The Problem of Information for the Theory of Evolution

Has Tom Schneider Really Solved It?

Royal Truman

Abstract

In several papers genetic binding sites were analyzed using a Shannon information theory approach. It was recently[1] claimed that these regulatory sequences could increase information content through evolutionary processes starting from a random DNA sequence, for which a computer simulation was offered as evidence. However, incorporating neglected cellular realities and using biologically realistic parameter values invalidate this claim. The net effect over time of random mutations spread throughout genomes is an increase in randomness per gene and decreased functional optimality. Structurally and quantitatively invalid scenarios characterize such evolutionary simulations as will be demonstrated here.

Background

Living organisms undergo non-random physical and chemical processes with apparent purpose, behavior not typical of inanimate matter. The growth of a seed, repair of a wound, digestion, replication of cells and so on, are performed reproducibly, with machine-like accuracy, and are necessary for survival. The scientist and layman recognizes at least intuitively the existence of 'information' as driving chemical and physical processes in manners necessary for life to be possible.

Since mutations randomize a genome over time, the question arises how a genetic code to store and process guiding information could arise. Further increases in specified complexity, as required of evolutionary models, to generate novel biological structures and chemical processes pose a major difficulty. Dr. Schneider, professor Dawkins and other evolutionist allies have chosen a biological example, the fine-tuning at DNA and RNA binding sites, and offer a computer program[1] as evidence that the natural, randomizing course of events might be overcome.

We shall examine the algorithm offered. One could write a computer program which "shows" that random natural processes would drive rocks from a quarry up a steep mountain in thousands of discrete steps, for every simulation run. One only has to use an unrealistic number of earthquakes and improperly model the effects not leading to our intended programming goal. The details matter very much to determine the true net outcome, as we shall find with the program[1] I am going to discuss. Overlooked details in such flawed simulations might not be obvious. Vast number of unrealistically hard earthquakes would affect not only the movement of our rocks but the surrounding mountain would be systematically destroyed.

A legitimate simulation must reflect what is being modelled with sufficient accuracy to justify decisions. Analyzing the financial feasibility of investing in an industrial project under various scenarios would be invalid if income streams are falsely represented and not all costs are taken into account. ***The true expected net outcome depends on the details***. The intention here is to provide some pedagogical guidance as to key factors which need to be incorporated realistically in developing models for evolutionary processes, to test the plausibility that random change plus reproductive advantage could explain various biological observations.

A binding site consists of a sequence of bases (A,C,G,T/U) which serve as addresses or locations on DNA or RNA which specialized recognizer proteins can identify and bind to. Too short a sequence, perhaps AC, would lead to many false addresses. Requiring that a very long, specific sequence (AACAGTCGGTATC.. TGGATCTA...) be 100% correct would not be robust against mutations. Hundreds of binding sites have been identified, such as[3]:

Organism	**Regulatory Protein**	**DNA Sequence Recognized**
Bacteria	Lac repressor	AATTGTGAGCGGATAACAATT
Yeast	GAL4	CGGAGGACTGTCCTCCG
Drosophilia	Krüppel	AACGGGTTAA
Mammals	GATA-1	TGATAG

Every position in the binding site need not have a specific base 100% of the time to permit correct identification of a binding location. Ambiguity introduced by such inexactness can be compensated for by lengthening the sequence.

Dr. Schneider writes[1] that, '*The **ev** model quantitatively addresses the question of how life gains information, a valid issue recently raised by creationists (R. Truman, http://www.trueorigin.org/dawkinfo.htm; 08-Jun-1999) but only qualitatively addressed by biologists)*'.

Mutations of an artificial "protein" were simulated with a computer program[1]. '*The simulation begins with zero information and, as in naturally occuring genetic systems, the information measured in the fully evolved binding sites is close to that needed to locate the sites in the genome.*' '*The purpose of this paper is to demonstrate that $R_{sequence}$ can indeed evolve to match $R_{frequency}$.*'

Caution. The reader must be warned that **the simulation cannot be mapped to a real biological scenario**. *'A small population (n=64) of "organisms" was created, each of which consisted of G = 256 bases of nucleotide sequence chosen **randomly**, with equal probabilities, from an alphabet of four characters (a, c, g, t).'*

What might these 64 living and reproducing organisms, with a total and unchangeable genome 1/4 the size of one typical gene, be? Careful examination of the characteristics assumed in the simulation and references demonstrate these cannot be single nor multiple cell life forms, nor virus nor any known organism. This prevents any kind of model validation.

We read later on, *'Given that gene duplication is common and that transcription and translation are part of the housekeeping functions of all cells, the program simulates the process of evolution of new binding sites from scratch.'*

Lets give this a little thought.

Of course, no attempt was made to show where these miniscule organisms with full transcription and translation machinery came from, nor does the simulation address the production of new genes in any manner. Let us play with the thought experiment anyway.

If such an ancestor, with a genome even smaller than the current 256 bases were to duplicate a "gene", it would waste energy and available material producing unnecessary extra protein during its lifetime and while duplicating its genome. Replication time would be longer than for its competitors and would have greater risk of failure. Even presently unnecessary DNA ballast needed for evolutionary trials and error to produce only a novel binding site represents a significant reproductive disadvantage. This worthless material would represent several percent of the 256 bases assumed for the genome, a very considerable handicap.

It is known [145] [146] that especially small genomes can shed chunks of unneeded DNA rapidly given that these members out-reproduce their competition.

Since sexual reproduction is not meant, let us use the known mathematics of budding or binary fission reproduction[144]:

$$x = \frac{x_0 e^{(st)}}{1 + x_0[e^{(st)} - 1]}$$

Suppose that only 1 of the 64 organisms eliminated or did not originally have a significant portion of junk not needed at the time and the remaining 63 continued trying to evolve a new binding site. Suppose that instead of proportionally 10 minutes generation time on average the streamlined member and ancestors now reproduce 10 seconds faster. The advantages of needing less energy, nutrients, less risk of interference with on-going cellular process, etc. we'll approximate by using a selectivity factor s= 0.0167 (based on 10 / 600 seconds shortened generation time). Since x_0 = 1/64 = 0,0156, we obtain from the formula above:

- After only 500 generations we do not have 64 members with superfluous DNA to evolve a new binding site, but one 1 survivor!
- After 680 generations her chances look dim: > 99.9% of the population no longer has the necessary DNA material for evolutionary experiments.

Obviously, an unnecessary gene duplication to provide DNA to experiment on would introduce proportionally far more worthless DNA than needed for a binding site. Those not suffering such a fate would be at yet a great reproductive advantage. This is a critical oversight in the simulation(1) which invalidates the whole exercise.

Alternatively, a more complex, free-living organism might tolerate an unnecessary gene better, but such creatures could not possibly survive the mutation rate assumed[1], 1 base per 256 throughout the whole genome, every generation.

For many organisms it appears that about 30% of the predicted proteins are unrelated to others in its own proteome or that of other organisms[4] and gene duplication is a rare phenomena commonly identified with various destructive disorders(1).

Overview of Flaws in the Model

Several flaws in the simulation disallow the conclusions claimed. We read, *'Then we need to apply random mutations and selection for finding the sites and against finding non-sites. Given these conditions, **the simulation will match the biology at every point**.'* [emphasis added]. This claim will be shown to be incorrect. Objections #1 - #6 document that biologically unrealistic parameter values are assumed by the computer program, which render any claims that binding sites could develop by chance invalid. We next establish that the model does not simulate random evolutionary processes (objections #7 - #26) in any biologically reasonable manner.

Biologically unrealistic parameter values are assumed.

Objection #1: **The mutation rate is unrealistically high.** *'At every generation, each organism is subjected to one random point mutation in which the original base is obtained one-quarter of the time. For comparison, HIV-1 reverse transcriptase makes about one error every 2000-5000 bases incorporated, only 10-fold lower than this simulation.'*

This is a remarkable statement in light of what the authors referenced[5] actually wrote: *'Our finding, that a limited number of mutations in the HIV genome after exposure to 5-OH-dC has a disproportionately large effect on viral lethality, substantiates the concept that the **mutation frequency of HIV is close to the error threshold for the viability of the quasispecies**.'*[6] [emphasis added]. References supporting this view were supplied[7] [8]. Indeed, *'most HIV virions in the blood appear to be nonviable.'*[9] The virus can only exist due to the huge number of HIV-1 copies produced in an infected individual, about 10^{10} virions per day[10] [11], and hardly 64 members as in the simulation!

Since *'transcription and translation are part of the housekeeping function of all cells...'*[1] and the 64 organisms supposedly survive autonomously, it becomes increasingly mysterious what these creatures with such a miniscule genome and unheard of mutations rates could possibly be.

There are reasons why these self-destructive mutation rates, which would rapidly accumulate, don't occur in the biologically relevant double-stranded DNA: *'In particular, E. coli DNA methytransferase, formamidopyrimidine-DNA glycosidase, and endonuclease III fail to repair efficiently altered substrates when present in the DNA strand of an RNA-DNA hybrid.'*[5]

When DNA is replicated, copying errors occur at about one per 10^8 to 10^9 nucleotide sites[8]. Since in the article[1] it is claimed a billion years would be sufficient for humans to evolve (presumably from some eukaryote-like, non-parasite organism), we need to postulate that a proto-yeast like organism is being alluded to. Let us see where this takes us. For the 13,478 kb yeast (S. cerevisiae) genome[12] comprising about 6217 ORFs (Open Reading Frames), ca. 40% of the ORFs (i.e., 2497[13] at the $1X10^{-10}$ p-level) have pressumed orthologues with the simplest multi-cell organism known[13] [14] and such proteins appear to be critical for survival. Then extrapolating backwards, the evolutionary common ancestor would have had a genome of at least 5.4 megabases with perhaps 2400 genes critical for survival, i.e., ***having virtually no room for error***.

The mutation rate of 1/256 used by the simulation indicates that proportionally 21094 random mutations per proto-yeast member on average would have occurred each generation! Many genes would be hit by 10 or more errors every generation (and the errors would multiply during somatic cell replacement in multicellular life forms during the following billion years). Error catastrophe would be inevitable.

In the simulation all of these mutations are dedicated to a single goal. This implies proportionally for the proto-yeast that 21094 mutations are dedicated to fine tuning one specific binding site every generation in every member, and all the individual point mutations are assumed to be flawlessly recognized by natural selection. Were this even remotely true one could easily dispense of simulations and offer empirical evidence.

A spore-forming bacterium from the permian Salado Formation considered to be 250 million year old within the evolutionary dating framework was reported recently to have a complete 16rDNA sequence of 99% similarity with current Bacillus marismortui.[15] This would indicate a base-pair substitution rate $< 10^{-10}$ per site per year, incongruent with the rate chosen by the simulation.

Objection #2: **The proportion of selectively useful single point mutations assumed is unrealistically high.** The computer program used a twos complement "points" scheme, assigned to each of the 4 nucleotides for each possible position within the receptor sequence of length L = 6 (see Table 2). Statistically, a point mutation on a random genome by these arbitrary rules (at either the DNA binding location or the protein represented by the weight matrix) would have the same chances of being positive or negative, to increase or decrease the viability of a genome. The biological statement would be (depending on the tolerance used) that about 50% of ***any*** point mutations would generate an improved new binding relationship the very first generation, starting from a totally random genome, followed by disminishing returns thereafter (the absurdity of this implied assumption should be apparent). The offspring then get flawlessly selected.

