PCID 3.1.1, November 2004

Evaluation of neo-Darwinian Theory using the Avida Platform. Part 2.

Royal Truman, PhD. Mannheim, Germany. Submitted: Oct 9, 2004.

Abstract

The development of new, ever more complex logic functions is very rapid under the usual Avida settings. It is computationally convenient to model with parameters which minimize the time needed for the simulations. These include: very high mutational rates; small population size; miniscule genome size; novel functions easy to generate by mutations; very little genetic material which can be damaged by mutations; very little indispensible genetic material; the exclusion of many services such as transcription, translation and physical reproduction from mutations; the presence of more complex functions easily attainable in a bootstrap manner; very high selectivities for new functions; and no penalty for carrying superfluous genetic material.

Such assumptions virtually guarantee the evolution of novel logic functions, however, due to these characteristics of the Avida platform. Whether the published papers are relevant to neo-Darwinian theory in real biological organisms is doubtful.

In project finance simulations the key costs and benefit factors need to be considered accurately enough to decide whether an investment is likely to result in net profit or loss. It is easy to create models and use parameters which predict either loss or gain, whichever one wishes. The details matter very much, and must be realistic enough for permit a valid conclusion.

Several factors would limit the possibility for random mutations in very small biological genomes to produce novel, complex functions. These are not considered in the Avida computer runs reported. Examples include no possibility for a graceful degradation of existing logic functions due to mutations, and no penalty for carrying superfluous genetic material. Extrapolating to biologically reasonable settings suggests neo-Darwinian theory is incapable of explaining more than rather inconsequential changes in cellular processes.

After just a few generations, Avida computer runs typically transform random strings of instructions to code for new logic functions. One suspects that the proportion of random strings able to code such functions is, in biological terms, unrealistically high. Under default conditions a population of 3600 digital members obtained a novel ***Not*** function on average in only 225, and ***Nand*** in 328 generations (Table 1). Genetic drift prevents most successful mutants from building up and fixing in the population. However, rewarding such lineages, as Avida default settings does, with an increased replication rate of about 2 for these two simplest functions increases the chances dramatically. Failed evolutionary attempts can quickly be replaced by new ones since these are easy for 3600 Avida organisms to generate.

Rapid increase in the proportion of members with the simpler logic functions facilitates build-up of the next more complex ones (Figure 1), in a bootstrap manner (Figure 2). It is therefor implied that the simpler functions are superfluous. For example, ***And*** can be selected for in spite of the total loss of the ***Nand*** from which parts were co-opted (Figure 1). Loss of simple functions in favor of more complex ones is also illustrated in Figure 2.

I observed during Avida runs that the average genome length, i.e., number of instructions present, varied over time (Figure 3). Deletions which occur within one of the initial, necessary 15 instructions for replication are more likely to kill that organism than is an insertion. Although both have 0.05 probability per replication, the net effect due to this effect is an increase in genome size on average. Another factor may also be at play. Point mutations can occur among the initial nop-C instructions provided at the start of most runs. Subsequent insertions and deletions in these regions are equally likely. Deletions would generally have no effect, but some insertions and point mutations could code for a premature *Divide*. The offspring would lack all instructions for normal replication. Overall, genome lengths tended to increase long before the first logic function was found by chance mutations (Table 1; Figure 4).

Without enough initially superfluous extra instructions, new logic functions cannot develop. Several runs were also performed in which no mutations were allowed, and confirmed that the software performs as expected: the population was invariant.

It was observed that the more challenging logic function ***OrNot*** was easily generated by random mutations with no need for build-up of intermediate, simpler functions (See Table 1). The first member was found in only 331 generations on average. This illustrates that even more complex Avida functions are statistically rather easy to find by chance mutations.

I explored next the origin of logic functions after lowering the rewards of all nine logic functions to $2^{0.01436}$ to represent a biologically more realistic relative fitness, ω, of about 1.01. As expected, the level of reward will not affect when ***Not*** & ***Nand*** first appear (Table 2). However, **OrNot** & ***AndNot*** take about 3 to 4 times more generations for the first member to be generated. Under the reduced reward conditions none of the higher logic functions appeared by 5000 generations, in sharp contrast to what was reported earlier (Table 1). More importantly, observe that higher functions which do appear are easily lost due to genetic drift. Appearance and loss would have to occur many times before each is fixed, requiring now many times more than 5000 generations (Figure 5).

I decided to estimate how often the two simplest logic functions arise with no reward at all. From Table 3 one observes that many are found by chance. But Figures 6 and 7 illustrate that without strong selection they are very unlikely to fix even after an inmense number of generations.

To illustrate how often logic functions can be lost in the absence of unrealistic rewards, as used routinely in Avida runs[<1>], a test was performed in which the four simplest logic functions (***Not***, ***Nand***, ***And***, ***OrNot***) were not rewarded, and a reward of 1.01 was used for the remaining five functions. From Table 4 one sees that most of complex logic functions were generated (since this is statistically easy). However, during the tens of thousands of generations, one observes that these lineages would repeatedly appear, then die out (Figures 8-10). At the end of the run none of the higher logic functions were present. Due to the large number of ways ***No***t & ***Nand*** can arise, significant proportions of these often built up even in the absence of reward (most of the experiments are not reported here). ***Or*** & ***AndNor*** arose often but in spite of a positive selectivity did not fix even after 36000 generations, the pre-set point to terminate the runs. At that point the average genome size was 63.5 instructions, so genome truncation had not occurred to hinder evolution of new logic functions.

The effect of starting with fewer nop-C instructions.

Without genetic "playing material" no new logic functions can evolve. In the next series of experiments (Table 5; Figure 11) I explore how much longer it takes for the first example of each logic function to appear starting with 5 nop-Cs instead of the usual default 35 instructions (Table 1). It now takes about 10 times more generations for one of the two simplest logic functions to be found. Figure 12 shows that even this was facilitated by the rapid increase in genome size in the absence of reward, which is simply an artefact of the Avida platform.

