Genome truncation vs mutational opportunity: can new genes arise via gene duplication?—Part 2

Royal Truman and Peter Terborg

In 1970, Susumo Ohno proposed gene and genome duplications as the principal forces that drove the increasing complexity during the evolution from microbes to microbiologists.[1] Today, evolutionists assume duplication followed by neo-functionalization is the major source of new genes. Since life is claimed to have started simple and evolved new functions, we examined mathematically the expected fate of duplicate genes. For prokaryotes, we conclude that carrying an expressed duplicate gene of no immediate value will be on average measurably deleterious, preventing such strains from retaining a duplicate long enough to accumulate a large number of mutations. This genome streamlining effect denies evolutionary theory the multitude of necessary new genes needed. The mathematical model to simulate this process is described here.

In Part 1,[2] we examined critically the evolutionary claim that gene duplication and subsequent divergence could account for the origin of a large number of new genes. Without many new genes an original, primitive life form cannot evolve new, complex biological functions. We pointed out that possessing an extra superfluous duplicate gene is selectively disadvantageous. These strains would be out-reproduced by their streamlined competitors and go extinct before being able to produce new biochemical functions.

Streamlining is especially effective for prokaryotes with smaller genomes, such as 2,000 or fewer genes. This means evolutionists face a serious difficulty in explaining a necessary large increase in genome size during the first two or three billion years of their theory. One can compare gene families of present bacteria and archaea and find tens of thousands of examples lacking any sequence homology at all. Where did these all come from? Since many fundamental processes are shared by known living organisms, the evolutionist must claim these evolved long ago. And many of these universal genes are sequentially unrelated.

In Part 1, we examined several scenarios.[2] The results were unexpected in the extreme. Even with generous assumptions not even a single prokaryote would accumulate *22* or more *mutations* on a single duplicate gene. More realistic assumptions suggested the maximum for any prokaryote may actually be only *15* or so mutations in a single gene. The calculations were based on average generation times of about twenty minutes. To put matters into perspective, we also explored the maximum number of *organisms* which would have possessed several mutations *at any particular generation*. Using huge populations of 10^{31} individuals, the most generous assumption scenario indicates that about 1.1×10^{10} members with ten mutations would have been present in at any time. This is a negligible proportion of about 1 out of 10^{21}, which natural selection would eliminate before additional mutations can occur, given their selective disadvantage. With more realistic assumptions the odds are far worse: only about one individual with ten mutations would have co-existed at any time.

Even if the quantitative results in Part 1[2] could be modified by factors of billions, the evolutionary claims could not conceivably be true. The number of trials needed to produce the number and variety of genes found throughout nature is vastly too great.

Methods and results

First, we will examine here the reasoning used in developing the model. One purpose is to critically evaluate the robustness of the claims and to permit the reader to effectively argue this case himself. The mathematical tools also permit examination of microevolutionary changes, those leading to slight modifications, whether positive or deleterious, of the same basic biological function.

The key parameters needed in Part 1 include: (i) the mathematical function to describe of number of mutations per generation; (ii) the number of Mutational Time Slices (*MTSs*) which could be produced; (iii) the total number of organisms, x, produced with m mutations; (iv) the rate of duplicate gene formation; (v) the rate of nucleotide mutation; and (vi) the selectivity factor, s, favouring streamlined genomes.

(i) Mathematical function to describe mutations per generation

Each nucleotide (nt) of a duplicate gene could mutate before forming part of the offspring cell. How many organisms would have $m = 0, 1, 2 \ldots$ mutations after several generations? This can be modelled using a binomial probability distribution, discussed in almost any general textbook on statistics. The general case is formulated as follows: suppose two distinct outcomes are possible, 'success' with probability p and 'failure' with probability $1 - p$. Since only these outcomes are possible, their sum is

one. A number n of trials will be carried out. What is the probability of obtaining m successes, $p(m)$, assuming each trial is independent of the preceding? This distribution is described by the binomial probability (1) distribution, shown below.

$$p(m) = \frac{n!}{m!(n-m)!} p^m q^{n-m} \quad (1)$$

where p = probability of a success per trial; $q = 1 - p$; n = number of trials; m = number of successes after n trials.

For example, what are the probabilities of all possible outcomes from tossing a fair coin five times: m can be 0 … 5 'heads', and the expected $p(m)$ values are displayed graphically in figure 1. (We use 'm' since 'success' will refer in this paper to a *m*utation occurring).

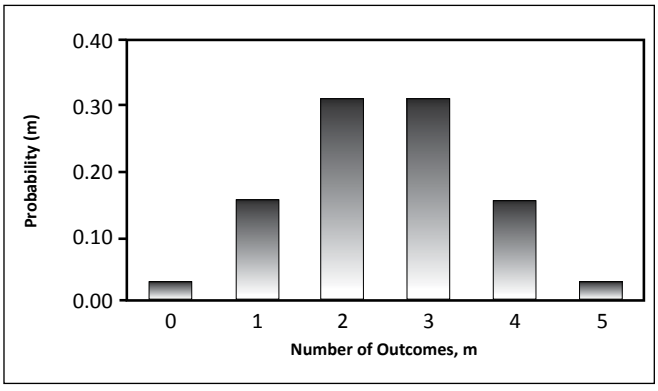

Figure 1. Example of a binomial distribution. Shown here: probability of success $p = 0.5$, outcome from five repeated trials. Possible outcomes: $m = 0$ to 5 successes. Sum of all $p(m)$ must be one. Expected or average outcome is sum of all $(m \times p(m)) = 2.5$.