Recalling that *'Generation of the weight matrix integers from the nucleotide sequence gene corresponds to translation and protein folding in natural systems'*[1] it is unrealistic to assume half of all possible point mutations on a random genome would automatically allow a "better", exactly L=6 bases long sequence, to be identified: '**Most single-base changes in promoters and ribosome binding sites decrease synthesis by 2- to 20-fold**' (*Mulligan et al., 1984;*

Stormo, 1986).[16] Random sequences, very far removed from a functional one, would continue generating non- functional sequences via random mutations >>99.9999... % of the time, and evolution cannot look ahead to select a suitable candidate.

Cell regulator activities must occur at the correct location, and the simulation badly underestimates the effects mutations have. Dr. Schneider pointed out correctly elsewhere, *'With this theorem in hand we can begin to understand why, under optimal conditions, the restriction enzyme EcoRI cuts only at the DNA sequence 5' GAATTC 3'* ***even though there are 4096 alternative sequences of the same length in random DNA****. A general explanation of this and many other feats of precision has eluded molecular biologists.'*[17] [emphasis added].

Recall that a nucleic acid binding site is supposed to be evolving via random mutations, as well as the recognizer protein. This must result in very precise three dimensional interactions which involve H-bonding, hydrophobic, and other stabilizing interactions. For gene regulatory purposes, at least one more domain must be present in the protein, capable of interacting with the transcription machinery[18]. Finally, the rest of the protein must ensure all parts fit together geometrically by folding properly. These requirements must be met concurrently to a very high level of precision before any kind of Darwinian selection can be invoked, inconsistent with the computer program which assumes instant "improvement" starting from a random genome.

In the discussion we shall see that a realistic estimate for the proportion of minimally functional to totally non-functional ***proteins*** is very small, on the order of 10^{-44}. The proportion of acceptable ***gene*** sequences coding proteins can only be even lower. The simulation would have to stumble on an acceptable sequence of such unlikelihood, beginning with a random state, over countless generations, before any kind of selection could enter into play. The approximately 50-50 chance assumed by the computer program is unjustifiable and gets evolutionary progress off to a roaring start precisely at the point where all evolutionary conceptual models have the greatest difficulty.

Of the vast number of possible folded geometries, a miniscule subset, on the order of 10^{-44}, would even have a properly folded topology[19] [20] [21] [22] [23] within which the recognizer site would have to be developed. Even should half of the 4 bases (A,C,G or T) be acceptable at every position of the mini-gene (to code for a stable folded protein), one expects for the 64 member population in the simulation a probability of roughly

$$64 \times (0.5)^{125} = 1.5 \times 10^{-36} \qquad (1)$$

per generation of obtaining the first candidate mini-protein (which is coded for by only 125 base pairs according to the paper[1]) upon which natural selection would have a chance to start working. Even assuming a generation time of 1 second, in 10 billion years ($< 3.2 \times 10^{17}$ seconds) we'd have essentially zero chance of even getting started.

This objection alone renders the whole exercise meaningless.

Objection #3: **Countless point mutations are assumed to instantly provide reliable binding interactions.** Unlike the fictitious positive and negative integers used in the simulation, in earlier papers the weight matrix was derived using real data on functional sites[10] [143]. Known binding sites were selected from genbank, lined up and the proportion of each of the 4 bases found at each position of a sequence was determined (see Appendix).

Binding of a protein to DNA or RNA is rarely the simple matter implied by the computer program, but generally requires cooperation with other carefully crafted proteins[11]. For example, transcription in eukaroytes is regulated by a group of gene-specific activator and repressor proteins[24] at specific binding sites. Simulating the production of one recognizer member of such ensembles by random point mutations has not been justified nor validated as being biologically conceivable. Instead, an arbitrary proportion of positive and negative integers in the computer program defined how to converge towards a short term goal flawlessly irrespective of any biological selective significance or stochastic effects.

How is chance to know a random mutation would lead towards developing a binding interaction? *'$R^{sequence}$ does not tell us anything about the physical mechanism a recognizer uses to contact the nucleic acid.'*[25]

Lacking any intelligence to choose, 3 dimensional shapes on the regulatory protein must be generated to permit the exact binding with a specific DNA sequence, like a well-meshed machine. That is why a methionine-carrying tNRA is able to identify a very short sequence on mRNA, AUG, and position a physically large m-RNA properly at the ribosome complex: it is due to the specialized geometry prepared at the ribosome's P site. There is nothing biologically remarkable about AUG alone. Crystallographic, molecular modelling and cryo-electron microscopy studies have shed insight as how such feats are possible. Translating an mRNA strand one codon at a time requires the whole ribosome complex to act in a synchronized fashion, aptly described as a rachet-like mechanism[26]. The cell's survival depends on ribosomes being able to locate the binding sites correctly[27][12].

Exactly how polypeptides are supposed to be able to identify that a location is or will become a useful binding site is deemed irrelevant: *'As mentioned above, the exact form of the recognition mechanism is immaterial because of the generality of information theory.'*[1] Quite the contrary, for a realistic evolutionary simulation such physical details are critically relevant, and is a fatal oversight in the simulation. It is assumed random point mutations provide half the 64 member population with a 100% effective survival advantage, based on fine tuning of a single type of binding site under development. This is geometrically and thermodynamically unrealistic. Developing such precise binding interactions, one random mutation at a time, has nothing to do with the mathematics of information theory and needs to be quantitatively simulated based on physical realities. Any assumption of recognizable Darwinian selectivity for the intermediate stages needs to be quantitatively justified.

The requirements on recognizer and binding site are generally very stringent a must be close to perfect to be of any use whatsoever[13].

Objection #4: **The rate of selection is unrealistically high.** A standard textbook on cell biology reports[28] the average times evolutionists assume are needed for one acceptable amino acid change per 100 in specific proteins. The fastest rate reported within the evolutionary model required 0.7 million years for fibrinopepetide, and the slowest was for Histone H4, with 500 million years. In another place we read, *'only about one nucleotide pair in a thousand is randomly changed every 200,000 years.'* [29]

This is incompatible with Dr. Schneider's claim that his simulation *'is within the range of natural population change.'* The computer program required only 704 generations to create a new binding site type at exactly 16 positions on a genome with its novel recognizer protein from scratch. After adding up all the "points" at each possible binding site using the current weight matrix (and a cutoff score of -58), the 32 members scoring lowest (selectivity s ≈1 !) magnanimously discontinue their ancestors' eons of hard evolutionary work. The half having the less desirable status, due to a single point mutation, get pin-pointed every generation 704 times in a row without exterminating the future of higher life forms.

Since real world selection coefficients (based on major mutations and not mere single point mutations) are proposed to be on the order of 0,01[30] or less, one would expect some justification for the 100-fold greater rate chosen. Simpson felt s=0,001 may be too low, but 0,01 could be taken as a "frequent value" (i.e., might occur now and then).[31] Artificial laboratory settings or antibody resistance in hospital settings (with necessarily much larger population sizes to avoid killing the population off) are not representative of a natural setting relevant to an evolutionary scenario.

A small and non-growing population of 64 members was chosen for the simulation. Fisher's analysis showed that a selection coefficient even as great as S=0,1 would have only a 2% chance of fixing in a population of 10,000 or more.[32] Lacking is the justification how the population would be limited to 64 members for at least 704 generations. I demonstrated above that instead of having 64 members with superfluous DNA to tinker with, long before 704 generations we'd have none at all. Presumably these organisms are submitted to catastrophic environmental conditions to justify the maniac selectivity coefficient implied, but the simulation disallows the possibility of failure, that a generation might not pass on viable progeny(7).

Objection #5: **Degeneracy of the genetic code, sexual dilution, and other factors are ignored.** The degeneracy of the genetic code has been neglected. A protein is being represented by the weight matrix, and an approximation (Table 3) suggests that on average roughly 24% of all point mutations would generate the same polypeptide starting from a random DNA sequence (this assumes for estimation purposes that mutational transitions and transversions are all statistically the same). With about 1/4 mutations producing the same amino acid, the credibility of the scoring assumptions is further strained.

Given the close correlation between number of synonym codons and proportion of corresponding amino acid present, the genetic code may have been designed partially to help retain protein functionality by protecting against point mutations(9).

Recessive mutations and dilution of point mutations by sexual reproduction are not considered although it is claimed the rates in increase in Shannon information content can be quantitatively extrapolated to explain the origin of the human genome[1].

Objection #6: **The final state is not stable.** Having somehow achieved the miraculous, how long might these organisms manage to stay balanced on Mount Improbable? '*When selective pressure is removed, the observed pattern atrophies (not shown, but Fig. 1 shows the organism with the fewest mistakes at generation 2000, after atrophy) and the information content drops back to zero (Fig. 2b). The information decays with a half-life of 61 generations.*'[1]

This confession closes the case decisively. Removal of what amounts to an intelligently driven selection allows the genomes to randomize rapidly. No naturally stable increase in Shannon information has been demonstrated.

This outcome is to be expected due to the catastrophically rapid mutations assumed with the guarantee the population will not perish. The highly contrived mathematical characteristics describing this small population has no resemblance to the multiple survival challenges real organisms face in nature.

The model does not simulate random evolutionary processes.

Objection #7: **Foreknowledge of the sequence length, L, is provided to the computer program.** Evolution somehow knows that one, and only one, binding site type only, of length exactly 6, is to be developed. In Table 1 we summarize some examples of binding sites with L ranging between 4 to 51 bases. Some recognizers, such as the H-NS protein, can interact with binding sites of various lengths, and is affected by the protein's concentration.

In addition, binding sites need not be contiguous, there may be spacers between conserved portions of the same binding site[2]. A legitimate simulation needs to consider competing sequences of at least L= 4 to ca. 51 ***concurrently*** with no foreknowledge as to an intended outcome: survival and increased reproduction rate can have multiple causes and cannot be simply attributed to the change we wish to favor. If a randomizing process can go in every direction at once, so be it.

Objection #8: **Foreknowledge of the required number of binding sites, λ, is assumed.** Although the author claims chance can generate binding sites "from scratch", finding the necessary number of a new kind of site through biological trial and error, in the face of multiple survival challenges, has not been simulated: this number was conveniently provided to the computer program[3].

Current physiology implies something already functional, which can hardly be a random starting point. A multitude of unrelated types of binding interactions exist to regulate genetic control elements and evolution cannot know in advance the necessary number of even one of them. We read,

'The bacterium Escherichia coli has approximately 2600 genes, each of which starts with a ribosome binding site. There have to be located from about 4.7 million bases of RNA which the cell can produce. So the problem is to locate 2600 things from a set of 4.7X10^6 possibilities, and not make any mistakes. How many choices must be made? The solution to this question, log$_2$(4.7 X 106/ 2600) bits, is "obvious" to those of us versed in information theory...' [=$R_{frequency}$ = 10.8 bits][27]

Beginning with random base and recognizer sequences, the binding interaction is to be optimized and converge on the needed number of bits of Shannon information to uniquely identify, but not unduly overspecify, an ensemble of addresses or locations on the genome. However, at any point in time the computer program already "knows" what $R_{frequency}$ value is biologicaly needed and no attempt was made to simulate a trial and error process of finding this value over many generations. Evolution has been provided with foreknowledge.

Random sequences on DNA will not interact reliably with random polypeptides in any biologically sensible manner: the selection being provided is strictly a mathematical artifact, nothing real is being simulated. Random mutations cannot know in advance that an $R_{sequence}$ of 4 and not 2.8 or 20.3 bits needs to be converged on, to permit λ= 16 binding sites to be located reliably.