The initial 35 nop-C instructions cannot perform logic functions until a large number of mutations have occurred. During this period genomes are also varying in size. These factors make more difficult an estimate of the proportion of random sequences able to perform a logic function. Some experiments were devised using initially random instructions instead of a string of nop-Cs. A different string was used for each of four trials, each with a different random seed. The Avida parameter **child_size_range** was set to 1.0 and both inserts and deletes were deactivated in the ***genesis*** file. This ensures invariant genome size and production of variability only through point mutations. The initial genomes (Table 6) had 15 random instructions. The average proportion of ***Not*** functions from among all random 15-instruction strings generated was $\mathbf{1.5X10^{-3}}$, and of Nand functions $\mathbf{4.5X10^{-4}}$.

In the next experiment (Table 7) 15 additional random instructions were added to each of the starting sequences reported in Table 6, and the same random seed was used pairwise across the four repeated runs reported in Tables 6 & 7. The average proportion of ***Not*** functions from among all random 30-instruction strings generated was $\mathbf{7.5X10^{-3}}$, and of Nand functions $\mathbf{7.5X10^{-3}}$. The later runs imply that about 0.1 to 1% of all 30-instruction length random strings would contain a logic function. The additional fifteen instructions more than doubled the proportion of logic functions found. The biological implication by extrapolation would be that many random polypeptide 300 residues long could perform at least one useful function. This unstated implication is one of the key reasons Avida runs produce new functions so easily, but this is biologically absurd.

Experiment 17iii from Table 7 was selected for a very long test, being close to the average of the four tests performed. The density of logic functions was decreased by only rewarding ***OR***, ***NOR***, & ***EQU***, with a factor corresponding to a relative fitness $\omega = 1.07$ (i.e., $2^{0.1}$). These settings were intended to help extrapolate further in the direction of biologically realistic constraints. A wide variety of logic functions almost always fix in the population within a few hundred generations when Avida runs are performed under default and typical parameter settings. However, now no logic function fixed during hundreds of thousands of generations. Some would be generated, occasionally build up to a handful of members, then die out. The run self-terminated at 1,000,000 updates, at which point 360,239 generations had been tested and 87,946,048 genomes produced. The average genome length in this run remained at 45 instructions. At the end of the experiment no functions were found among any of the 3600 organisms.

A lineage containing ***OR*** did not fix in the population by simpler stepping-stones in this experimental setup. Nevertheless, it was generated by random mutations a large number of times, since this is statistically not especially difficult. The first example occurred during update 177,788 (62,874 generations and 15,295,879 different random genomes tested). Although a build-up of a few members did occur frequently, due to the relatively high $\omega = 1.07$ used (a maximum of 14 members in the 3600 size population in one case), total lineage extinction occurred soon afterwards in all cases. Note that discovering a brand new, complex

cellular function in a 37 random codon sequence (i.e., 30 Avida instructions X 26/21) in $1.5X10^6$ mutant trials is unrealistic by many orders of magnitude for real biological proteins.

The yet more complex **NOR** was also discovered many times in this experiment. The first example occurred in update 177,737 (62,856 generation and 15,291,357 random genomes). Multiple new attempts followed by extinctions occurred. A maximum build-up of 4 out of 3600 members was observed. No examples of the final target, EQU, were present.

To reinforce the point being illustrated: it is the very dense set of logic functions; their use to build upon to create yet more complex functions; and the unrealistically high fitness rewards assigned to the ever higher functions, which permit Avida digital genomes to race up a fitness slope. I repeated the experiment just described, using the same starting sequence and random seed, but this time excluding only the two foundational logic functions (***Not*** & ***Nand***) and assigning the default Avida rewards to the remaining: ***And***, ***Orn***: 4; ***Or***, ***Andn***: 8; ***Nor***, ***Xor***: 16; ***Equ***: 32. All kinds of logic functions built up in this run. Whereas after hundreds of thousands of generations not the remotest possibility of fixing **Or**, nor ***Nor*** was found above, these default parameter settings (i.e., as installed with the Avida software) led to fixing of both functions within only 7,910 generations (26,567 updates; 1,610,936 different genomes tested).

Rewarding also the simplest, foundational ***Not*** & ***Nand*** logic functions with values of 2, using the same starting sequence and seed as just analysed in the two preceding runs, showed that ***Or*** and ***Nor*** fixed within only 376 generations (4,593 updates; 74,593 different genomes tested). This later experiment reflects the default settings used when Avida is installed and used. I started with a random string of instructions instead of nop-Cs for clarification purposes.

Fitness as a function of genome compactness. Avida digital organisms initially can only reproduce, and perform no logic functions. Strings produced which contain instructions able to code for logic functions are rewarded by faster reproduction rates. In the usual settings[<1>], non-functional instructions are not penalized, but rather extra "SIPS" are made available in proportion to genome size. As mentioned in Part 1, for such miniscule genomes this is biologically incorrect and furthermore, the effect of genome truncations using Avida runs has been documented in the literature[<2>]. A curve[<2>] shows that about 2/3 of the digital genome is removed in less than 100 generations (ca. 400 updates). This is consistent with experiments *in vitro* using small RNA strands[<3>], and the known deletional bias observed in bacterial genomes[<4>].

Incidently, genome truncation due to small deletions in bacterial size genomes, which are thousands of times larger than Avida digital genomes, would be much slower. (Larger deletions do, however, occur often enough that species sizes tend to remain compact[<4>]). But these living systems must code for transcription and translation proteins and RNA; must code the proteins responsible for cell division, and for some indispensible other house-keeping proteins. All this genetic material is subject to mutational risk, unlike Avida genomes. A much larger proportion of mutations would be deleterious, so that the number of random attempts needed to produce brand new functions becomes inmense.