In the case of nucleotide (nt) single base mutations, there are two possible outcomes, 'mutated' with probability p or 'not mutated'. We assume the probabilities are independent. Each of n positions along the duplicate gene can mutate. For an average size gene this involves about one thousand nt. The probabilities $p(m)$ now refer to the total number of mutated positions after having replicated the duplicate gene during a generation.

Are the probabilities of nt mutation really independent of previous history and context? Not entirely. If an nt has mutated already, a future mutation could revert the preceding change. In a sense mutations would be 'wasted'. Evolutionary theory requires as many trial-and-error attempts as possible, so to calculate upper bounds we neglect this factor and pretend each mutation must always generate something new. In addition, the immediate physical-chemical context of an nt can affect both the probability and nature of its mutation. This means that nature is in reality restricted in its ability to explore the space of possibilities in an attempt to discover a useful combination of mutations.

For computational convenience one can calculate $p(m)$ over intervals such as every 100 or 1,000 generations. The number of binary trials, n, is then the number of nt in the duplicate gene multiplied by the number of generations. For high selectivity values, such as $s = 0.001$ favouring smaller genomes, strains with duplicate genes go extinct rapidly. In this case fewer computations are needed and shorter intervals of 100 generations were used, and *vice-versa*.

(ii) Mutational Time Slices (MTS)

Throughout nature duplicate genes would arise and go extinct continuously. Of course one cannot model countless individual fates. To simplify but still model realistically, one can work with averages. We assumed in Part 1[2] that initially half the world's prokaryotes, x_0, would initially possess a fresh duplicate gene, not necessarily the very same gene. Given the selective advantage enjoyed by streamlined competitors, on average the proportion with a duplicate decreases and eventually dies out, unless a fortunate ensemble of mutations on the duplicate were to provide a significant selective advantage.

We examine the proportion having some number m of mutations, typically more than five, because fewer are not going to generate a new biochemical function. Since the probability of an nt mutation is exceedingly small, using selectivity coefficient values s between 0.001 and 0.0001 revealed that eventually plateaus are reached at which virtually no more mutations will accumulate. There are two reasons for this. First of all, the number of prokaryotes carrying a duplicate gene decreases rapidly due to natural selection. Secondly, the binomial distributions disfavour the extreme cases far from the expected value, see figure 1. The chances of additional mutations decreases with the number m already present since these are so rare. The odds for individuals with e.g. $m = 18$ mutations on the duplicate to receive yet another mutation on that gene are far lower than for the much larger proportion of individuals which have accumulated e.g. m = two mutations.

Incidentally, in these kinds of statistical studies one needs to check whether the software is producing rounding-off errors. Microsoft Excel and Open Office Calc spreadsheets, used in this study, generate round-off errors after about 99.9999999999999% of the population no longer possesses a duplicate gene. This is not a problem here, however, since additional mutations in one of these rare survivors are statistically very unlikely. Furthermore, the chance of one of these actually fixing in such a huge population (10^{31} members) would be negligible.

Plots of number of individuals with highly mutated duplicate genes vs generation were examined in Part 1.[2] The generation number where it became obvious that an asymptote had been reached (i.e. at which additional individuals with that number m of mutations are not being generated) was identified. The generations from origin of duplicate to negligible further increase in number of mutated individuals defined an MTS ('Mutational Time Slice'). The next MTS begins with half the population again endowed with a duplicate gene.

Table 1. Example to illustrate numerical results using equations (1), (2) and (3) in the main text. Parameters used: number of mutations, $m = 19$; probability of mutation = 10^{-9}; nucleotides in a duplicate gene able to mutate = 1,000; initial population with a duplicate gene, $x_0 = 5 \times 10^{30}$.

(a)	(b)	(c)	(d)	(e)
0	50000000000.000E+30			
500	4.875E+30	7.637E-50	9.75E-42	9.751E-42
1000	4.750E+30	3.900E-44	2.11E-38	2.106E-38
1500	4.626E+30	8.416E-41	4.86E-36	4.883E-36
2000	4.502E+30	1.936E-38	3.31E-34	3.362E-34
2500	4.378E+30	1.306E-36	1.05E-32	1.080E-32
3000	4.256E+30	4.054E-35	1.94E-31	2.050E-31
3500	4.134E+30	7.362E-34	2.44E-30	2.647E-30

(a) Number of generations, *t*.
(b) Organisms retaining a duplicate gene after *t* generations of natural selection. Based on eqn. (1)
(c) Organisms with *m* mutations at end of generation number *t*, based on eqn. (2), times factor (b).
(d) Numerical integration: organisms having *m* mutations generated in current generation interval (MTS).
(e) Cumulative number of organisms with *m* mutations after *t* generations.

This analysis takes into account the rareness of gene duplication, but the rate at which this occurs is not known. In a much cited study Lynch and Conery estimated[3] gene duplication rates to range from about 0.02 down to 0.002 per gene per million years, depending on the species. Baker's yeast, a single cell organism, was included in their analysis. This is commented on later.

Selectivity factors, *s*, on the order of 0.001 to 0.0001 rapidly cause disappearance of inefficient prokaryote genomes carrying redundant genes. From table 1, an MTS lasts at most 360,000 generations, and in Part 1[2] we assumed very short generation times, 26,000 per year, to produce as many highly mutated organisms as possible. Therefore, all MTSs used lasted less than 14 years. Assuming Lynch and Conery's average estimate has some relevance to other single cell organisms in addition to baker's yeast would suggest about 20 new duplicate genes for a small prokaryote (< 2000 genes) per organism would be produced in a million years. This rate is far too low to replenish each MTS with half the population carrying a new duplicate gene.