Objection #9: **Binding sites must be correctly located with respect to the genetic element being regulated.** Real binding addresses must be judiciously placed at suitable locations on DNA to permit specific cellular processes to be regulated. This is very different than merely having the correct number, λ, of binding sites. The correct sequence at the wrong place can cause havoc, and the trial and error process to get these placed corrected was not simulated: *'Level 1 theory explains the amazingly precise actions taken by these molecules. For example, the restriction enzyme EcoRI scans across double helical DNA (the genetic material) and cuts almost exclusively at the pattern 5' GAATTC 3', while avoiding the 4^6 - 1 = 4095 other 6 base pair long sequences. How EcoRI is able to do this has been somewhat of a mystery because conventional chemical explanations have failed.*[27]

A trial and error process would have destroyed countless organisms and removed such destructive evolving machinery long before stumbling on a working scheme[(4)],[(4b)]. Obtaining a suitable number of recognizers to find a matching binding site is a necessary but insufficient cellular requirement. We read, however,

'... as a parameter for this simulation we chose γ=16 and the program arbitrarily chose the site locations.'[1]

Even a perfectly functional binding site cannot simply be placed anywhere to regulate a specific gene! (We'll ignore the matter of where these genes being regulated came from in the first place, and whether they would function without the regulatory elements still to be evolved.) However, the twos complement scoring scheme[1] permits the computer program to pick binding locations arbitrarily, and calculates "points" based on how well the evolving sequences and regulating element match up. It is simply assumed the non-binding portion of

gene being regulated. No trials and errors are simulated to get a proper binding sequence located at an acceptable distance.

Objection #10: **Selection is intelligently driven.** Careful reading reveals not a simulation but a designed convergence algorithm. The two matrices of numbers, plus a tolerance score, define goals which can change slightly across generations. The immediate goals are instantly known and flawlessly acted upon by the computer program, with no consideration to survivability uncertainties. By retaining half the highest of 64 scores every generation the process of being intelligently guided. The 2 matrixes converge far more quickly than random changes are allowed to separate them. The rules established are:

'A section of the genome is set aside by the program to encode the gene for a sequence recognizing "protein", represented by a weight matrix consisting of a two- dimensional array of 4 by L = 6 integers. These integers are stored in the genome in twos complement notation, which allows for both negative and positive values... By encoding A = 00, C = 01, G = 10 and T = 11 in a space of 5 bases, integers from - 512 to +511 are stored in the genome... Each base of the sequence selects the corresponding weight from the matrix and these weights are summed. If the sum is larger than a tolerance, also encoded in the genome, the sequence is "recognized" and this corresponds to a protein binding to DNA.'[1]

I worked out the scoring matrix according to the rules[1] used by the simulation (Table 2). Should this not be correct I hope for clarification. A binding site of L=6 could score between 6 X 512 and -6 X 511 "points". For λ=16 sites, the range for a genome falls between -49056 and +49152. The 32 high scorers always kill off exactly the 32 low scorers with enviable military precision and no collateral damage. A one point difference can flawlessly make the difference between life and death, no stochastic effects are allowed. Incredible as this level of selection appears to be, *'To preserve diversity, no replacement takes place if they are equal.'*[1] Evolution has been granted skills beyond even Maxwell's Demon: a correct choice is made between entities which are quantitatively indistinguishable! The effects on the simulation of this innocent appearing decision has not been discussed[], but the Pascal source code found on the web site[2] does shed some light: *'SPECIAL RULE: if the bugs have the same number of mistakes, reproduction (by replacement) does not take place. This ensures that the quicksort algorithm does not affect who takes over the population.* ***Without this, the population quickly is taken over and evolution is extremely slow!'***[2] [emphasis added].

Identifying the most suitable sequences to serve as binding sites is physiologically not so straightforward, and the biologically optimal binding sites are not always the strongest physically(5). Indeed, nature presents us with many examples of biologically sensible solutions which are unexpected if derived under natural, unguided conditions. For example, Weindel has pointed out[33] that under pressumed Ursuppe conditions (Formosa reaction) many sugars are generated whose nucleotides form stronger base-pairing (A::T; G:::C) than occurs with D(+)Ribose (as determined by melting point studies in his laboratory).

Although thermodynamically preferred when compared to existing RNA-DNA and DNA-DNA interactions, these chemical options are biologically unsuitable since the strands would not be separable as required by cells. Such observations cast severe doubt on the claim enough time and the right conditions suffice to explain the origin of life. An evolutionary process cannot plan for the future and choose to ignore known chemical kinetics and thermodynamics.

Objection #11: **No provision is made for the proportionally greater destructive possibilities.** In an earlier paper we see how sensitive binding sites can actually be towards a single point mutation: *'For example, the **E. coli** genome should contain about 1000 **Eco**RI restriction enzymes sites (G-A-A-T-T-C), but that same genome should also contain about 18,000 sequences one nucleotide removed from an **Eco**RI site. Site recognition by and action of **Eco**RI within **E. Coli** must include enough discrimination against the more abundant similar sites to avoid a fragmented genome.'*[16] Not only is the proportion of useful point mutations unrealistically modelled[1], but the proportion of almost correct (but deadly) to acceptable sequences is very large, and this has not been accounted for in any manner in the simulation.

Other examples include the 6-base TATA box for which a single base mutation drastically damages transcription by RNA polymerase II[34].

Objection #12: **The simulation assumes all organisms in that population face one and the same goal which is to be optimized.** *'The organisms are subjected to rounds of selection and mutation. First, the number of mistakes made by each organism in the population is determined. Then the half of the population making the least mistakes is allowed to replicate by having their genomes replace ("kill") the ones making more mistakes.'*

The nature of this highly focused selection which could drive fine-tuning of a single kind of binding process was not discussed. There are many possible reasons for an organism to die without producing offspring given that each organism faces a variety of challenges. The effects of a mutational proportion of 1/256 bases in the genome could affect any of many cell processes in the real world and will hardly allow optimization for a single and the same goal every generation, driving fine-tuning of one kind of binding site, base-pair at a time.

Should all evolutionary selection be focused on one goal, deleterious mutations elsewhere would not be eliminated. The net effect would certainly not be an overall net decrease in gene sequence randomness.

Furthermore, overlapping binding regions serving unrelated functions by different proteins could not be selectively identified and fine-tuned using this biologically over- simplified single-goal scenario(6). A particular base mutation at one binding site may facilitate recognition by one recognizer, but be selected against since another one has become less effective.

It has been suggested that the existence of higher sequence conservation than needed to locate the same binding type sites implies the existence of other recognizers interacting at those locations also. An Intelligently Designed possibility could be entertained: extra robustness had been built in and that randomization has not proceeded long enough to remove the excess Shannon-type information.

Chauvin has pointed out[35] that fly resistance to a single substance can be developed but cannot occur if the population faces 5 toxic products simultaneously. He believes such claims of co-evolution are merely laboratory artifacts. Whether countless random mutations could be guided by differential reproduction to produce structural novelty is highly improbable. Careful thought could begin with examples where no plausible selective advantage can be offered for either the intermediate steps not final result, such as the fact that some edible varieties of butterfly can mimic in appearance perfectly another species which is perfectly edible[36].

Objection #13: **Binding sites generally require many novel biomolecules to function.** In a recent article[37] the atomic structure of the large ribosomal subunit *of Haloarcula marismortui* was reported. 3045 nucleotides plus 31 proteins are involved. Ribosomes can be inactivated by cleaving of a single covalent bond in the SRL (sarcin-ricin loop) of the 23S rRNA component. As the authors point out, *'[The] ribosome assembly must be accompanied by a large loss of conformational entropy' and 'Of the 2923 nucleotides in 23S rRNA 1157 make at least van der Waals contact with protein.... to immobilize the structures of these molecules.'* Only now can 3 recognizers, which participate directly in the protein synthesis, perform properly at the intended binding sites: *'Rather than being included in the ribosome to ensure that the RNA adopts the proper conformation, it seems more appropriate to view the RNA as being structured to ensure the correct placement of these proteins.'* Precisely as expected if Intelligently Designed.

One cannot first create one protein then start developing the other components afterwards(14) while hoping the first remains intact over time. As an example, UBF activates transcription by relieving repression caused by an inhibitory factor which competes for binding of TIF-IB to the rDNA promoter. This is not left to chance, but UBF can be interfered with by pRb which seems to act as a signal which links the cell cycle with multiple components of the transcriptional machinery[38]. As a rule, several protein interact and these must be able to penetrate the nuclear membrane where gene expression can be regulated.

The simulation makes no attempt to see whether chance could attain a minimum level of functionality to enhance viability and upon which selection could then begin to work.

Objection #14: **Structural features of DNA may serve as relevant or incorrect binding sites.** Where the binding site is located is biologically critical, and there are various possibilities(15). A legitimate simulation needs to mimic the trial and errors needed to identify a binding address and all the attempts to create a useful cellular outcome under the control of such binding interactions.

The same or similar binding sequences on different portions of the genome can produce very different or even contradictory effects, no weight matrix exists *a priori* to guide evolution towards a parsimonious, multi-goal state.

Objection #15: **The same binding location can be used by different proteins to regulate important processes.** How often one point mutation at a single binding site would really lead to a selective advantage when this affects the address multiple proteins use has not been simulated. Sharing or competing at the same or overlapping sites is well known[16].

Objection #16: **The same proteins can affect many unrelated genes concurrently.** For example, mutations in E. coli hns alter the expression of many genes with unrelated functions[39][17]. The same binding protein can interact in different regulatory complexes at the same binding site: mycN/max heterodimers probably activate and max/max homodimers repress transcription of, as yet, unidentified target genes upon binding to the DNA sequence CACGTG.[40]

Objection #17: **The same protein may contain multiple recognizer sites which can be used for unrelated binding purposes.** Should a protein already have been fine-tuned for a specific function, adding *post facto* another recognizer site without interfering with the geometry, folding order and so on of the previous function would require a multitude of random trials. Alternatively, building multiple recognizer sites concurrently creates formidable constraints. The existence of multiple sites is well-established, such as the A/B pocket and the C-terminal domain of pRb[38].

Objection #18: **The same protein can be a transcriptional activator and repressor depending on the gene it acts on.** Examples abound of proteins accelerating transcription of one gene and slowing down that of another[18]. Notice how the program trivializes such realities. The reader is invited to give careful thought as to how many random attempts might be needed until chance mutations were to stumble, through only viable intermediate regulator protein structures, on solutions compatible with the contradictory cell requirements.

Objection #19: **Regulatory proteins need to be transferred to the correct location in the cell.** Regulatory activities can occur in different organelles, and biochemical activities in various portions of a cell can determine whether a protein will penetrate the nucleus and then locate the intended binding site[19]. How this requirement, which many binding sites must fulfill before they can function, could be developed by trial and error point mutations is missing in the simulation.

Objection #20: **Regulation pathways by binding proteins can involve multiple proteins.** Representative examples include: ***Xvent-1*** (a homeobox gene, and goosecoid interact in a cross-regulatory loop suppressing each other's expression)[41]; ***E1A*** proteins (transformation and transactivation are mediated through binding to pRb, p107 and p130, and the TATA box binding protein TBP[42].