One cannot perform the relevant Avida runs to explore the effect of replication time vs. genome size. In the documentation (genesis.html) one reads, "***Merit is typically proportional to genome length otherwise there is a strong selective pressure for shorter genomes (shorter genome => less to copy => reduced copying time => replicative advantage").***" No parameter

options are available to the Avida platform for the "genesis" file which would permit truncation in some realistic manner, proportional to the genome length. As demonstrated in Part 1 and reported in the literature[2], [4], [3] this truncation would be very rapid. In such organisms superfluous genetic material to produce new, complex biological functions would not be available.

Biologically realistic values for relative fitness.

Avida type experiments assume that complexity increasing functions are easily attainable via random mutations. These new functions provide the fortunate mutant with enhanced fitness. Fitness increases the proportion of the endowed lineage within the population, either through superior survival or reproductive chances, or both. Evolutionists usually assume relative fitness values, ω, of about 1.01, which is orders of magnitude lower than usually assumed in Avida experiments. I point out in Appendix 1 the difficulty in accepting large ω values for bacterial alleles, based on designed, artificial laboratory conditions. It is suggested that under natural conditions typical bacterial alleles would confer relative fitness values of essentially $\omega \approx 1$ averaged over time.

Other examples of rapid selection in higher organisms have been offered as evidence for evolution. However, examples of changes in population proportion which represent net degradation are contrary to the kinds of examples needed to support evolutionary theory (Appendix 2).

The proportion of useful proteins among all polypepdide sequences.

Generating novel logic functions by random mutations of individual instructions in Avida runs is statistically very much easier (Tables 6 & 7) than generating real proteins in such a manner. What proportion of random amino acids chains might perform a biologically useful service? Those of an evolutionist persuasion must argue that finding these by random chance plus natural selection is not in the realm of the miraculous. Several arguments have been proposed.

Some argue that enzymes may have had other or similar functions earlier. For reasons discussed in Appendix 3 it is very unlikely the neo-Darwinian theory can account for the origin of the complex enzymatic networks.

To work properly, a protein must fold in the same three dimensional structure reliably. Any polypeptide chain containing *n* residues could, in principle, fold into 8^n conformations[5]. Most of the evidence suggests that only a very small proportion of polypeptides would fulfill this minimal requirement to be a useful protein. Studies claiming this proportion may be high enough to be relevant for evolutionary purposes are not plausible (Appendix 4). The designed examples are neither representative of natural proteins nor can the data be extrapolated in any meaningful way. It is likely that Sauer's estimate[6] of about 1 out of 10^{-65} chains of typical protein length might fold reliably is closer to the truth. This contrasts sharply with Avida logic functions: I estimated that on average about a 0.1% of random strings of protein length would be expected to possess at least one logic function.

Some mutations do generate novelty. The typical examples, however, are rather trivial variations on a working theme, leading to no increased complexity. These mutants can be typically reached easily by one to three mutations (Appendix 5). This does not permit

extrapolation to the demands of producing novel proteins which are sequentially far removed from any other, since these cannot be developed via a large number of selectively advantageous intermediates.

Genes may actually display little sequence variability.

If this is the case, the assumption of many convenient steppings close together is false. As an example, I consider the recycling of proteins by proteosomes (Appendix 6) , as practiced in all eukaryotic cells, with the help of ubiquitin. An exhaustive search[<7>] for all reported ubiquitin sequences provided data for about 160 organisms, covering plant, animal and yeast life forms. Besides methionine, **28** of ubiquitin's 76 residue positions were found to be invariant (Appendix 6) and for a large number of the remaining positions only one example was found with a single different amino acid at one position and for only one organism (these might actually be a post-translational modification and not the product from original mRNA. The number of invariant residues may actually be higher). In only two residue positions was significant variability observed, and these was limited to very few amino acids.

What are the putative preceding functional stepping stones for ubiquitin and the large number of other proteins involved in this function? Probabilistic jumps on the order of $< (1/20)^{28}$ (i.e., $4X10^{-37}$) stand in stark contrast to the ease of producing new Avida logic functions as pointed out above.

Finally, the net effect of mutations needs to be considered critically. Notice that Avida logic functions are defined by discrete instructions with a specific fitness. Inexact sequences results in total loss of a logic function and a dramatic loss in fitness (Figure 13). This prevents a graceful slide down a fitness slope. Variants of biological proteins, however, can function with slightly less effectiveness. I propose that in nature many non-deadly mutations can accumulate with relative selectivities too small to measure, given the large number of challenges each organism faces over its lifetime. The number of less effective sequences over all proteins greatly outnumbers the better ones[<8>,<9>], making this process statistically probable (Figure 14).

As slighly deleterious mutations over a large number of different proteins accumulate, the average fitness of the population decreases, making additional degrading mutations yet easier. The net effect of mutations is predicted to be loss of function and specificity over time, and not the creation of novel complex biological functions.

Summary.
Avida computer runs under typical parameter settings produce new logic functions easily. The simple ones are very easy to generate by random changes in individual instructions. The proportion of random strings possessing at least one logic function was examined as a function of instruction string length. Extrapolation to protein size sequences shows that allegedly a considerable proportion of biological polypeptides based on the 20 natural amino acids would supposedly perform a useful function. However, truly novel, complex biological functions demand many new genes to be created. It has been estimated[<6>] that perhaps 1 out of 10^{-65} random polypeptides fold reliably, and random mutations are unlikely to discover initial starting points for natural selection to act upon.