Assuming half of all prokaryotes will be graced with a new duplicate gene in < 14 years is very generous. Depending on the scenario modelled far shorter time spans were usually used.

(iii) Total number of organisms with m *mutations produced during an MTS*

Natural selection will favour streamlined genomes. The fraction of individuals, *f*, which carry a duplicate gene after a specific number of generations, can be calculated using an equation developed by Hoyle[4] (2):

$$f = \frac{x_0 e^{st}}{1 + x_0\left(e^{st} - 1\right)} \qquad (2)$$

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 here refers to number of generations; $x_0 = x$ at $t = 0$

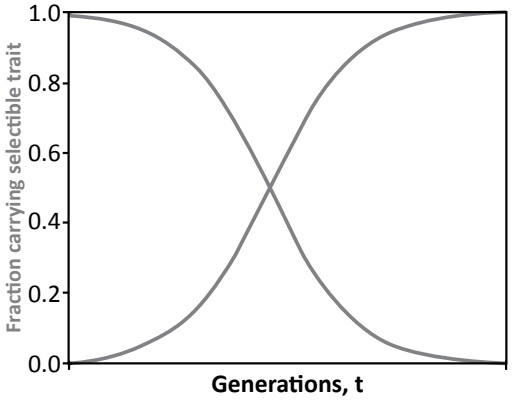

Figure 2. Build-up of a trait by natural selection for fixed population size. Fission (non-sexual) reproduction, using Hoyle's equation (2). For every generation the sum of both curves must add up to one.

In figure 2 we show that *s* can be positive or negative. The figure also illustrates how increases in the number of those individuals possessing a trait come at the expense of those lacking it, and *vice-versa*. If this were not the case, Hoyle's mathematical functional would be incorrect: since there are only two possibilities, both proportions must always add up to unity.

Eq. (1) permits calculation of the distribution of number of mutations which occur on a duplicate gene after *t* generations. The fraction *f* of individuals carrying a duplicate gene after *t* generations is given by eq. (2). We assume a huge population of 10^{31} prokaryotes on average. Combining these three factors leads to function $O_{(m,t)}$:

$$O_{(m,t)} = 10^{31} \frac{(1000t)!}{m!(1000t - m)!} p^m q^{(1000t-m)} \frac{x_0 e^{st}}{1 + x_0\left(e^{st} - 1\right)} \qquad (3)$$

where t = time in number of generations; s = selectivity favouring duplicate gene elimination, typically 0.001 or 0.0001; p = probability of an nt mutation; $q = 1 - p$; $1000t = n$ = number of mutational trials, assuming 1000 nt in a duplicate gene; m = number of mutations.

Figure 3 shows an example of $O_{(m,t)}$ using $m = 14$. Integration over variable t from 0 to the number of generations in an MTS gives the total number of individuals during this interval which possessed m mutations. The mathematically inclined can use table 2 to check the understanding of equations (1)–(3).

Numerical integration was performed with Excel spreadsheets. The number of generations in an MTS depends on the number of mutations, *m*, being evaluated. These generations are divided into intervals of 50 generations (for $s = 0.001$) or 500 generations (for $s = 0.0001$). In solving

Table 2. Summary of predictions based on mathematical models discussed in the main text.

Mutations, m:	19		20				21		22	
Mutation rate / nt each generation (a):	1E-8	1E-9	1E-8	1E-8	1E-9	1E-9	1E-9	1E-9	1E-9	1E-9
Selectivity factor, s (b):	0.0001	*0.0001*	0.0001	0.001	*0.0001*	0.001	*0.0001*	0.001	*0.0001*	0.001
Years available (c):	4E+9	4E+9	4E+9	4E+9	4E+9	4E+9	4E+9	4E+9	4E+9	4E+9
Maximum having m mutations in any generation (d):	1.5E+10	*7.5E-09*	1.3E+09	7.3E-11	*7.3E-11*	8.7E-31	*7.0E-13*	8.5E-34	*6.8E-15*	8.3E-37
Generation t with maximum surviving mutants (e):	172 727	188,106	181,819	19,802	197,984	19,981	207,921	20,981	217,601	21,976
Total different mutants per MTS (f):	1.48E+15	*8.16E-04*	1.35E+14	8.05E-07	*8.05E-06*	9.8E-27	*8.01E-08*	9.6E-30	*7.91E-10*	9.5E-33
Plateau for new mutants, generations (g), (h):	350,000	350,000	360,000	36,000	360,000	36,000	370,000	37,000	380,000	38,000
Maximum mutants ever produced (i):	4.40E+23	*242,469*	3.90E+22	2.33E+3	*2,326*	2.8E-17	*22.5*	2.7E-20	*0.22*	2.6E-23

(a) Average number of nucleotide mutations between parent and daughter each generation. Drake estimated[3] for prokaryotes about 10^{-10} /nt each generation.
(b) Natural selection favours smaller genomes, ceteris paribus. Selectivity coefficient, *s*, to remove unnecesssary duplicate genes is about inversely proportional to the number of genes present.
(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) Eq. (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 geneneration at which virtually no new mutants form with a specific number of mutations, by visual inspeaction. See figure 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*).

(3), the value of *t* at the start and end of each interval was used, and the average value was multiplied by the size of the generational interval (50 or 500 generations). The principle is shown in figure 4.

To give the evolutionary model the maximum chances of working, we shall assume that all mutations generated during all MTSs would be different.