Furthermore, regulatory proteins often form symmetric dimers (two identical proteins) or asymmetric ones[18], with each member binding to different regions on DNA (see Figure 1). Notice that symmetric schemes now require duplicate sequences on DNA, both at the correct location, which would have to develop by random mutations while providing biological functionality during the whole process.

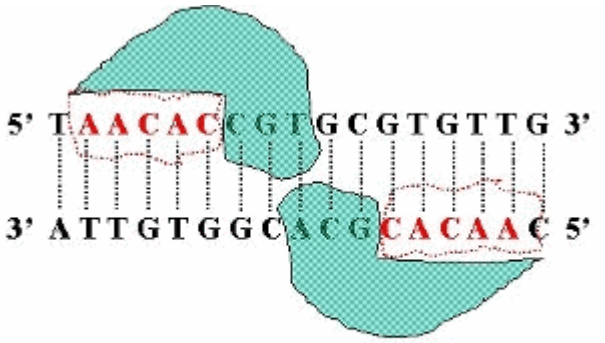

Figure 1. DNA sequence recognized by 2 lambda cro protein monomers. (142) p. 410

Objection #21: **Regulation at binding sites may require fine-tuned interaction with other chemical processes.** One may consider ***SL1***, which is inactivated by cdc2/cyclin B-directed phosphorylation, and reactivated by dephosphorylation. This allows SL1 to work as a switch to prevent pre- initiation complex formation and to shut down rDNA transcription at mitosis[43]. Requirements such as these illustrate the large number of neglected trials needed before a binding interaction can fulfill a minimum functionality.

Cations are also used as part of regulatory signals. *'The restriction enzyme EcoRI is a protein which cuts duplex DNA between G and A in the sequence 5' GAATTC 3'. In the absence of magnesium, binding is still specific but cutting does not occur.'* (17)

Since Dr. Schneider's simulation uses a binding length of L=6, we can consider a well-known process of this length which relies on selective methylation. *'In vivo cellular DNA is protected from EcoRI by the actions of another enzyme called the modification methylase. This enzyme attaches a methyl group to the second A in the sequence GAATTC, so that EcoRI can no longer cut the sequence. In contrast, invading foreign DNAs are liable to be destroyed because they are unmethylated. The methylase is precise, attaching the methyl only to GAATTC and not to any of the sequences, such as CAATTC, that differ by only one base from GAATTC... How a single molecule of EcoRI can achieve this extraordinary precision has not been understood.'*[17]

'*For example, if the restriction enzyme EcoRI did not reliably and repeatably recognize one pattern, GAATTC, the bacterium might die by the destruction of its own genetic material. Likewise, if a DNA polymerase did not reliably insert adenosine opposite every thymidine, many mutations would occur.*'[44]

This scheme could only work after the modification methylase were already present and fine-tuned to attach under the correct circumstances. EcoRI, a binding sequence, and additional components must all be in place within an acceptable tolerance before any kind of selective advantage would be measurable.

Objection #22: **Regulation often needs to be achieved for a specific (or across different) cell type.** Consider as an example ***IL-4*** (inappropriate multi-organ expression leads to autoimmune-type disease in mice)[45]

In gene therapy the administered protein is a less than satisfactory substitute for a protein physiologically regulated by its origination in a specific tissue. That is why injected insulin cannot control blood glucose sufficiently well to prevent all diabetic crises, let alone the slow tissue damage and complications that lead to premature death.[46] However, the simulation[1] assumes chance only needs to generate two regulatory components: the binding site and part of a protein.

Proportions of mRNA generated needs to be carefuly regulated, and can vary considerably according to specialized cell type. For example, the alpha-fetoprotein gene in a mouse results in 200 times more mRNA in the yolk sac than the gut[47]. The regulation of expression is fine-tuned according to cell type. In addition, the pattern of expression of a specific gene can differ significantly depending on exactly where it is placed in a genome.

Objection #23: **Regulation needs to be achieved according to stage in cell life.** As an illustration, ***H-NS*** functions as a global inhibitor of gene expression during the cell's exponential phase of growth[48] The trial and error attempts to be minimally functional has been neglected in the computer program.

Objection #24: **Different promotors can act on the same gene producing isoforms.** These can be tissue and cell dependent(20). The scenario of development by one random point mutation at a time leaves such observations unexplained. As an illustration, six hERalpha mRNA isoforms are produced from a single hERalpha gene by multiple promoter usage. All these transcripts encode a common protein but differ in their 5'-untranslated region as a consequence of alternative splicing. A differential pattern of expression of the hERalpha gene in human tissues and cell types was found.[49]

Objection #25: **An acceptable proportion of regulatory binding protein or complex must be generated and regulated as needed before natural selection can act.** Vastly different levels of protein are found in cells and these change as needed(21). Table 4 demonstrates the wide distribution of mRNA molecules in a typical mammalian cell, which ranges from about 5 copies to over 12,000[47]. An overabundance would prevent fine-tunning of binding sites[39] (21b), whereas too small an amount could preclude enough cellular value to be selectively identifiable(22). Literature abounds demonstrating gene expression cannot be too low nor high(23).

Objection #26: **Irreducible complexity, a fact of cellular processes, is glossed over.** The simulation allegedly *'is representative of the situation in which a functional species can survive without a particular genetic control system but which would do better to gain control* ***ab initio****. Indeed, any new function must have this property until the species comes to depend on it, at which point it can become essential if the earlier means of survival is lost by atrophy or no longer available.*

I call such a situation a "Roman arch" because once such a structure has been constructed on top of scaffolding, the scaffold may be removed, and will disappear from biological systems when it is no longer needed.'

Using lack of evidence as proof for an argument is rarely convincing (such as Punctuated Equilibrium being true due to the lack of transitional forms in the fossil record). Before a particular polypeptide could be available for a new function, evolution is now required to have produced it plus additional components for a preceding use, also by chance mutations. Each individual precursor now also requires an ensemble for a yet earlier functioning complex. This argument requires at best a starting point and at worse merely increases the implausibility of obtaining each needed biochemical component.

Objection #27: **Recognition of binding sites does not cover the miracles evolution is suppose to explain.** We read, *'Second, the probability of finding 16 sites averaging 4 bits each in random sequence is* $2^{-4X16} \cong 5 \text{ X } 10^{-20}$ *yet the sites evolved from random sequences in only ~* 10^3 *generations, at an average rate of ~ 1 bit per 11 generations.'* As pointed out, this was achieved by distorting cellular realities to the point of biological irrelevance. But now an extrapolation is made from what is a relatively trivial fine tuning challenge for evolution to a grandiose claim:

'Likewise, at this rate, roughly an entire human genome of ~4 X 10^9 *bits (assuming an average of 1 bit/base, which is clearly an over-estimate) could evolve in a billion years...'*

This is indeed a remarkable extrapolation! A mutational rate of 1/256 bases on average throughout the whole genome would have to apply to multi-cellular organisms also. Visualize what your child would look like after cell fertilization and the following 50 or so cellular duplications. Almost 0.5% of the bases get scrambled 50 times in a row (recall that a perfectly random distribution of bases on DNA implies a 1/4 chance any base will show up at each position). Should even one gene in a somatic cell remain functional, subsequent cell replacement during the lifetime is sure to wipe it out also. This process is then to be repeated to produce lovely, bouncing grandchildren. A free-living organism would not last very long with such flawed DNA duplication and error correction mechanisms.

Major issues for which no plausible solutions by chance mutations have been offered to date have not even been addressed. Examples include: how sexual reproduction could have arisen; the existence of multi-cellular organisms[30] with specialized cells and integrated functionality; and biological novelty demanding the interaction of large numbers of genes, such as in sonar and sight. Even had a plausible simulation been offered which demonstrated that one type of binding site could be generated ab initio via random mutations, an extrapolation to evolutionarily unexplained and unrelated problems is unwarranted.

DISCUSSION

Evolutionary theories need to account for the creation of novel biological functionality, which includes explaining how new genes might arise. Consulting the Munich Information Center for Protein Sequences, we determine that yeast, the simplest eukaryote cell known, has a sequenced length of 13.5 megabase[12] coding for about **5929** different kinds of proteins, about 30% with no known homologues[50]. The worm Caenorhabditis elegans is the simplest multicellular animal showing complex development and a differentiated nervous system[51] and has 959 cells. Its 97 megabase genome[52] codes for about **19,099** proteins[51] (three times more than yeast). Chervitz et al.[13] compared the sequences of yeast and the worm. Around 40% of yeast ORFs (Open Reading Frames) appear to have counterparts in the worm, and 20% of worm ORFs were found in yeast and seem to be indispensible. Most significant is that 34% of the predicted proteins are found only in other nematodes[51]. Conversely, many important proteins in yeast are not found in the worm[13]. In fact, a large number of domain structures are not shared at all (Table 5).

Somehow a vast number of correct base pair sequences need to be incorporated into genomes without producing chaos. For 'information' as used here[1], an increase in genome size and restriction of genes to subsets of allowable sequences represent increases in information content as defined by Shannon. Fine tuning of one kind of binding site is admitedly a very modest part of what needs to be explained.

One cannot brush off the objections introduced above by reasoning that "the principle is what matters", and the alleged convergence merely requires a much greater number of generations if modelled more accurately. Mutations by nature randomize those acceptable DNA sequences responsible for biologically useful functions. Increase in information content, as defined, to optimize or create new function requires this trend to be reversed. Any claim that random mutations plus selective reproduction would work must be realistically and quantitatively modelled to justify the claims the statistically unexpected trend could actually have occurred.

The problem can be broken down into two components. (a) Random trials would be simulated until all constraints outlined above are satisfied unto the minimum point where reproductive selection could be sensed in a Darwinian sense; (b) thereafter, additional trials would be simulated where selection, unguided by a long-term goal, would increase favorable mutations throughout a population. The simulation neglects aspect (a) entirely, by assuming extraordinarily fast mutations rates and proportions of useful point mutations available to random sequences, which would instantly be selectively acted upon with no possibility of extinction. Let us identify some of the minimal constraints binding sites must satisfy before any kind of selection were to be possible.

Since one or more proteins will be involved in binding to a portion of nucleic acid polymer strand we need a realistic probability of getting a protein with minimal functionality. We begin with random amino acid sequences, since evolution eventually starts with a biologically non-functional state. Yockey has done extensive calculations using Shannon's information theory on the cytochrome c family:

'Cytochrome c is the best candidate for the first application for a number of reasons. The list of sequences reported in the literature includes the largest number of species for any protein and also covers a wide range in the taxonomic scale.'[53] (24) It is possible additional synonyms could be tolerated, resulting in lower Shannon information.

Cytochrome c seems reasonably representative of presumably very ancient genes '*We find in Dayhoff's list (1978) that proteins which are regarded as ancient or even precellular such as certain domains and structure of glyceraldehyde 3-PO4 dehydrogenase, lactate dehydrogenase, glutamate dehydrogenase ferredoxin and the histones have a mutation rate which is nearly the same or smaller than that of cytochrome c. It is therefore reasonable to believe that they have the same or larger information content.'*[53]

Other studies confirm the intersymbol independence of protein residues and restricted number of functional members within a protein family[54].

To ensure we are not being too demanding, let us assume that all the possible varieties of cytochrome c would have been usable by the first organism in which it supposedly first evolved. This is unlikely to be true. Yockey has pointed out that *'Fitch & Markowitz (1970) have shown that as the taxonomic group is restricted the number of invariant position increases.'*[53]

To the list of all currently available sequence data, Yockey generously added all amino acids which might be tolerated by cytochrome c at each position. This allowed him to calculated[147] via Shannon's information theory the number of minimally functional cytochrome c members. He also calculated the total number of polypeptides sequences 110 residues long (excluding sequences very unlikely to be generated, having many residues rarely used in nature).

