times 10^{71}$ alternatives. The logic is: an average protein has about 333 amino acids. Any position could undergo a mutation to any of 19 other amino acids, leaving 332 other positions for the next mutation.
32. $19^{13} \times 333 \times 332 \times \dots \times 321 = 2.05 \times 10^{49}$ alternatives.
33. Wilson, C.A., Kreychman, J. and Gerstein, M., Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, *J. Mol. Biol.* **297**:233–249, 2000.
34. Wilson et al., ref. 33, see figure 7A.
35. Wilson et al., ref. 33, see figure 7D.
36. Pawlowski, K., Jaroszewski, L., Rychlewski, L. and Godzik, A., Sensitive sequence comparison as protein function predictor, *Pac. Symp. Biocomput.* **8**:42–53, 2000.
37. <www.ncbi.nlm.nih.gov/Entrez/>
38. <www.expasy.ch/sprot/>
39. Pawlowski et al., ref. 36, p. 50.
40. Devos, D. and Valencia, A., Practical limits of function prediction, *Proteins* **41**:98–107, 2000.
41. Holm, L. and Sander, C., Mapping the protein universe, *Science* **273**:595–602, 1996.
42. <www.creationontheweb.com/gen trunc1_ref42>.
43. <www.creationontheweb.com/gen trunc1_ref43>.
44. <www.creationontheweb.com/gen trunc1_ref44>.
45. <www.creationontheweb.com/gen trunc1_ref45>.
46. <www.creationontheweb.com/gen trunc1_ref46>.
47. <www.creationontheweb.com/gen trunc1_ref47>.
48. <www.creationontheweb.com/gen trunc1_ref48>.

Royal Truman has bachelor's degrees in chemistry and in computer science from SUNY Buffalo, an M.B.A from the University of Michigan, a Ph.D. in organic chemistry from Michigan State University and post-graduate studies in bioinformatics from the universities of Heidelberg and Mannheim. He works for a large multinational in Europe.

Peter Borger has an M.Sc. in Biology (Hons. biochemistry and molecular genetics) and a Ph.D. in Medical Sciences from the University of Groningen, The Netherlands. He is currently working on the cellular and molecular aspects of pulmonary diseases, such as asthma and COPD, and is an expert on the molecular biology of signal transduction and gene expression.