(iv) Rate of duplicate gene formation

Lynch and Conery suggested[3] for eukaryotes an average rate of origin of new gene duplicates on the order of 0.01 per gene per million years, or 1 gene per hundred million years. This was an average value based on organisms which includes single-cell yeast, but nevertheless only eukaryotes. The estimates relied on phylogenetic evolutionary assumptions, e.g. that several similar genes could not have been present *ab initio*. In other words, alternate sequences are assumed to only have arisen from an initially identical gene, a premise we do not accept.

From an evolutionary theoretic framework, a reasonable expectation would be that at some point starting two to three billion years ago all genomes possessed less than two thousand genes. Therefore, duplication events would only occur about once every ten thousand to hundred thousand years per organism, instead of only about ten years as used in calculating MTS times in Part 1. One must not overlook that for duplicates to be able to evolve into a new function we are only interested in those which are expressed (e.g.

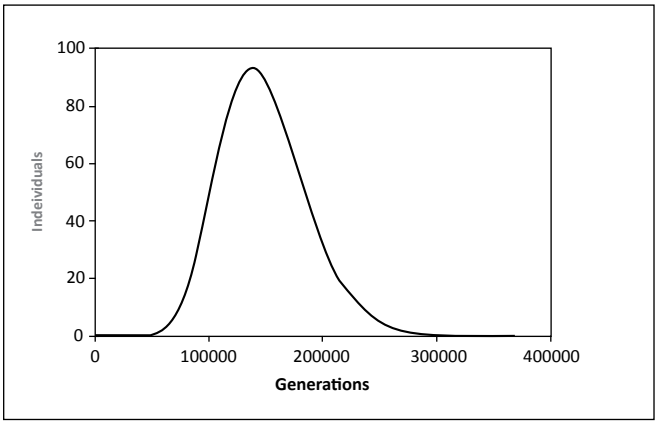

Figure 3. Individuals with $m = 14$ mutations after t generations. Selectivity factor $s = 0.0001$, initial population size 10^{31}, with half initially possessing a duplicate gene.

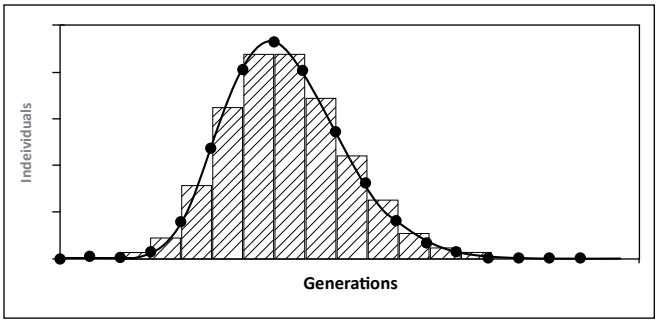

Figure 4. Principle of numerical integration. The x axis is divided into discrete interval widths. The height of each is the average of the y value at the beginning and end of each interval. All the areas are then summed.

in a suitable promoter environment); which don't destroy other genes; that result in protein stoichiometries which are acceptable; and for which mutations will not destroy the expression during the large number of future generations as mutations begin to accumulate. For very compact prokaryote chromosomes functionally harmless gene duplications are surely the exception!

These estimates persuade us that we have been generous in our model settings. Very short MTS times were used in which, however, half the population would actually not be able to replenish with new duplicate genes lost by natural selection.

It is questionable whether vastly higher gene duplication rates would favour an evolutionary model. Such poor chromosome replication would add a major stochastic factor in determining who survives. In most cases this process is deadly. Suppose an advantageous mutation on a duplicate gene were to occur sometime. If this locus or any other part of the genome were to undergo constant genetic insults in the many subsequent generations of this new strain, then survival would be primarily a random effect. An evolutionist can not assume that a constant, reliable positive selection would be available every generation (which would facilitate subsequent fixation in the population) since a random influx of predominantly 'bad' mutations would dominate the actual outcome.

(v) Rate of nucleotide mutation

Drake has estimated[5] nt errors per replication for various prokaryotes. For *Escherichia coli* he reported a value of 5.4×10^{-10} and for *Neurospora crassa* 7.2×10^{-11}. To provide evolutionary models with the best chance of succeeding, a higher rate of 10^{-9} was used in Part 1 of this study.[2] We are also ignoring the protection provided through redundancy in the genetic code. Many nt mutations will actually not end up coding for a different amino acid and therefore not be able to evolve into a new function.

It would take on average about a million generations per organism (10^{-9} mutations/nt $\times$ 10^3 nt/gene) to produce a single mutation on each average sized duplicate gene. Clearly, producing variants with multiple mutations in just a few generations is highly unlikely.

(vi) Selectivity factor, s, favouring genome truncation

Loss of unnecessary genetic material by prokaryotes is selectively advantageous. Whether the extra gene arose from Horizontal Gene Transfer (HGT) or chromosomal gene duplication is irrelevant, as long as the duplicate is passed on consistently to the progeny. 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 contribution to genome streamlining, a further difficulty in the evolutionary model. We are only considering competition between original pristine genomes lacking a duplicate gene vs the new mutants.

Natural selection will disfavour lineages with larger genomes *ceteris paribus*: (1) there is a significant metabolic cost and (2) the generation times will be longer. One would intuitively expect a value of s for prokaryote lineages unburdened by an unnecessary duplicate gene to be about the reciprocal of the number of chromosomal genes. There is now scientific evidence to support our view. Quantitative evidence supporting this expectation has been offered in two Appendices.