His work shows that for every functional member, random mutations would have to generate and test

$$5 \times 10^{43} \qquad (2)$$

non-functional variants.

We can now evaluate objectively the claim[1] that 64 random genomes could produce a novel binding site, with regulator protein, in 704 generations by random point mutations, starting from total random sequences. Furthermore, I have already pointed out that within just a few generation we would no longer have all 64 members with necessary but presently superfluous DNA material to develop the new binding site.

This is the estimated proportion of minimally functional to worthless residue sequences for the best studied protein to date. Is this proportion unduly small for genes overall? For histone H4, alcohol dehydrogenase or glyceraldehyde-3-phosphate dehydrogenase it is orders of magnitude too generous[57] [58] and other considerations suggest comparably infinitesimal proportions must be overcome on average to produce new proteins before selection could begin to fine-tune(25).

It is noteworthy that some domains, which are highly invariant but key portions of proteins which interact with specific DNA sequences, although only part of the protein are alone larger than cytochrome c[59]: POU (~160 amino acids), CTF DNA binding domain (132 acids) and CFT proline-rich domain (143 amino acids). Proteins containing such domains must not only be properly folded but possess additional functioning domains to be of any biological use.

Vague references to co-evolution using existing parts for new purposes merely shifts the problem elsewhere. If the odds of obtaining that protein, but for a different ancestral purpose, is similar then nothing has been solved. One merely introduces additional difficulties, such as the need to explain how that protein and the members of the earlier function arose. Should n = 5 structurally unrelated genes be involved in the preceding function, then the odds of obtaining a functional ensemble becomes a number such as[2] raised to the 5^{th} power. Thereafter one needs to demonstrate there is a viable path accessible by random mutations which can connect the preceding ensemble of components with that protein's new use, and that all evidence for the ancestral complex was then conveniently eliminated.

Experimental studies on acceptable sequences based on protein folding by Sauer using arc repressor[19] and lambda repressor[21] suggest Yockey's estimate[2] is far too generous, at least for average-size proteins. Sauer estimated that about one out of 10^{65} of polypeptides he studied are able to fold properly (one of many requirements for useful proteins), a number Behe[22] has compared to successfully guessing a grain of sand in the Sahara desert three times in a row. Let us accept all known evidence and accept that a very small proportion of polypeptides would be biologically useful. Let us tentatively accept 5×10^{43} as representative also for DNA base sequences. The simulation[1] needs to account for the generations needed to produce the first minimally functional protein from a random sequence.

Assuming a generation time of only 10 minutes for a billion years would provide $10^9 \times 365{,}25 \times 24 \times 6 = 5.3 \times 10^{13}$ generations. A population 64 members on average would have a chance on the order of

$$\mathbf{6.8X10^{-29}} \qquad (3)$$

of stumbling on one minimally acceptable gene sequence before selection could start optimizing a binding interaction (the need for it to also have binding sites and be minimally regulated has been neglected). This assumes all point mutations could occur and do not concentrate on a limited number of hot spots[60].

It is apparent that 704 generations in total could not possibly suffice for 64 descendants of 64 random genomes to produce a novel binding site optimally at 16 locations as claimed[1]. This illustrates the fact that the simulation[1] is not dealing with anything biologically relevant. The process has essentially zero probability of even getting started.

Having one minimally functional gene, a realistic computer model would next simulate competition between degradation of this sequence and developing a novel binding site.

Since evolution cannot look ahead, multiple sites of varying lengths L, generated by random point mutations, must be tested by trial and error concurrently, each with a full complement of regulatory elements. Suitable point mutation selectivities and population genetics assumptions need to be identified.

What kinds of odds are faced in developing just one of the hundreds of already identified DNA recognition sites[3], each used by a different specific gene regulatory protein (or set of regulatory proteins), starting from scratch using random point mutations? In total an eucaryotic cell has thousands of different gene regulatory proteins. A realistic simulation must mimic the process of satisfying multiple requirements by the recognizer protein and DNA/RNA binding site at a minimum level of functionality before any kind of Darwinian selection could be assumed. Some of the factors neglected by the simulation under consideration have been identified.

In **Stage 1**, a true simulation would run through random trial and errors attempting to satisfy several constraints before selection could be invoked. Any and all forms of selection for any biological purpose which would increase the proportion in a population is meant here, not only with respect to a specific future binding interaction. Once all constraints are met, the second stage, with selective advantages, would be simulated.

Proportion of recognizers before selection of any kind would be measurable	
P_1	With an acceptable stable tertiary structure
P_2	With an acceptable recognizer site
P_3	Generated reliably within an acceptable concentration range
P_4	Not interfering with other genetic processes (repressor vs. activator)
P_5	Transferred to the correct cellular compartment
P_6	Located in the correct cell type of multicellular organisms
P_7	Acting during an appropriate portion of the cell life
P_8	With at least one additional functioning domain besides for the binding site
P_9	With minimal operational regulation such as by phosphorylation, cations, methylation, etc.

Proportion of binding sites before selection of any kind would be measurable	
P_{10}	With acceptable binding length, vs. L=4 to ca. 51 unaccepable alternatives
P_{11}	With suitable base sequences for each particular length, L
P_{12}	In an acceptable location with respect to genetic elements to be regulated
P_{13}	In an acceptable concentration range in the genome
P_{14}	Biologically compatible with already existing recognizers.

Only now does selection become relevant for any organisms meeting all constraints.

In **Stage 2** the same considerations apply but the proportion of better to lesser tuned possibilities decreases steadily. The possibility of overall loss of Shannon information content in the genome by decrease in gene specificity via random mutations must be permitted in such a simulation. Thus, if a very high mutation rate is permitted, it must be treated as truly random across all genes.

In particular, realistic selection coefficients need to be used especially if one is dealing with point mutations. At the borderline level for selection to be measurable they would be essentially zero. Ironically, the incremental improvement will decrease once acceptable functionality has been attain- ed, even as the proportion of improved configurations a- vailable become vanishingly small (Figure 2).

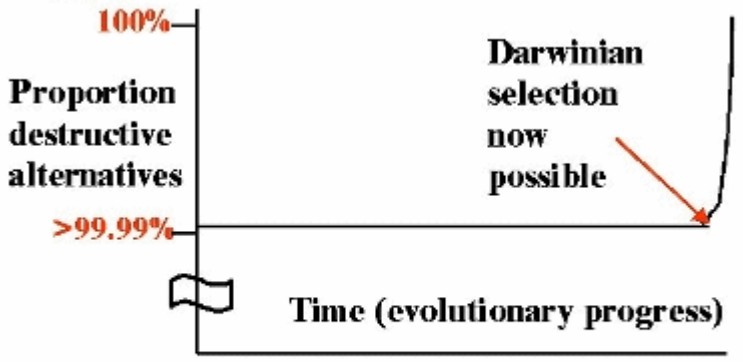

Fig 2. Destructive alternatives for binding sites increases with Shannon informational content.

It was a special creationist, Edward Blyth[61] who introduced in 1835, long before Darwin, the notion of natural selection, as a way of preventing major errors from being passed on to offspring. The selectivity coefficient, s, would be large when comparing a viable state to a major genetic disaster, but s for a base pair change would be near zero when attempting to distinguish between 'working quite well' and 'slightly better'. Eventually this resembles placing a ball on an almost vertical slope and hoping enough earthquakes would roll it up further uphill. The implausibility is strictly a statistical matter.

Realistic population genetics would need to be included in the post-selection simulation stage since even the rare good mutation has only a very small probability of being fixed in the population.

Admitedly even this proposed simulation would not model the generation of multi- gene, novel biological functions, via random mutations. Over-simplified models, such as Dawkins' example of mutating English letters, based on selfish gene notions, have no biological relevance(26) and the logical and mathematical flaws have been pointed out[62] [63].

Conceptually, Dawkin's example resembles fixing a magnet and allowing it to relentlessly attract a metal object, although very fast at first and slower towards the end. The distance can never increase between "generations".

Schneider's refinement allows the metal piece or magnet to move a very small distance sideways between generations, bringing the two relentlessly together. Progress at the beginning is also very rapid.

On average each generation must increase its Shannon information content, due to the way the algorithm was programmed, until reaching the intended plateau. Occasionally a given generation may be farther from the goal than the preceeding but the unrealistic parameter settings used guarantee success. Both programs have been intelligently designed to disallow failure to converge to the intended result given enough iterations.

We recognize repeatedly two remarkable assumption hidden in such simulations: chance mutations have a huge proportion of useful options at every step linking initially random base pair sequences and currenty observed genetic sequences; and these intermediate steps are selectively recognized with uncanny skill. This is quantitatively not consistent with what we know about mutations. ReMine[64] has criticized such unrealistic evolutionary assumptions in considerable detail(27).

Spetner has also questioned how many single nucleotide changes may actually be available with a measurable selective value[65] (recall that about 24% of these would code for the same amino if mutated randomly[1]). He points out[66] the dilemma this assumption causes: if faced with such a rich variety of useful mutations at all times, each heading off in different evolutionary directions, then long-term convergence to similar functional structures won't occur, contra what evolutionists claim. If each reasonably sized genome had a million felicitious mutations available then stumbling on similar organs (as observed for unrelated mammals and marsupials) via a multitude of random mutations is statistically absurd.(28),(28b)

Whereas step-wise development of binding sites by random mutations is not reasonable, one could entertain the notion that initially over-engineered binding sites had been Designed to provide robustness against random mutations. Point mutations could squeeze out excess Shannon information until the limit is reached where the locations can be unambiguously identified. Further destructive mutations would render the organisms non-viable and be selected against. Such a proposal would be inconsistent with an evolutionary viewpoint but consistent with Special Creation or Intelligent Design.

Suitability of Shannon's Definition of Information in Biology.

Few creationists or members of the Intelligent Design community view Shannon's work in telecommunications as an adequately comprehensive theory of information in biology, in spite of its mathematical virtues. It is certaily true that of all amino acid sequences which can occur, only a small subset fulfill a useful biological function, and the mathematics developed by Shannon, Tribus, Brillouin and others help with various probabilistic calculations. The word 'information' carries powerful and often inconsistent associations and I hope to provide a more useful and comprehensive theory of biological information later(29). Repetitive and reliable guidance of complex processes necessary for organisms to survive has no parallel in the non-living chemical world.

Examples include the production of highly specified proteins via the genetic code; coordination of multi-cellular processes (heat regulation, signal transmission, etc.); cell duplication; animal instincts; guidance of biochemicals to specific organelles across various membranes; production and delivery of energy packets (ATP). We have concentrated here on what may be the first attempt by evolutionists to model with a computer program the creation of a specific, new biological function: a novel binding site, by random point mutations and Darwinian selection.

There is currently intense discussion as to how 'information' should be defined and its properties[67] [68] [69]. Gitt[68] has examine many aspects of coded information, and concluded that information obeys many laws, one of which is that a coded information system can only arise by intelligent agency.