Avida runs typically reward organisms with new logic functions with dramatically faster reproduction times, and those lineages can then easily bootstrap up to ever more complex

logic functions. Simultaneously lowering the density of convenient stepping stones and the relative fitness these are rewarded with presented a totally different picture. It was shown that using more typical relative fitness values (such as ω in the 1.01 to 1.1 range) led to repeated extinctions of logic functions found by chance mutations. These lower ω values reflect the fact that most organisms face many survival challenges simultaneously and can die for reasons unrelated to the factor being selected for. Instead of consistently fixing in a few hundred generations, at the end of hundreds of thousands of generations (when the trials were terminated) no logic functions at all were present.

Other serious considerations render the picture more dismal. The more compact genomes of these kinds of rapidly-reproducing asexual organisms would quickly out-reproduce those carrying superfluous junk, removing the necessary genetic substance for evolutionary trials, as shown both experimentally[<3>, <4>] and with Avida[<2>].

It was pointed out that Avida organisms do not code for the physical infrastructure to actually perform cellular processes: RNA and dozens of proteins are necessary for translation, transcription, replication and other house-keeping activities, and these are genetically coded for in biological cells and thus subject to mutational risk. All these services are provided by external software written in C++ and by computer hardware in the Avida platform. Mutations cannot affect these, which permits such a high mutation rate, as discussed in Part 1. Time constraints limits the number of mutation possible and in nature most of these would be "wasted" due to destruction of necessary, genetically encoded, functions.

Slowing down the mutational rate a few orders of magnitude to realistic levels would offer far more generations in which genome truncation can occur, since those lineages reproduce more quickly, preventing the generation of novel functions. In fact, the loss of large numbers of genes in bacteria has been deduced from comparative sequence analysis[<29>, <30>].

Avida does not permit logic functions to be almost correct, unlike real proteins, rendering it impossible to model a graceful slide down a fitness slope. Loss of a logic function due to a single instruction mutation is generally so heavily penalized (e.g., up ca. 32 times slower reproduction rate per generation for that lineage in the default setup!) that most deleterious mutations get eliminated. In nature, mutations with almost neutral selectivities would permit degradation to occur. It is proposed that as populations degrade across all proteins in which this is possible, the average fitness will be lowered, permitting yet more degradation without weed-out by natural selection. *Contra* the views of many Avida reseachers, the net effect of mutations in nature appears to be destruction of specificity and functionality, and not the creation of increasingly complex genomes.

The typical parameter settings and assumptions used in Avida based research guarantee the desired outcome: mutations and selection produce novel complex functions. Demanding calibration to biological realities shows the conclusions are not legitimate.

Methods.
Avida VERSION_ID 2.0b6 was download from http://myxo.css.msu.edu/papers/nature2003/
At the DOS command line the software was initiated with: primitive –g my_genesis.txt

The original *genesis* file was slightly modified to conform to the details reported in[<1>] and renamed to my_genesis.txt (Table 10). Tables 11- 13 show the default settings used in this study.

The computer runs were performed on Windows 2000 and Windows XP hardware. Random strings were generated by producing random numbers between 1 and 26 with Excel 2000 and converting these to the corresponding Avida instruction number. A large number of experiments could not be reported here due to space constraints. This explains the gaps in experiment number among the tables provided in this paper. These unreported computer runs were designed mostly to address the question of proportion of random strings which can perform a logic function; and failure to build up complex logic functions due to repeated extinctions (genetic drift). The latter are tested by requiring lower proportions of logic functions and that their rewards be made more biologically realistic.

Acknowledgements.

I wish to thank Dr. Lenski for some clarifications he kindly provided and Mr. Kaben Nanlohy for helpful technical comments on Avida. Critical exchanges with Dr. Charles D. and Dr. RBH forced more careful rewording. Mr. Salvador Cordova stimulated / provoked much thought. Mr. David Coppedge's suggestions improved an earlier draft.

References

<1> Lenski, R.E.; Ofria, C.; Pennock, R.T.; and Adami, C., "The evolutionary origin of complex features", ***Nature***, **423**(8), (2003), 139.
http://myxo.css.msu.edu/papers/nature2003/

<2> Ofria, C.; Adami, C.; and Collier, T.C., "Selective Pressures on Genomes in Molecular Evolution", *J. theor. Biol.* **222,** (2003) 477-483.
http://arxiv.org/abs/quant-ph/0301075.pdf

<3> Mills, D.R.; Peterson, R.L, & Spiegelman, S., "An extracellular experiment with a self-duplicating nucleic acid molecule", *Proc. Natl. Acad. of Sci. U.S.A.* **58**, (1967), 217.

<4> Mira, A.; Ochman, H.; and Moran, N.A., "Deletional bias and the evolution of bacterial genomes", *Trends in Genetics* **17**(10), (Oct., 2001), 589-596.

<5> Lodish *et al.*, "Molecular Cell Biology", 4th ed., W. H. Freeman and Company, New York, p. 62, 2000.

<6> Reidhaar-Olson, J. and Sauer, R., "Proteins, Structure, Function and Genetics", **7**, (1990), 306; Bowie, J. and Sauer, R., *Proc. Nat. Acad. Sci. USA* **86**, (1989), 2152; Bowie, J., Reidhaar-Olson, J., Lim, W. and Sauer, R., *Science*, **247**, (1990), 1306; Behe, M., "Experimental support for regarding functional classes of proteins to be highly isolated from each other"; in: *Buell J. and Hearns, G.* (Eds), "Darwinism: Science of Philosophy?", Haughton Publishers, Dallas, 1994, pp. 60–71. Discussed in: Behe, M.J., Dembski, W.A. and

Meyer, S.C., "Science and Evidence for Design in the Universe*",* Ignatius, San Francisco, 2000.

<7> Unpublished personal research.

<8> Axe, D.D., "Extreme functional sensitivity to conservative amino acid changes and enzyme exteriors", *J. Mol. Biol.* **301**, (2000), 585.