Recent studies showed that the metabolic costs for most genes of the single cell eukaryotic microbe *Saccharomyces cerevisiae* and an estimated total amount of energy per second generated were published in 2005.[6] These data provide (Appendix 1) a basis to estimate the penalty of carrying an unnecessary, extra, expressed duplicate gene for single-celled organisms. In addition, the effect of longer chromosome replication time is examined in Appendix 2. Both independent factors favour lineages with streamlined genomes. For bacteria having a few thousand genes the conclusion is that s will be somewhere between 10^{-4} and 10^{-3} per generation. The smaller the genome the stronger the truncating effect would be. This is particularly problematic for evolutionary theory which assumes that, for hundreds of millions of years, only primitive organisms with but a few hundred genes would have lived. Any increase in genetic material would be proportionally most detrimental for these small genomes. (The quantitative evidence supporting this expectation is presented in Appendix 1 and 2).

Discussion

Although it is generally assumed that gene duplication and adoption is an evolutionary mechanism for genomes to increase complexity/information, our model demonstrates the opposite: a rapid loss of duplicates is inevitable for small, simpler organisms.

In these kinds of models one must be careful to develop internally consistent scenarios. Parameter tinkering must then take all impacted effects into account. One may wish to provide more mutational opportunities on a duplicate gene by assuming very small genomes and thereby a greater

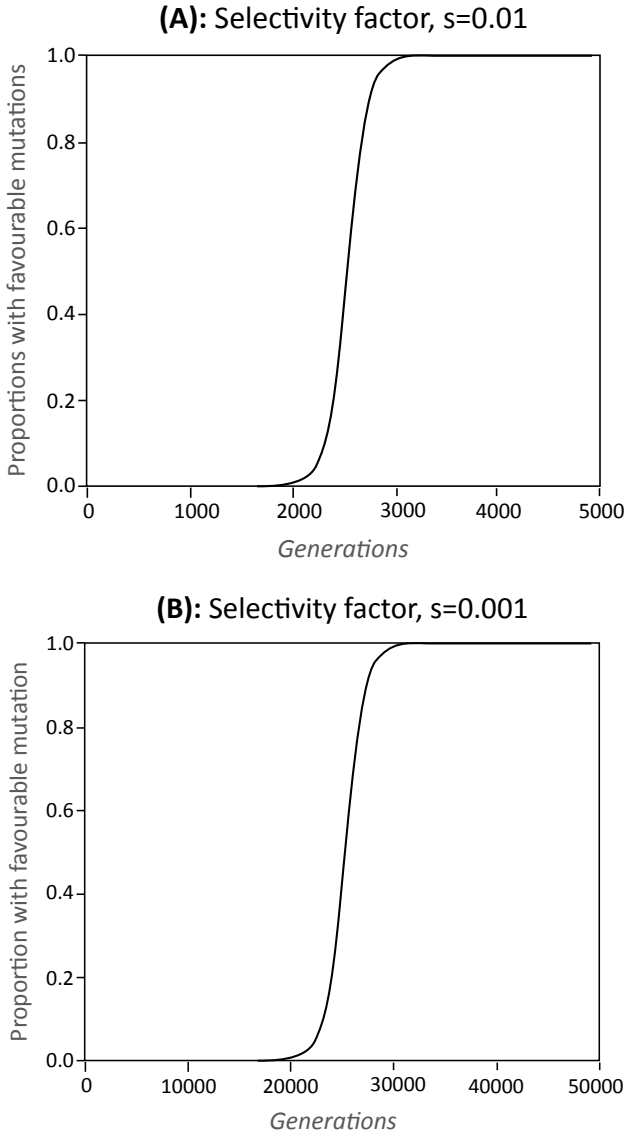

Figure 5. Buildup of proportion with selectively favourable trait is initally very slow. Small population size assumed, $x_0 = 10^{11}$ organisms, fission form of reproduction.
(A) High selectivity factor assumed, $s = 0.01$
(B) Lower selectivity factor assumed, $s = 0.001$

number of MTSs, since generation times would be shorter. This would mean, however, that the relative disadvantage of carrying a duplicate gene would increase, and s >> 0.0001 *must* now be used. This would *decrease* the number of mutations which could accumulate before the descendants of a gene duplication event go extinct.

An evolutionist may wish to assume much higher mutation rates. Let us assume the new mutant strain is eventually produced, and enjoys a net selective advantage (which compensates for the disadvantages of carrying an extra expressed gene) of between 0.1% and 1%. For some 2,000 to 20,000 generations this strain would remain a negligible component of even a tiny population of 10^{11} members (figure 5). Consider a mutation rate of 10^{-7}/nt each generation for a smallish genome of 2,000 genes (c. two million nt). Long before such strains could fix each individual must survive some 200 to 2,000 random mutations, in addition to genetic drift, which by sheer bad luck could wipe out in just a few generations a handful of 'good' mutations present on the duplicate gene. The survival criteria would be dominated by chance, and not positive selection.

This lineage must increase in size quickly or soon be eliminated by random drift. To provide the best chances one could assume this lucky event occurred in a small isolated population, since the chances of fixing in populations of sizes such as 10^{31} are remote.

Fixing an advantageous mutation would be easier in small populations, but fewer organisms would provide less total number of mutational opportunities. Without a large number of mutations new functions cannot arise.

A larger number or size of duplicate genes would generate more mutations on average per organism. But the disadvantages must not be overlooked. The negative selectivity factor, *s*, increases and the potential for cellular interference becomes much greater.