Conclusions

Dr. Schneider has identified a phenomenon which certainly needs explaining. After aligning 149 E.coli ribosome binding sites, *'We get: $R_{sequence}$ = 11.0 ± 0.4 bits per site, which is almost identical to the value of $R_{frequency}$, 10.8, we found earlier!* ***There is just enough pattern at ribosome binding sites (Rfrequency) for them to be found in the genetic material of the cell*** *($R_{sequence}$). These data imply that there is no excess pattern, and no shortage of pattern.'* (27)

The existence of patterns of minimal and highly conserved size, for which a protein has been precisely tailored, often aided by additional enzymes, displays a remarkable level of fine-tuning, and raises the issue who or what produced such a feat. Reliable identification of short patterns is only possible by precise 3-dimensional *Übereinstimmung* between recognizer and DNA site.

The simulation described was rigged to converge by using a large number of assumptions which are biologically unrealistic. Many cellular constraints were not included in the simulation, such as: the need for binding sites to be placed correctly with respect to pre-existing genetic elements which are to be regulated; the need for multiple new enzymes for recognizers to be able to work; the need to provide recognizers within an acceptable concentration range. Unrealistic parameter settings were used, including: the rate of mutation; the proportion of available useful mutations; the flawless effectiveness of natural selection.

Finally, the model is biologically fatally flawed in many ways: organisms with very small genomes which inherit superfluous DNA would be rapidly out-populated by those without it; the organisms are assumed to face only one survival goal; multiple and often inconsistent use of binding locations and recognizers was overlooked; recognizers are assumed to automatically be in the correct cellular compartment (organelle); and all details which could allow the simulation to fail, such as including randomizing mutations elsewhere in the genome, or error catastrophe, were excluded.

The limited goal of producing novel binding sites from scratch by random point mutations has not been demonstrated by this paper[1]. This can be easily demonstrated by sensitivity analysis (i.e., by testing various parameter settings) even using this flawed model as a starting point.

The reader is invited to test or consider the effects on generations needed by increasingly merely 3 parameter values, still far below what is biologically relevant:

Parameter	Value Used in Simulation	Realistic Value	For Sensitivity analysis
Selectivity coefficient[a]	ca. 1	0.001 to 0.01	ca. 0.05
Mutation[b] rate/ nucleotide / generation	1 / 256	$< 1 / 10^8$	ca. $1 / 10^5$
Proportion useful polypeptides[c], (147)	ca. $5X10^{-1}$	10^{-44}	ca. 10^{-15}

a. To avoid reprogramming, one could "kill" 1 or 2 low scorers instead of 50% every generation, and replace equal scorers by the descendant mutant. Since relative improvement becomes less noticeable as the binding interaction is optimized, and genetic drift is neglected, this value for validation purposes is obviously absurdly generous once the process begins.
b. HIV-1, with 10^{10} virions generated per day, was determined to be at the upper limit of possible mutation rate, with one error every 2000-5000 bases. 64 free living genomes could never tolerate this rate. For bacteria 1/10E8 to 1/10E10 is estimated[8]. So my proposed value is generous.
c. A proportion of 10^{-15} random polypeptide sequences with s=0.1 is absurdly generous, and if true would be verifiable empirically. The proportion of functional cytochrome c proteins (which are about 1/3 the size of average proteins) to all polypeptides of comparable length was calculated by Yockey to be on the order of 10E-44.

Such changes clarify how dramatically great the number of generations would become should realistic parameter settings be tested. **During these single goal iterations, function-destroying mutations would accumulate in other portions of the genome since the population is not allowed to perish. With a population of only 64 members natural selection cannot weed <u>all</u> flaws accumulating througout the genomes. I predict the net effect, if simulated realistically, will show a net destruction of functional specificity, meaning <u>a net decrease of information</u> as defined>[1] over all genes as time increases.**

It is apparent that the extrapolation to claim a billion years is sufficient to produce human beings by chance, starting from a random DNA sequence, given than even one novel binding site could not possibly be generated as proposed, is not justified.

Footnotes

(1) For example, gene duplication has been identified with: disorders of the PNS associated with duplication of the peripheral myelin protein-22 (PMP22) gene locus[70]; Charcot-Marie-Tooth disease (CMT), associated with a partial duplication of chromosome 17[71]; amplification of the MYCN gene is frequently observed in human neuroblastomas[40]; consider also the frequenin gene of Drosophila[72]. [RETURN TO TEXT]

(2) Such as the bacterial mRNA AUG initiation codon which is separated from the six nucleotide Shine- Dalgarno sequence: 'the procaryotic initiation codon, which is predominantly AUG, also has GUG and UUG on occasion.'[73] There may be 'spacing between parts of a binding site, as with ribosome binding sites (Shine and Dalgarno to initiation codon) or procaryotic promoters (-35 to -10).'[73] [RETURN TO TEXT]

(3) Dr. Scheider has alluded to such effects in an earlier paper: 'Secondly, the information patterns are different for the various repressors. LexA and TrpR have high peaks three bases wide, while ArgR has double spikes and cI/Cro have single spikes. These distinctive morphological differences probably reflect the location and strength of structural contacts between the different repressors and their cognate sites.'[25] [RETURN TO TEXT]

(4) 'Likewise, the number of sites is approximately fixed by the physiological functions that have to be controlled by the recognizer. So Rfrequency is essentially fixed during long perids of evolution. On the other hand, Rsequence can change rapidly and could have any value, as it depends on the details of how the recognizer contacts the nucleic acid binding sites and these numerous small contacts can mutate quickly. So how does Rsequence come to equal Rfrequency? It must be that Rsequence can start from zero and evolve up to Rfrequency. That is, the necessary information should be able to evolve from scratch'.[1] [RETURN TO TEXT]

(4b) For the C. elegans genome an average of 5 introns has been estimated for the approximately 19099 genes. Unregulated joining of the resultig exons provide a vast potential to generate worthless and interfering polypeptides.[51] p. 54. [RETURN TO TEXT]

(5) 'At any particular time in the history of a natural population, the size of a genome, G, and the number of required genetic control element binding sites, γ, are determined by previous history and current physiology, respectively, so as a parameter for this simulation we chose γ =16 and the program arbitrarily chose the site locations, which are fixed for the duration of the run.'[1]

'Probabilities computed from the individual information distributions are curious because sequences with evaluations significantly higher than the mean have low probabilities of being real sites, as can be seen in the distributions. Strong sites are less likely to appear in the set of natural sites. Evidently the sites evolve to what is required for their function rather than to become the strongest binder.'[74]

[RETURN TO TEXT]

(6) Dr. Schneider is aware that only the immediately functional will be tolerated: 'If a protein were unnecessary, mutations in its gene would eventually destroy it, and the ribosome binding site at the start of the gene would atrophy. Likewise, if the ribosome were to start translation in places that it shouldn't, the cell would waste energy making useless proteins. Thus it would make biological sense if the only places ribosome binding sites exist is in front of functional genes.'[27] [RETURN TO TEXT]

(7) From statements in ealier papers it is clear Dr. Schneider should have taken this into account. 'First, when two or more recognizers have binding sites that are always in the same register with respect to each other, the sequence conservation is higher than expected from the size of the genome and the number of binding sites. If a thorough information analysis has been done, the situation is easy to detect and in such cases it is unwise to use the individual information matrix because it does not represent a single entity. Second, when nearby sites are not in the same register, the sequence conservation of one site is blurred out in the alignment of the other site.'[74]

And elsewhere he writes, 'In comparison to the other binding sites, the pattern at T7 promoters in the phage genome is dense and contains more information than one would expect. However, when an experiment is performed to determine what components are important to the RNA polymerase, only half of the pattern remains. The excess pattern is thought to represent the binding of another DNA binding protein.' [73] [RETURN TO TEXT]

(8) The population is not allowed to go extinct to guarantee the intention of the compute program is met. We read, 'The fact that the population cannot become extinct could be dispensed with, for example by assigning a probability of death, but it would be inconvenient to lose an entire population after many generations.' Model accuracy to reflect the true state of affairs should be the correct criteria. [RETURN TO TEXT]

(9) Kimura[75] estimates for mammals with about 50 cell divisions along the germ line from the fertilized egg to a gamete between $50X10^{-8}$ to $50X10^{-9}$ per nucleotide pair per generation. Kondrashov[76] estimates for humans a higher value, between $2X10^{-8}$ and $1X10^{-7}$.

In bacteria the mutation rate per nucleotide has been estimated to be between 0.1 and 10 per billion transcriptions.[77] [78] 'For organisms other than bacteria, the mutation rate is between 0.01 and 1 per billion.'[79] [80] [RETURN TO TEXT]

(10) Sonneborn (1965) suggested[56] that the degeneracy redundance helped protect against protein error. This effect is very weak compared to other error correction mechanism available in the modern organisms. [RETURN TO TEXT]

(11) 'First, a matrix is created from the frequency f(b,l) of each base b at position l in the aligned sequences, according to Riw(b,l) = 2 + log2f(b,l) - e[n(l)], where e[n(l)] is a small sample correction. Second, this matrix is used to evaluate the individual information content of each site. That is, after aligning the matrix with a sequence, each base in the sequence selects one of the four weights in Riw(b,l), and all weights for the site are summed for all positions l to produce the "individual information", Ri. Third, if the individual information values of all the aligned sequences are averaged the result is Rsequence.'[81]

'Only functional sites are needed to create an individual information matrix...A set of 20 or more examples generally gives a reasonable sequence logo and weight matrix.'[81] [RETURN TO TEXT]

(12) Consider an illustrative example, such as the binding of transcription factor UBF to an

rDNA promoter[82]. This relieves repression by an inhibitory factor which competes for binding of TIF-IB to the rDNA promoter. To prevent runaway transcription and cell proliferation, the product of the retinoblastoma gene, pRb, can bind to UBF interfering with UBF's binding to DNA. pRb seems to act as a signal transducer connected to the cell cycle clock via its phosphorylation state. 'All three classes of cellular RNA polymerases are targets of pRb-mediated transcriptional repression.'[38] Defective RB genes can result in cancer in the retina or pancreas. Simulating the origin of the UBF factor or its binding site one random point mutation at a time, neglecting the regulatory environment, misses the essence of what is supposed to be modelled. [RETURN TO TEXT]

(13) Dr. Schneider is aware that binding sites depend on providing suitable geometric shapes: 'The messenger RNA which the ribosome is searching is shown as a string of a's, c's, g's and u's... The ribosome is depicted as an ellipse with two hook-like pieces. The left one represents the 3' end of the 16s rRNA, which is the part of the ribosome which recognizes the Shine and Dalgarno sequence ["ggag"], and the right one represents the first transfer RNA to which is attached a formylmethionine (fMet), the first amino acid of the new protein... the initiation codon [is] "aug".'[27] [RETURN TO TEXT]

(14) 'A "randomization" experiment was performed in which OxyR protein was used to gel shift 30 base pair equi-probable random sequences. Unfortunately this gave a dismal logo, possibly because the protein was prevented from binding properly by the flanking constant sequence of the vector.'[83] [i.e., rest of the protein is also relevant!].