<9> Truman, R.; and Heisig, M., "Protein families: chance or design?", *TJ*, **15**(3), (2001), 115.

<10> Hartl, Daniel L. and Clark, Andrew G., "*Principles of Population Genetics*", Third Edition, Sinauer Associates, Inc., Sunderland, Massachusetts, 1997.

<11> Hartl, D.L. and Dykhuizen, D.E., "Potential for selection among nearly neutral allozymes of 6-phophogluconate dehydrogenase in Escherichia coli." *Proc. Natl. Acad. Sci. USA* **78**, (1981), 6344-6348.

<12> Catchpoole, David, "Worm evolution in pollution?", *Creation* **26**(3), (2004), 54-55.

<13> Levinton, J.S., Suatoni, E., Wallace W., Junkins, R., Kelaher, B. and Allen, B.J., Rapid loss of genetically based resistance to metals after the cleanup of a Superfund site, *Proceeding of the National Academy of Sciences USA* **100**(17), (2003), 9889-9891.

<14> Alberts, Bruce et. al., "Molecular Biology of The Cell", Third Ed. Garland Publishing, 1994, chapter 3.

<15> Knowles, J.R., "Tinkering with enzymes: what are we learning?" *Science* **236**, (1987), 1252-1258;
Knowles, J.R., "Enzymes catalysis: not different, just better". *Nature* **350,** (1991), 121-124.

<17> Fell, David, "Understanding the Control of Metabolism", Portland Press, London and Miami, 1997.

<18> Davidson, A.R., and Sauer, R.T., "Folded proteins occur frequently in libraries of random amino acid sequences", *Proc. Natl. Acad. Sci. USA* **91,** (March 1994), 2146-2150.

<19> Hall, Barry G., "The EBG system of *E. coli*: origin and evolution of a novel β-galactosidase for the metabolism of lactose", *Genetica* **118, (**2003), 143-156.

<20> Ref. <5>, p. 66-67; 503-504.

<21> Hershko, A. and Ciechanover, A., "The ubiquitin system for protein degradation", *Annu. Rev. Biochem.* **61**, (1992), 761-807.

<22> Rechsteiner, M.; Hoffman, L.; and Dubiel, W., "The multicatalytic and 26S proteases", *J. Biol. Chem.* **268**, (1993), 6065-6068.

<23> Varshavsky, A., "The N-end rule". *Cell* **69**, (1992), 725-735.

<24> Stryer Lubert, *Biochemistry*, Fourth Edition, W. H. Freeman and Company, New York, 1999, p. 942-943; 988.

<25> Ref. <14> p. 218-221.

<26> http://www.ncbi.nlm.nih.gov/BLAST June 16, 2004. Use Protein-protein BLAST (blastp)

<27> ClustalX downloaded from: http://www.ebi.ac.uk/clustalw/

<28> Ref. <24>, p. 977.
"The amino acid sequences of H4 from pea seedlings and calf thymus differ at only 2 sites out of 102 residues. "

<29> Scherer, S., "Kleinstes Genom einer photosynthetischen Bakterienzelle als Hinweis auf einen genetisch "komplexe" Vorfahen?", *Stud. Int. J.* **11** (2004), 29-31.

<30> Dufresne, A. *et al.*, "Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome". *Proc. Natl. Acad. Sci. USA* **100**, (2003), 10020-10025.

Appendix 1**. Large relative fitness values under laboratory conditions are not representative of nature**.

In haploid, asexual organisms such as bacteria, ω can be determined by competition studies in a chemostat. For simplicity, one can assume discrete generation growth for two genotypes[<10>]:

$$\mathbf{A_t = A_{t-1} + rA_{t-1}} \qquad [1]$$

where **A** is the population size, t is a generation period, and r is the intrinsic rate of increase. This leads to the relationship $\mathbf{A_t = (1 + r)^t A_0}$ where A_0 is the population size at the beginning of the experiment.

The ratio of A to B type bacterial populations is given by:

$$\frac{A_t}{B_t} = \left(\frac{1+a}{1+b}\right)^t \frac{A_0}{B_0} = \omega^t \left(\frac{A_0}{B_0}\right) \qquad [2]$$

where **a** and **b** are the intrinsic rates of increase for genotypes A and B, and **ω** is the relative fitness (1+ a)/(1+b).

In laboratory tests it is convenient to transform into logarithms:

$$\log\left(\frac{p_t}{q_t}\right) = \log\left(\frac{p_0}{q_0}\right) + t\log(\omega) \qquad [3]$$

where **p** and **q** are the frequencies of genotypes A and B at generation t, and **ω** is the relative fitness. Plotting $\log(p_t/q_t)$ vs. t should give a straight line with slope equal to log ω.

As a concrete example[<11>], E. coli strains containing either allele gnd(RM77C) or gnd(RM43A) were compared. This gene is polymorphic in natural populations and codes for the enzyme 6-phosphogluconate dehydrogenase (6PGD) which is part of the path to metabolize gluconate. In the chemostats only gluconate was provided as a source of carbon and energy. This enzyme is not needed in the metabolism of ribose. Controls in which only ribose nutrient was provided led to a value of ω = 0.999 based on about 35 generations, which indicates that the strains are otherwise essentially identical in fitness.

When only gluconate was available, gnd(RM43A) conferred a selective advantage, with ω = 1.0696, i.e. a selective advantage of about 7%.

Evaluation of such examples.
There are several reasons I cannot accept a selective advantage of this size as being typical to develop new complex biological functions in small, haploid organisms, under natural settings.

The initial proportion with the RM43A version was 0.455 at the beginning of the experiment and this increased to 0.898 within only 35 generations, just a few hours. A relative fitness of ω = 1.0696 under natural conditions would cause the less fit to disappear quickly. For example, assuming a generation time of one hour, eqn. [2] indicates that the more fit would dominate by a proportion of $\mathbf{1.4X10^{256}}$ to one within a year, a number greater than all atoms in the universe (about 10^{80}).