The date summarized in table 1 and figure 6 offers a fatal challenge to the evolutionist claims. Several studies have shown[7] that often only very limited subsets, such as *1 out of* 10^{50} random sequences, would lead to a minimally functional gene variant. But novel metabolic networks typically involve at least five totally unrelated genes, with translated proteins participating in a fine-regulated feedback inhibition scheme. The number of mutations which could accumulate on a duplicate gene could not reasonably provide enough random trials to develop such a complex system.

Given the negative results of our model and the data obtained in the laboratory with bacteria thousands of generations apart,[8] it must be concluded that gene duplication events do not provide a solution to explain how novel, complex biochemical processes could arise.

Conclusion

Ohno's conjecture of gene duplication followed by neo-functionalization may appear intuitively possible when evolutionary assumptions, such as huge time scales, are made. Mathematical rigour casts irrefutable doubt on this claim when examined in detail, even with huge timescales. The fact that expressed duplicate genes in small genomes are clearly selectively disadvantageous prevents such strains from surviving long enough to generate large numbers of mutations on the duplicate. Analysis of modern genes shows huge sequence differences from each other. Pairwise comparisons reveal that generally at least 30 judiciously placed mutations are needed[2] before a new function would be suspected, and for these suitable ones to arise a much larger number of random trials would be needed for natural selection to evaluate. But evolutionary theory needs to account for thousands of sequentially unrelated genes,

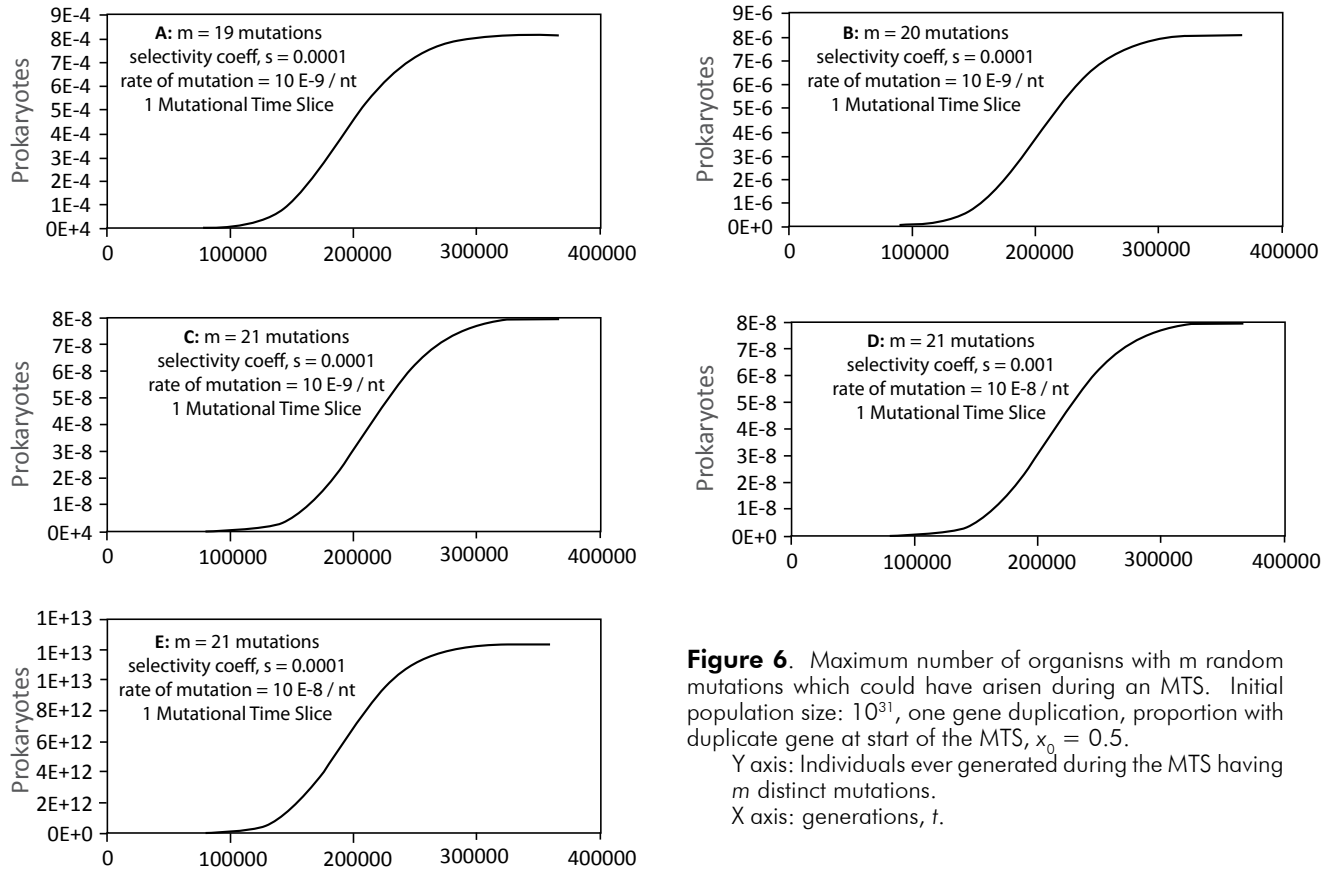

Figure 6. Maximum number of organisns with m random mutations which could have arisen during an MTS. Initial population size: 10^{31}, one gene duplication, proportion with duplicate gene at start of the MTS, $x_0 = 0.5$.
Y axis: Individuals ever generated during the MTS having *m* distinct mutations.
X axis: generations, *t*.

and these are the indispensable building material for new, complex biological functions.

In summary, the solution spearheaded by Ohno in 1970[1] does not only fail according to our model, but there is now ample biological evidence that gene duplication never was an important evolutionary mechanism.