'To clarify this situation, the randomization experiment was repeated with 45 base pair equi-probable random sequences which were then aligned by an information theory technique using the malign.p program. Only some of the patterns evident in Fig. 2a [based on biological OxyR binding sequences] were confirmed by this experiment, whereas others became more predominant.'[83] 'The almost insignificant weak preferences for A at ±6, T at ±9, and T at ±18 of wild-type appear amplified. Additional conservation not seen previously appeared at ±8 (?), ±12, ±16, ±19 and ±22. The reason for these quantitative discrepancies between the wild-type sequence logo and the logo from experimentally selected sites is unknown, but might be accounted for by the small sample sizes.'[83] [RETURN TO TEXT]

(15) 'Genetic control systems often work by one molecule binding to a spot to prevent another molecule from binding there.'[27]

Other examples have also been offered: 'The lac repressor protein will bind the operator only if it is not also binding an inducer.'[44]

In vitro experiments may fail to mimic in vivo observations if not all binding members are provided. 'For example, no spermidine was used in the gel shift experiment, yet it is well known that spermidine is important for precise recognition by other DNA binding proteins.'[83]

'Cui et al. investigated Lrp by using the SELEX (systematic evolution of ligands by exponential enrichment) procedure, an in vitro method that is used to identify binding motifs. In the SELEX procedure, a specific protein is used to select binding sequences from random

synthetic sequences.'[84] 'Surprisingly, the sequence logos for natural Lrp binding sites determined by footprints or mutations do not closely resemble the sequence logos obtained by SELEX.'[84] 'To explain these major discrepancies between the natural and the SELEX sites, we suggest that three proteins are binding in the SELEX experiment...Two of these are the complementary regions separated by 10 bp at -7 to -5 and +5 to +7...'[84] 'The strongest sites, such as those found by SELEX, are not "optimal" when viewed on an information theory scale.'[84] [RETURN TO TEXT]

(16) 'OxyR is a tetrameric protein that binds to the DNA of several promoters in Escherichia coli and activates transcription of genes encoding antioxidant enzymes.'[83] Furthermore, 'When a protein is in contact with a major groove, the two base pairs and their two orientations can be distinguished, as recognized by Seeman et all, so the protein is capable of "choosing" one of the four possibilities: A=T, T=A, C≡G, or G≡C.'[83] 'This choice of 1 possibility in 4 can be made with 2 bits of information. (This is calculated as $\log_2 4/1 = 2$)'.[83] 'In contrast to the major groove, contacts in the minor groove of B-form DNA allow both orientations of each kind of base pair so that rotations about the dyad axis cannot easily be distinguished.'[83] 'Because only 2 of the 4 possibilities can be distinguished, when a B-form minor groove is probed by a protein no more than 1 bit of information ($\log_2 4/2 = 1$ bit) can be obtained.'[83] [RETURN TO TEXT]

(17) Ultrabithorax and even-skipped homeo domain proteins (UBX and EVE) of Drosophila melanogaster exert active and opposite effects when bound to a common site upstream of a core promoter[85]; factors capable of binding to Sp1-binding sites including Sp3, Sp4, BTEB and BTEB2[24] (of a number of putative Sp1 target genes tested,

only two showed decreased expression in the absence of Sp1 which indicates other components are needed to regulate only the intended sites); occupancy of a operator by the tryptophan repressor blocks access to the promoter by RNA polymerase[86].

Dr. Schneider has alluded to this fact in earlier papers: 'It is possible for two such recognizers to have the same base preferences. Since we use sequences to estimate the probabilities of bases at each position, the analysis will give the same information content for two entirely distinct mechanisms.'[25] And, 'Most single-base changes in promoters and ribosome binding sites decrease synthesis by 2- to 20-fold (Mulligan et al., 1984; Stormo, 1986). Binding to similar sites would degrade the function of the entire system. For repressors, binding to pseudo-operators would increase the chances of gratuituosly inhibiting transcription and may also serve as a sink for the recognizer.'[16] [RETURN TO TEXT]

(18) Consider also Sp1 promoter selectivity (chromatin, TAFs, and CBP are required for synergistic activation by Sp1 and SREBP-1a.)[87].

As another example, 'Fis [Factor for Inversion Stimulation] is a pleitropic DNA-bending protein that enhances site-specific recombination, controls DNA replication, and regulates transcription of a number of genes in Escherichia coli and Salmonella typhimurium.'[88] [RETURN TO TEXT]

(19) SAL[89] (development of the fly's gut); Dorsal[90]; Cubitus interruptus (Ci) (drosophila limb development by regulating different sets of Hh target genes)[91] [92]; RelB (acts like a transactivator with p50, whereas RelA-mediated transactivation is reduced by RelB)[93]; GLI3 (Mutations are known to alter the balance between its activator and repressor function)[94]; VDR (novel coactivator complex of VDR called DRIP was reported, which seems to confer cell-type specific effects)[95]; Cro and cI (are DNA-sequence specific activators and repressors)[96] [97]; Sp3 and BTEB (24); GA (both an activator and repressor of ribosomal protein gene transcription)[98]; Sim[99]; Krüppel (activates transcription as a monomer through an interaction with TFIIB, represses by interacting with TFIIE)[100] [101] [102] [103] [104]; SpoIIID (sporulation in B. subtilis)[105]; Tax (an HTLV-I oncoprotein)[106]; p53 (a tumor suppressor)[107] [108] [109]; Fos (activator and repressor of transcription due to differential Fos phosphorylation)[110]; Retinoic-acid receptors and Thyroid- hormone[111]; Retinoic acid receptors (RARs) and retinoid-X receptors (RXRs) (activate or repress transcription by binding as heterodimers to DNA-response elements that generally consist of two direct repeat half-sites of consensus sequence AGGTCA. Ligand-dependent transactivation by RAR on DR + 5 elements requires the dissociation of a new nuclear receptor co-repressor, N-CoR, and recruitment of the putative co-activators p140 and p160)[112]

Dr. Schneider writes: 'Lrp [the leucine-responsive regulatory protein] binds to multiple sites in a number of operons, including dad, fanABC, papBA and ilvIH. Leucine can invoke either positive or negative transcriptional control by Lrp.'[84] 'Lrp is known to both activate and repress transcription, so sequence logos for both Lrp activation and repression sites were made. There are no major differences observed between the sequence characteristics or activation and repression sites, except for more strongly conserved bases at the -10, -9, -2, -1 and +6 positions in the activation logo and a more strongly conserved A at the +2 position and T at the -4 and +12 positions in the repression logo.'[84]

27 E.coli Lrp binding sites were identified, which act on a large number of different genes or operons[84]. Of these, Lrp had an activation effect in 17 cases and a repression effect in 9 cases (for trxB the effect was not known).

'We found that we could not predict repression versus activation. The failure of this bootstrap test, for all sites, suggests that either the activation and repression sites are essentially identical or that more examples are needed to distinguish between them.'[84] [RETURN TO TEXT]

(20) SREBPs, Sterol Regulatory Element-Binding Proteins (regulated by SCAP cleavage-activating protein which forms complexes with SREBPs in membranes of the endoplasmic reticulum (ER). In sterol-depleted cells, SCAP facilitates cleavage of SREBPs by Site-1 protease, thereby initiating release of active NH(2)- terminal fragments from the ER membrane so that they can enter the nucleus and activate gene expression)[113]; Sphingomyelinase (balance between cholesterol and sphingomyelin regulated by proteolytic cleavage of SREBPs. Sphingomyelinase causes a fraction of cellular cholesterol to translocate from the plasma membrane to the endoplasmic reticulum. Sphingomyelinase prevents the nuclear entry of sterol regulatory element binding protein-2 (SREBP-2))[114]. [RETURN TO TEXT]

(21) There are other interesting examples. Multiple forms of the chicken estrogen receptor-alpha protein (ER-alpha) are transcribed from a specific promoter that is located in the region of the previously assigned translation start site. The resulting cER-alpha forms I and II differ in their ability to modulate estrogen target gene expression in a promoter- and cell type-specific manner.[115]

Four different cER alpha mRNA isoforms are under the control of four different promoters in chicken tissues to modulate the levels of expression of the chicken ER alpha gene in a tissue-specific and/or developmental stage-specific manner.[116]

Splicing exons to produce the proper isoforms can be controlled according to type of tissue, such as reported for the three variants of GPDH.[117] [RETURN TO TEXT]

(22) The transcription factor OxyR, can sense elevated levels of hydrogen peroxide and induce the expression of a transcript encoding another transcription factor Fur, raising its concentration from about 5,000 molecules/cell in E. coli to about 10,000 after oxidative stress[118]. This contrasts with concentrations of other transcription factors in E. coli, such as about 10 to 20 copies of the LacI repressor and 50 to 300 copies of the Trp repressor[86] per cell.[118] Note that there are on the order of one billion proteins molecules in total distributed throughout cells.

'The previously identified 9.9 bit Fis site at -66 overlaps Xis site 2... Fis is involved in both integrative and excisive recombination and will stimulate excision when the concentration of Xis is low. Fis binding to the 9.9 bit site excludes Xis binding at Xis 2 and stimulates Xis binding at Xis 1.'[81] [RETURN TO TEXT]

(22b) 'Therefore, at high concentrations H-NS can bind to DNA fragments in a nonspecific fashion, but at low concentrations it shows a clear binding preference for the proU regulatory region.'[39], p.6580. 'At slightly higher H-NS concentrations, we found that also the 192 bp EcoRI-BglI fragment from pΔ651 (comprising proU sequences from +24 to +202 bp) was efficiently retarded'[39], p.6580 'When the H-NS concentration was further increased, both the 933-bp fragment derived from the vector and the 992-bp fragment from pBK20 carrying the proU promoter were bound at the same protein concentration.'[39], p.6581 'We note that H-NS occupation of the extended binding region at the 5' end of proV begins at relatively low protein concentration (0.22 μm), whereas the protection of the region around the proU - 35 sequence from Dnase I digestion requires a substantially higher H-NS concentration (6.2 μm)'[39], p.6582 'When interpreting the H-NS footprinting data for the various promoters, one needs to consider that binding of H-NS to a particular target sequence is concentration dependent. This is illustrated by the more than 20-fold higher concentration of H-NS required to protect the G + C-rich region around the proU -35 sequence

(-22 to -29 bp) in comparison to the highly A + T-rich extended H-NS binding region at the beginning of the proV gene (+64 to +109 bp)'[39], p.6584 [RETURN TO TEXT]

(22c) Proteins do not always bind exclusively at DNA sequences of the same length. 'We carried out an additional footprinting experiment with a DNA fragment (fragment III) that

allowed us to monitor H-NS binding to sequences upstream of the proU promoter. Again, *several protect regions with variable size and spacing were visible.*'[39], p.6582 [emphasis added] 'However, a striking feature of all the DNA segments protected by H-NS is their high A + T content and the presence of uninterrupted stretches of 5 or more A• T base pairs.'[39], p.6582 [RETURN TO TEXT]

(23) 'Another difficulty with this kind of experiment is that it always contains at least one unknown parameter, the stringency of selection. If the concentration of OxyR protein were large, then its non- specific binding should cause more DNA sequences to shift in the gel. This would lead to a sequence logo with low information content relative to the natural sequences. However, a low concentration of OxyR protein should lead to a much higher measured information content, perhaps higher than is naturally found.'[83] [RETURN TO TEXT]

(24) Examples: low-level increase in proteolipid protein (Plp) gene expression (causes CNS disease)[70] [119]; peripheral myelin protein-22 (PMP22)[120]; beta-globin-gene (underexpression is also worthless)[46]; flies overexpressing hop cause ectopic wing veins and duplications, eye defects and melanotic tumors[121]; frequenin in Drosophila[72]; PAI-1 (mice which lack or overexpress revealed a correlation between the level of PAI-1 and the extent of lung fibrosis after injury)[122]. That a minimal amount of mRNA is needed to be biologically detectable is apparent, but even small decreases from the required may be no better than none. Tumorous cellular proliferation is prevented via interference of the UBF transcription factor by pRb. Too much UBF factor (or too little pRb) does not prevent transcription activity[38].