There are many nutrients available in the intestines to E. coli. It is not surprising that some alleles would be more effective in metabolizing the breakdown of one or the other. By excluding all but one the researcher can "get what one selects for" under artificial laboratory conditions.

One could experimentally determine relative fitnesses under more natural conditions over time. Fluctuations in proportions of nutrients can be expected to favor one allele over the other. Suppose that after about one year, or $t \approx 8766$ generations, the relative proportion shifts by one percent, a rather high value. From eqn. [2] this would imply a miniscule $\omega = \mathbf{1.000001}$.

This has several evolutionary theory implications.

• A radical environmental change, such as presence of a single new nutrient, may indeed favor a particular allele, but only if it is already present or very easy in statistical terms to generate by mutations.

• Relative selectivities under realistic and variable environmental conditions for the kinds of experiment described above are typically far lower than 7%. Rearranging eqn. [2] allows us to calculate the number of generations needed on average for a single initial mutant to double in a new lineage (Table 8), as a function of relative fitness. $\omega = 1.01$ would require about 70 generations, during which a one-member "lineage" would often die out. Genetic drift effects would require that an initial mutant forefather be recreated a vast number of times on average until a minimal fraction has built up.

• The kinds of examples described above show no increase in biological complexity. They are merely variability within an existing theme. All the enzymes needed to metabolize the nutrients are already present and the necessary gene expressions already well regulated. Processing other nutrients which are chemically similar usually requires but small changes in geometric and electronic characteristics at a single enzymatic catalytic site. This is a considerably simpler matter than creating the foundational molecular machines, which are composed of multiple complex proteins.

• Strong selection for a single trait for many generations permits those deleterious mutations which do not determine survival to build up. If $\omega = 1.1$ and 10^9 bacteria are present (about one liter of water in nature) then eqn. [2] predicts some 169 generations would be needed on average for a single mutant's lineage to increase to a fraction of 1% of the population. It is not surprising that antibiotics-resistant bacteria in hospitals fare poorly when forced to compete in natural multi-challenge environments. Single factor strong selection permits other less damaging mutations to build up, and the "good" mutation are generally deleterious under natural settings, since typically over-production of protein occurs or binding site specificities are damaged.

Degradation of other functions or of gene specificities as a price for short term survival would surely result in an increase in sequence entropy. Loss of functioning genes, such as found in blind fish in dark caves or loss of wings in some island beetle species, illustrate a trend in the opposite direction than as required by neo-Darwinian theory.

Appendix 2. **Some evolutionary examples actually reflect degradation**.

There are many reports of rapid genetic adjustment to environnmental changes. "***Between 1953 and 1979, a battery factory released approximately 53 tons of cadmium and nickel hydride waste into Foundry Cove on the Hudson River. Cadmium became very concentrated (up to 10,000 parts per million) in the riverbed sediments. Despite such high levels of toxic cadmium, a riverbed population of a worm species, Limnodrilus hoffmeisteri, survived the pollution – i.e. it was said to have 'evolved resistance' to the cadmium.***"[<12>, <13>] Subsequent cleanup efforts lowered cadmium to less than 10 ppm levels. The fraction of (original) non-resistant genotypes began to rise, so that between only 9 to 18 generations the proportion of that of other areas not affected by such pollution resulted.

Worms resistant to high cadmium concentrations were already present in low proportions in the original population and catastrophic selection favored the elimination of the predominant type. The key gene produces metal-lothionein-like protein which binds cadmium. It appears that the original well-regulated expression of this gene is occassionally damaged by mutations leading to over-abundance of the protein. According to the Shannon definition of information, this represents a loss of information, because a larger variety of binding sequences is now present.

Such examples are precisely the opposite of what neo-Darwinian theory needs. The more precise construction was damaged, causing waste in energy and materials due to the over-production of a protein, rendering such variants clearly less fit to compete under a wide variety of challenges. Adjustment to a catastrophic single factor comes at too high a price for evolutionary theory purposes. Extending the argument, additional pollutants could be introduced sequentially, leading to a crippled lineage which eventually cannot survive under any natural circumstances.

Appendix 3. **Extant enzymes may have had a slightly different function earlier**.

Enzymes are precisely crafted catalysts. The chance of being present among a random sequence is essentially zero. Could they have arisen in a neo-Darwinian manner?

The exposed residues of the folded protein(s) which make up an enzyme need to form weak, noncovalent bonds with ligands. Typically a large number of such interactions are necessary, which places severe constraints on three-dimensional and electronic details of the binding site[<14>]. For example, peptide bonds can be hydrolyed using either acid or base catalysis. In a chemical flask these would neutralize each other. However, enzymes can use differently charged amino acids on the same chain in such a manner that they interact simultaneously with the ligands using both catalytic effects[<15>,<16>]. In one enzyme, replacing a glutamic acid with an aspartic acid shifts the position of the catalytic carboxylate ion a mere Angstrom (about the radius of a hydrogen atom) yet this reduced the enzymatic activity by a thousandfold[<14>].

Reaction rates so fast that they are difusion controlled are well known[<14>]. To avoid the need for greater enzymatic concentrations and optimize the thoroughput yet more, assemblies of multienzyme complexes exist, molecular machines built along the principles of assembly lines.

Now, both the substrate and genes need to be present at the same time and place for an organism to profit from such equipment. A gene producing non-functional polypeptides which are almost correct in terms of amino acid sequences would be deleterious. The odds of

obtaining an evolutionary starting point natural selection could act upon would be improved if a number of similar alternate ligands were available, processing any of which would be advantageous.

This is the intuition behind reports of mutant genes found to be useful in processing novel substrates, even artifical man-made ones.