Appendix 1: Metabolic costs to carry an unnecessary expressed gene

The expression of a gene presently not needed will waste energy and material, decreasing a prokaryote lineage's survival chances. We would like to estimate this in terms of a selectivity factor, *s*, to permit mathematical analysis.

In 2005, Wagner[6] provided the data needed to quantify the energy cost of expressing an additional average sized new gene in a single-celled organism. The cost of expressing all genes for the eukaryotic microbe *Saccharomyces cerevisiae* in relation to the total energy produced was then estimated.[6] The energy demand for a protein-coding gene consists of two components. (a) The first is the energy needed to manufacture the various building blocks: ribonucleotides for DNA and mRNA, and the amino acids used by the new gene's protein. These energy demands can be calculated since the biosynthetic pathways are known. (b) The second involves polymerizing the monomers to produce mRNA and polypeptide chains.

These costs result for each additional expressed gene. ATP, involving activated phosphate bonds (~P), will be consumed.

The amount of mRNA and protein produced per second in a cell can be estimated from kinetic studies, using experimentally determined rates of decay and synthesis and the number of polymer molecules present in the cell. Some doubt exists in the estimated decay rates for proteins, but the two methods ('RO' or Ribosomal Occupancy and 'HL' or Half-Life) used to estimate this value agreed to within an order of magnitude.[9]

S. cerevisiae synthesis of all *mRNA* was calculated to require about 6.69×10^5 ~P molecules per second, and *protein* synthesis still more (RO method: 1.55×10^7; HL method: 6.22×10^6 ~P molecules per second).

Gene expression is responsible for most of the energy consumed by the yeast studied, for several reasons. First, 51.3% of its biomass consists of RNA and protein,[10] and *'76.6% of the total adenosine triphosphate (ATP) cost of polymerization is invested into RNA and protein polymerization'*.[6] Thus, to a good first approximation, the energy demand for a duplicate gene is on average about inversely proportional to the number of expressed genes in that genome.

Having a duplicate gene would not be selectively neutral. Based on population genetics considerations,

Wagner pointed out[10] that there is a boundary *critical selection coefficient*, s, which is less than 1.47×10^{-7}. This number is based on a very small effective population size, and the critical value may be significantly smaller. This implies that if the presence of a duplicate gene penalizes more than this value, then natural selection will eliminate that lineage. For *all S. cerevisiae* genes Wagner studied, a duplicate gene carried an energy penalty far greater than the critical selection coefficient.[10] This assumes, of course, that the extra gene is not performing a useful function. If it were, it would not be free to mutate into a new gene anyway.

Using 4,346 yeast genes for which the necessary data was available,[6] a median selective penalty of ca. $s = 5 \times 10^{-5}$ for a duplicate gene was estimated. Individual nucleotides and amino acids require very different amounts of energy to be synthesized, so depending on composition and length, the amount of ATP needed to produce and express various genes can vary significantly.

Wagner considers the median selectivity disadvantage reported, $s = 5 \times 10^{-5}$, as too low.[11] One reason is that data for only 4,346 of the genes was available, and it was assumed in the estimate that those not measurable (40% of the genome)[12] would have average expression levels corresponding to the median of the others studied. But one reason kinetic data for these genes was not available is probably due to extremely low expression levels. In addition, a large proportion of the genes are expressed only under exceptional circumstances. These two considerations imply that the total energy produced per second was overestimated. Thereby, the incremental effect of expressing a duplicated a gene has been understated, and the true s must be considerably larger than 5×10^{-5} for *S. cerevisiae*.

We would like to draw attention to another consideration not addressed in Wagner's study.[6] The nutritional microenvironment during a microbe's lifetime can change considerably, both in the direction of great surplus and desperate shortage. Both can be detrimental in a relative sense to lineages possessing a duplicated gene. During plentiful nutritional conditions bacteria can initiate multiple rounds of chromosome duplication before dividing into a daughter cell.

> 'Fast-growing bacteria have growth rates requiring replication re-initiation before the round in progress is complete. In this way, *E. coli* can attain growth rates of 2.5 doublings/h.'[13]

This would compound the disadvantages of a larger genome not offering some immediate advantage. The cost of an extra duplicate gene, for which multiple copies would at least temporarily be present and expressed, would be larger under those environmental conditions. The physical chromosomal replication times (discussed below) would also be proportionally greater. Conversely, there will be periods of insufficient nutrients for all. Envision a case where only a minority of a population possesses duplicate genes. Under starvation conditions and deep time there will be many opportunities for natural selection to eliminate the borderline cases. The total population size would decrease, preferentially eliminating members with less efficient use of available energy. Once a lineage with duplicate genes has been totally eliminated or decreased to the level where genetic drift leads to extinction, the process must start all over. But the much larger 'normal' population can simply replace the unfortunate members once the nutritional conditions have improved. Population genetics analysis working with an average s value can be misleading, if extinction of the desired trait requires a long wait for a new evolutionary attempt.

The characteristics of proteins coded for by mutated variants of duplicate genes will be similar to the original, optimized version which is valuable for the cell. Our analysis has neglected the deleterious effect of interference by similar but inferior proteins. This factor would further increase the selectivity advantage of the streamlined organisms.

We are, of course, assuming the extra duplicated gene is not providing a benefit, such as a protein dosage effect. Mutations on such valuable genes would generally be strongly detrimental and therefore not accumulate to provide an opportunity to evolve into a brand new gene anyway.

The key intuition in this appendix is that the energy budget would be less efficiently used if unnecessary duplicate genes are present, and not be available for normal cell growth. This increases the risks of starvation and also increases the time needed before such organisms would grow to the necessary size before they would begin the replication process.

We can now make a reasonable estimate for the net penalty prokaryotes with a duplicated gene would experience. This is the inverse of the number of genes in the genome. For *S. cerevisiae* the selectivity disadvantage would be approximately $1/6{,}300 = \text{c. } 2 \times 10^{-4}$, about half the calculated value the author stated was too conservative. For putative microbial ancestors, with less than 2,000 genes, an s value between 0.0001 and 0.001 would then be expected per duplicate gene. An additional independent contribution to the selectivity coefficient, which further favours lineages lacking duplicated genes, is discussed in Appendix 2.

Appendix 2: Generation time of microbes

In Appendix 1 we quantified the selective disadvantage of prokaryotes carrying an extra gene, based on the less efficient use of available energy generated metabolically. Another factor to consider is the additional time needed to replicate a chromosome containing an additional gene. Under optimal laboratory conditions, modern microbes reproduce quickly, for example: *E. coli* (17 minutes), *B. megaterium* (25 minutes) and *S. lactis* (26 minutes). In the intestine *E. coli* has a generation time of about 12 to 24 hours.[14]

Interestingly, it may well be that smaller genomes often take longer to replicate their chromosome.

> 'For example, in E. coli the fork progresses at ~1000 nt /s, in *Pyrococcus abyssi* at ~300 nt /s and in *Mycoplasma capricolum* at ~100 nt /s'.[9]

Assuming 1,000 nt for a duplicate gene, a total replication rate of 1,000 nt /s implies this lineage would take about one second per generation longer to reproduce. Is this a significant issue? It depends on the average lifespan. For a generation time of 24 hours the reproduction time is increased by a factor of $1/(24 \times 60 \times 60) = 10^{-5}$. Under nutritionally rich conditions, multiple rounds of chromosome replication can occur before daughter cells are produced.[9] The streamlined genomes could save two or three seconds in a generation.

If average prokaryote generation times were indeed on the order of a day during billions of years, then the penalty of having a duplicate gene on chromosomal replication time would not be so severe, and only the factors discussed in Appendix 1 would be of primary interest. On the other hand, in Part 1 of this series we introduced the notion of a Mutational Time Slice (MTS). This is the average time for which if half the world's population of prokaryotes had a duplicate gene, virtually no additional members would add more mutations on an already highly mutated duplicate gene. To favour the evolutionary perspective we assumed a generation time of merely 20 minutes, so as to provide a very large number of MTS and thereby generate as many different highly mutated variants as possible during 4 billion years. Generation times of a day instead of 20 minutes would decrease the number of mutations which could accumulate, hindering the search for new genes by chance. With this in mind, Heisig and Truman generously proposed[15] an average generation time of 10 minutes, which of course presupposes a smaller genome. Using this estimate leads to an increased generation time factor of $1/(10 \times 60)$, or $s = c.\ 2 \times 10^{-3}$, for members with a single duplicate gene.

Evolutionary theory requires simpler prokaryotes to have preceded the extant ones. It is reasonable to assume the ancient DNA polymerases would have been initially less optimized for speed, and the chromosomes smaller. Then for hundreds of millions of years the relative penalty of carrying an extra duplicate gene would be considerable. In the case where chromosomal replication represents a large fraction of the prokaryote's lifespan, each additional gene would slow replication down by a factor roughly proportional to the number of genes present. Primitive prokaryotes with 1000 genes which lack an extra duplicate gene would have a selective advantage of on the range of 0.0001 to 0.001 per generation.

The effects described in Appendix 1 and 2 are independent and mutually reinforcing.

References

1. Ohno, S., *Evolution by Gene Duplication*, Springer-Verlag, Berlin, 1970.
2. Truman, R., Genome truncation vs mutational opportunity: can new genes arise via gene duplication? Part 1, *Journal of Creation,* **22**(1):99–110, 2008.
3. Lynch, M. and Conery, J.S. The evolutionary fate and consequences of duplicate genes, *Science* **290**:1151–1155, 2000.
4. Hoyle, F., *Mathematics of Evolution*, Acorn Enterprises LLC, Memphis, 1999. See equation 1.6 on p. 11.
5. Drake, J.W., Charlesworth, B., Charlesworth, D. and Crow, J.F., Rates of spontaneous mutation, *Genetics* **148**:1667–1686, 1998. See p. 1670.
6. Wagner, A., Energy constraints on the evolution of gene expression, *Mol. Biol. Evol.* **22**(1):1365–1374, 2005.
7. Truman, R. and Terborg, P., Searching for needles in a haystack, *Journal of Creation,* **20(**2):90–99, 2006.
8. Cooper, T.F., Rozen, D.E. and Lenski, R.E., Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*, *Proc. Natl. Acad. Sci. USA*, **100**:1072–1077, 2003.
9. Wagner, ref. 6, p. 1366.
10. Wagner, ref. 6, p. 1367.
11. Wagner, ref. 6, p. 1365.
12. Wagner, ref. 6, p. 1367, 1369.
13. Rocha, E.P.C., The replication-related organization of bacterial genomes, *Microbiology* **150**:1609–1627, 2004.
14. Todar, K., Growth of bacterial populations, University of Wisconsin-Madison Department of Bacteriology, 2007, <textbookofbacteriology.net/growth.html>.
15. Truman, R. and Heisig, M., Protein families: chance or design? *Journal of Creation* **15**(3):115–127, 2001.

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.

Peer Terborg 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.