Evaluation. Although this argument does increase the chances of providing an initial starting point, it does not really lend very strong support to the view that such neo-Darwinian processes indeed created the molecular machines observed without any intelligent input.

• The ancestral genome at some point in the distant past would have survived without that biochemical process, so the necessary genes were not present and the relative selectivity to build these via random mutations one at a time would surely not be very large.

• Providing any starting point for an evolutionary process seems to have been slightly improved, but by how much? An enzyme also catalyses the back reaction. As a rule, enzymatic metabolism involves a well-regulated linked network[<17>] of reactions designed in a manner to hinder equilibrium back to the starting point. A stand-alone catalyst is not automatically a useful feature.

• An inexact catalyst able to affect several similar substrates facilitates destruction of cellular material needed for other purposes.

• How does natural selection choose which substrate to specialize on? The relative concentrations may change over the considerable number of generations needed before the multiple, fine-tuning, necessary mutations occur. Natural selection is presented with a shifting target.

• In haploid organisms evolutionary improvements in one lineage competes with unrelated improvements in a different lineage. Mutations in one member of a future multienzyme complex competes against a parallel attempt to optimize a different component in a different lineage. Generally only one would win out and the other attempts would be lost. Improvement must be for the most part sequential along the winning lineage only. However, there are only so many mutational attempts available and most would be "wasted".

• Once an enzyme has been optimized for one metabolite, exactly how does the species shift to a new mutant allele? One does not observe a proliferation of mutant gene copies in bacteria. In most bacterial genomes typically about half of the sequences do not show any sequence similarity to other genes on the chromosome, and for those which do, similar design constraints in the protein domains can be expected to require similar solutions.

If a small number of alleles were present, each able to metabolize a different nutrient with slightly different effectiveness, then natural selection would generally not be presented with a single, fixed goal over a large number of generations to fine-tune each enzyme variant.

Appendix 4. **The proportion of folded polypeptides might be higher than expected**.

Cassette mutagenesis studies[<6>] by Sauer at MIT allowed an estimate of the proportion of polypeptides able to fold properly. For the cases studied, a proportion of about 10^{-65} was reported, although no biological function was shown to exist even for that subset. This corresponds statistically to guessing correctly one atom in our galaxy.

In some studies, however, much higher probabilities were estimated. As an example, Sauer[<18>] studied a class of synthetic genes which code for 80 to 100 residue polypeptides composed mainly of glutamine (Q), leucine (L), and arginine (R). Glutamine is hydrophilic, leucine hydrophobic and arginine is charged, and added to increase solubility of the peptide in the cell. Ampicillin-resistance gene was included in the plasmid to permit selection of *E. coli* colonies possessing the synthetic vector. The product from three QLR genes were isolated and analysed and may have indeed folded stably. CD spectra of all three indicate the presence of α-helical secondary structure.

Protease resistance suggests a stable hydrophobic core had resulted in the three variants studied and gel filtration resulted in migration primarily as a single species. The authors conclude that this study implies "***that a significant fraction of random sequence proteins should fold into unique structures under native conditions***."[<18>]

Evaluation.

- The presence of α-helixes is not surprising, since as the authors point out[<18>] the three amino acids selected have a high propensity to pack in this manner. This is not representative of random sequences from among 20-plus natural amino acids.

- The fractions of hydrophobic and hydrophilic residues were chosen on the basis of reasonable hydrophobic and hydrophilic proportions for folding purposes. Notice the low variability among the proportions for the three polypeptides isolated (Table 9). This is clearly a carefully selected class of DNA sequences and not at all illustrative of random sequences.

- None of the proteins show significant loss of α-helixes up to 90°C, the highest temperature tested, in the presence of 6.0 M Gdn•HCl. No natural proteins are known to display such stability, and no biological use for them can be envisioned. In the same manner that the DNA double helix cannot be too stable to be useful for biological purposes, proteins must also not be too rigid (they must travel through membranes, permit entry of ligands, etc.).

- The gel filtration elution profiles indicate that trimers and tetradecamers were formed in two of the three QLR variants. In the third a single oligomeric species was formed.

- The variants were not soluble in aqueous media such as present in cells.

- A large number of "genes" coding for 50% Q (hydrophilic), 40% L (hydrophilic) and 10% E (charged) will often produce long "runs" with only Qs and Ls on strictly statistical grounds. For example, one of the three polypeptides generated displayed the sequence[<18>]:

QLLLLQQQQQQLQQQQLQQLILQLQLRQLQLLLLLLRLLQQLLQIRWLQLLLLQLQ
QRLQQQQQQLQRL

This explains the creation of stable folded structures: the very hydrophobic "blocks" will remain together for long periods of time as the polypeptide chain searches the thermodynamic folding space.

• The number of different random sequences generated was not reported nor was an estimate of the fraction thereof which folds stably. This makes it impossible to extrapolate to random polypeptides based on the 20 natural amino acids.

• None of the sequences produced anything biologically useful.

Appendix 5**. Mutations which produce real novelty.**

In some studies evolutionary change appears to be remarkably easy. For example, Hall reports[19] that strains of E. coli developed the ability to catabolize β-galactosidase sugars after the lacZ gene was knocked out. This is indeed a remarkable finding.

Evaluation.
• A fortunate mutant was provided overwhelming advantages under a lab setting since the wild type is denied necessary nutrients.

• The geometric details around the enzymatically cleaved β-glycosidic bonds of the sugars tested are very similar[19]. In other words, no increase in complexity is developed, but rather variants on an already working theme are selected for.

• It is one thing to select for a mutant in a carefully crafted environment in which one to three mutations only are necessary and a very similar function is already fully operational. Building up things like the arabinose operon itself (with 8 genes) in this manner from scratch under natural settings offers statistical barriers of a totally different dimension.

• Everything needed biochemically is already present, a minor change at the reactive site is usually all that is necessary. In the report selected[19], the previous unknown gene **ebgA** was already present on the bacterial chromosome, and so was its fully functional repressor **ebcR**. *E. coli* can already generate a virtually identical β-galactosidase (especially similar in terms of the relevant reactive site) by lacZ. Either of two single base pair mutations was sufficient to restore lactose hydrolysis once lacZ was knocked out. One of these mutations also permitted a novel lactulose hydrolysis. Both mutations were required for galactosyl-arabinose hydrolysis.

The ebcR repressor was easily deactivated through insertions of IS sequences. And all the components to break down the resulting monosaccharides are already present in the cell.

• The true meaning of the experiment is open to interpretation. What is the function of **ebgA**, which is so conveniently present? Was this Designed to provide robustness, the chance for an alternative enzyme to save a bacterial population? Does lack of nutrient initiate directed mutations at key hot spots, increasing the chances of survival in the presence of other carbon and energy sources? Why does **ebgA** also require the product from **ebgC,** although no other known β- galactosidase do?

• The best Class I Ebg enzyme developed[19] under guided laboratory selection possessed about 0.7% of the E. coli LacZ gene functionality. Without a strong enough selective advantage in vivo it is unlikely that further selection could fine tune the function much more. This brings up the question, as to why so many natural enzymes have attained difusion

controlled perfection, and how random evolutionary trials managed to avoid starting mutational trajectories which cannot lead to optimal enzymes.

Appendix 6. **Some proteins are actually sequentially far apart**.

Damaged (e.g.,. via oxidation) or misfolded proteins are removed from cells. And the concentration of other proteins are altered rapidly with stage of a cell cycle, as needed for present metabolic processes.

The half-life of some proteins may be but a few minutes, such as for mitotic cyclins which help regulate cell passage through mitosis. Or as long as the organism lives, such as the case of proteins in the lens of an eye[<20>]. Chains of ubiquitin are attached to lysines on a protein which is to be degraded. This requires the aid of 3 distinct and very complex enzymes: the *ubiquitin-activating enzyme*, E1; a *ubiquitin-conjugating enzyme*, E2; and *ubiquitin ligase*, E3[<20>], in addition to the proteasome which is a molecular machine consisting of many protein subunits, including multiple proteases and at least 10 types of polypeptides at the entrance[<21>, <22>].

Various enzymes recognize different degradation signals in target proteins, such as: Arg-X-X-Leu-Gly-X-Ile-Gly-Asx and regions enriched in proline, glutamic acid, serine, and threonine (PEST sequences) [<20>]. Furthermore, the particular amino acid found at the N-terminus determine the proteins' degradation time, the ***N-end rule***[<23>, <24>]. Short-live ones degraded within 3 minutes generally have Arg, Lys, Phe Phe, Leu or Trp at the N-end. Cys, Ala, Ser, Thr, Gly, Val and Met resist degradation for more than 30 hours[<20>]. All eukaroyte proteins have methionine initially as the N-terminal residue. Methionine aminopeptidases will remove this amino acid only if the second amino acid is stabilizing according to the N-end rule[<25>]. Furthermore, certain destabilizing amino acids, such as aspartate and glutamate, need to be first modified by the enzyme arginyl-rRNA-protein transferase[<25>].

The multiple ubiquitin chains attached to the protein to be degraded are left intact in the proteosome[<24>], and are released for subsequent further use.

Might such a system have developed by neo-Darwinian methods? The apparatus is ingenious and highly effective, once operational. But it surely did not arise by millions of small stepping stones, leading to ever increasing complexity. A newly evolving method to degrade proteins is potentially deadly, and until fully regulated would surely destroy countless lineases and be selected against.

Natural selection would have had nothing to work on. Ubiquitin without the necessary enzymes is worthless. The protease complex requires many parts to be fully functional, and it must recycle intact ubiquitin for reuse. While natural selection is suppose to be fine-tuning enzyme turnover through increased specificity, one needs to argue that at the same time the degradation of enzymes themselves, leading to decreased concentration and therefor metabolite thoroughput, was being selected for.

The N-end rule is problematic. Before the degradation pathway would have existed, a large number of proteins would have been present with each of the 20 possible residues distributed randomly in the first three positions. As soon as the necessary proto-enzymes became

available, some of the "correct" protein's half-lives would have been shortened along with some of the "wrong" ones. Natural selection has no guidance what to focus on.

Since all eukaryotes possess almost identical ubiquitin sequences, the evolutionist must argue there was a single-cell common ancestor for which the whole process was already fine-tuned. This means that individual evolutionary improvements had to be fixed sequentially through the vast single-cell population world-wide. Let us consider only one component of this scheme, the ubiquitin protein. A blast search[<26>] was performed, and all candidates collected. Incomplete annotations (e.g., unknown organisms) and post-translational modifications forced some of the data to be discarded. This lead to non-redundant sequences for ca. 160 organisms. Optimal sequence aligment with ClustalX[<27>] showed that to date **28** residues (plus methionine) are invariant; and in 11 organisms only one example of a modified residue was found for a specific residue position. Virtually all known data for plants, animals and yeast life forms showed no more that three residue differences from the consensus sequence, the best candidate for a single ancestral common protein.

Since presumably thousands of other proteins were able to diverge widely in far less time, this suggests that ubiquitin must have almost exactly this structure. If the sequence cannot now diverge, the evolutionary options for its putative original development via random mutations would be very constrained. Where did the ub domain come from? Surely not a gigantic coincidence, since obtaining 28 correct residues (each with 1/20 chance) at the exact right position, borders on the miraculuous. But what other gene with an unrelated function might have served as a starting point (along with all the other necessary enzymes)? And where did that ancestral gene come from?