'Because Fis bends DNA when it binds, the multiple DNA contortions might exclude RNA polymerase and silence transcriptional initiation. As levels of Fis protein decrease in the cell, the physical blockage would be relieved and transcription could proceed again.'[88] [RETURN TO TEXT]

(25) Yockey has shown[123] [124] that the number of synonymous codons allows a reasonable estimate of amino acid frequencies (limiting coding of arginine to only codons AGA and AGG). Experimental data based on known residues reported by other authors was reported[123] to establish the relative distribution of amino acids. The average Shannon information was determined to be about 4.153 bits/residue for proteins as a whole. This figure does not take into account structurally similar amino acids at various positions which would allow a protein to perform its intended function, and thus is an upper limit.

The protein sequences of all known cytochrome c were collected to identify possible alternative synonyms at each position. A sequence of length 110 amino acids was used after excluding those positions for which some organisms lacked an amino acid there[147]. [RETURN TO TEXT]

(26)

- Evolution cannot know in advance the length the new protein is supposed to be, so the stop codon needs to be suitably located to generate a protein which is neither too short nor long.
- Non-translated binding sequences, whose position can vary greatly according to gene, need to be available to allow transcription to be regulated. For example, translation of the GluR-B gene appears to initiate approximately 430 nucleotides upstream of the translational start codon, with no intron in the 5'-untranslated region of the gene[125].
- Other proteins involved in regulating transcription must already be available and located in the nucleus. Transcriptional initiation in eukaryotes as a rule requires the ordered assembly of a large number of protein factors into a functional preinitiation complex[126] [127]. In addition, untranslated regions of mRNA upstream (5'UTR) and downstream (3'UTR) of the open reading frame, and the mRNA precursor can carry important regulatory sequences[128]
- The estimated polypeptide possibilities[55] [56] [147], is based on Shannon's formula,

  $$2^{nH} = 2^{(n \times 4.153)}$$

  where n refers to the number of residues; this excludes a very large number of possibilities based on sequences which use many residues of lower probability.
- Removing introns correctly is not taken into account. The GluR-B gene spans more than 90 kilobase pairs and harbors 17 exons. Four alternatively spliced mRNAs are generated from the primary GluR-B transcript[125].
- Whether all protein variability found across all organisms would be acceptable in finding the first one, by trial and error, with measurable functionality, is unlikely[82]. For example, ribosomal gene transcription is markedly species specific since the RNA polymerase I transcription apparatus relies on different promoter-recognition properties.

The extreme intolerance of some enzymes to alternative residues has been documented[57] [58]. From the abstract of the latter reference: *'The amino acid sequences of enzymes like alcohol dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase are strongly conserved across all phyla. We suggest that the amino acid conservation of such enzymes might be a result of the fact that they function as part of a multi-enzyme complex. The specific interactions between the proteins involved would hinder evolutionary change of their surfaces.'* It makes no sense to try developing each functional gene by random point mutations one after the other. Behe might consider these as also examples of "irreducible complexity". **[RETURN TO TEXT]**

(27) Each time a letter which lines up correctly to the pre-arranged target, 'Methinks it is like a weasel' showed up, that 'parent' became flawlessly the starting point for n offspring, each of which are allowed to improve in the pre-targeted direction. If no additional matches showed up, the parent sequence was retained, and a fresh crop of offpring was given another

allowed. Of all offspring, the one which made the most progress towards the sentence already stored in the computer was always retained as the new ancestor.[129] [RETURN TO TEXT]

(28) 'First, selection in nature is not perfect... When selection is weaker, the substitution requires more time. Second, beneficial mutations are not easily produced. They are rare. A population of 100,000 is not likey to receive a major one every generation. Third, the effect of harmful mutations has not been counted. These must be eliminated by differential survival, and this raises the cost of the process. Fourth, time was not deducted for periods when the population is stuck on a local fitness peak, undergoing little if any change.'[64] [RETURN TO TEXT]

(29) 'There are thousands of examples of convergence that I could give...One amazing example of convergence is the ultrasonic echolocation systems (like sonar) found in animals scattered through the vertebrate phylum... Bats have an echolocation system, and so do toothed whales and dolphins. The system is also found in some birds. According to the experts, these systems could not all have been derived from a common ancestor...South-American electric fish and African electric fish both "see" in dark murky water by measuring the distortion of the electostatic fields they generate in the water around them. These two groups of fish are believed to have developed their electrostatic-imaging systems independently.'[130] [RETURN TO TEXT]

(29b) Striking examples of molecular convergence have been pointed out[131]: in fish antifreeze proteins[132], cytokinases[133], and apolipoproteins[134]. At a higher functional level there are many examples: unrelated fish (Eigenmannia and Gymnarchus) which produce electric signals to confuse predators[135] [136]; remarkably similar structure of receptors between arthropods and vertebrates based on different binding proteins[137]. [RETURN TO TEXT]

(30)

- The origin of the code and decoding apparatus is neglected. The later has many properties worthy of technical analysis and incompatible with chance origin
- The mathematics handles properties of coded messages and not the outcome. Coded information systems can be set up to require more or less input from a coded message to achieve the same goal. For example, a collection of 20 amino acids which is to produce a functional protein must indeed be placed in an acceptable sequence. But intramolecular reactions between the amino and carboxyl ends of the same chain must be precluded during the polymerization process. One could either design the decoding apparatus to prevent this possibility (as ribosome's geometry does) or more complex (= greater Shannon information content) messages are needed to unambiguously specify the desired outcome from among a now greater range of undesired possibilities. Also, since the peptide bonds used by proteins are generated only about half the time under unconstrained chemical reactions of amino acids, the decoding apparatus must either prevent this also (as ribosome does) or a yet more demanding coding scheme (more bits to specify intended outcome) would be needed to provide the necessary guiding intructions.

- All coded information systems work in a specific context, where much can be assumed without the need for clarification in the form of longer messages. The genetic code "knows" the t-RNA will already be optically pure[138]; that water and other interfering chemicals will be excluded (the message does not need to guide processes to ensure this).
- By 'information' a biologist has an intuition along the lines of "that which guides behavior to produce chemically and physically unexpected results", such as instincts (e.g. eel migration; nest building; mating); cooperative efforts (e.g., the foraging bee's waggle dance; pheromones); and the production of novel structure (e.g., eyes, blood circulation).
- According to Dr. Scheider, total information at a position x of a gene is its 'surprisal' upon determining which base is found there. The surprisal notion is given by:

  $$R_{sequence}(x) = H(x)_{max} - H(x)$$

  where $H(x)_{max}$ bits refer to the 4 possibilites (a,c,g or t) being of identical probability.
- This can be modified in several ways which have no relevance to what one is thinking about when using the word 'information':

  a. Errors which allow new bases to appear or existing ones to be modified on the genome would increase $H(x)_{max}$. Even if a specific gene is not be affected, total information supposedly increases.
  b. Should a catastrophy spare a portion of a colony living in a small region, like the immediate relatives of a recent ancestor, the variability of bases at a given position will be less, implying these and their ancestors would automatically have a greater information content, even though the variety of that gene/allele may be of biological lower "quality" than for the original population.
- Increasing the length of a gene with junk would always lead to more total Shannon information
- A small gene producing a protein with many biological functions would have less Shannon information than a larger one with a single function

[RETURN TO TEXT]

(31) 'Second, most of the signaling and regulatory genes known or expected to be involved in multicellularity have no yeast orthologs, even though they may contain domain sequences shared with yeast. Thus, virtually all biological processes characteristic of multicellular life are performed by proteins that are not close variants of proteins responsible for the core processes, even though they might share some domains.' [13] p. 2027.

Biologically usable polypeptides are only a small subset of possible true proteins, given the need for proper folding and other constraints to be useful. Repeated use of common themes across unrelated proteins, especially those for which no plausible common ancestry can be proposed, is consistent with the view intelligent agency was involved. This view is particularly persuasive when one realizes many protein domains, which display very little variability in all their positions, can consist of chains 200 - 300 amino acids long. [RETURN TO TEXT]

Table 1. Lengths (L) and number (γ) of some reported binding sites.

Factor	**Binding Length, L**	**Number in genome, γ**	**Ref.**
c-Jun	4	Not reported	(59)
*Hae*III	4	Not reported	(139)
EcoRI	6	Not reported	(27)
GATA-1	6	Not reported	(3)
Sp1	6	Not reported	(3), (59)
CTF/NF-1	6	Not reported	(59)
RNA polymerase II (TATA box)	6	Not reported	(34)
AP-2	8	Not reported	(59)
CREB	8	Not reported	(59)
OCT-1; OCT-2	8	Not reported	(3), (59)
GCN4	9	Not reported	(3)
C/EBP	9	Not reported	(59)
bicoid	9	Not reported	(3)
MAT $\propto$2	9	Not reported	(3)
Krüppel	10	Not reported	(3)
human donor splice junctions	10	Not reported	(74)
Symmetry	14	Not reported	(139)
GR	15	Not reported	(59)
CAP	16	Not reported	(3)
Lambda repressor	17	Not reported	(3)
GAL4	17	Not reported	(3)
CI/Cro	19	12	(139)
LexA	20	22	(139)
SRF		Not reported	(59)
ArgR	20	22	(139)
Lac repressor	21	Not reported	(3)
Fis	21	Not reported	(88)
human acceptor splice junctions	28	Not reported	(74)
TrpR	38	6	(139)
RNA Pol	42	83	(139)
LacI	43	2	(139)

Ribosome	45 40	2574 Not reported	(139) (74)
OxyR	45	Not reported	(83)
H-NS	46	Not reported	(39)
*Hinc*II	51	Not reported	(139)

Table 2. Scoring Matrix implied in [1] to evaluate binding interactions.

Pos.	A	C	G	T	Comments
#1	tcttt	gcacg	ctaag	Tttgt	The starting random weighting matrix
#2	cagga	attgt	aaaca	Cctaa	
#3	tccgt	ccatg	atttg	Tctga	
#4	cctac	attgt	tggac	Gagaa	
#5	gctca	tcggg	tatgc	cagcg	
#6	gggct	ggacg	gtcaa	tggca	
#1	-0 10 00 00 01	-1 10 11 10 10	+1 11 00 00 10	-0 00 00 01 01	Two's complement based on: A = 00 C = 01 G = 10 D = 11
#2	+1 00 10 10 00	+0 11 11 10 11	+0 00 00 01 00	+1 01 11 00 00	
#3	-0 10 10 01 01	+1 01 00 11 10	+0 11 11 11 10	-0 10 00 10 00	
#4	+1 01 11 00 01	+0 11 11 10 11	-0 01 01 11 11	-1 11 10 00 00	
#5	-1 10 00 11 00	-0 10 01 01 10	-0 11 00 01 11	+1 00 10 01 10	
#6	-1 01 01 10 01	-1 01 11 10 10	-1 00 11 00 00	-0 01 01 11 00	
#1	-129	-442	450	-5	Example: **catctt** scores as: (-422+296-136+251+294-92) (=171,>-58, the cutoff score)
#2	296	251	4	368	
#3	-165	334	254	-136	
#4	369	251	-95	-480	
#5	-396	-150	-199	294	
#6	-346	-378	-304	-92	

Table 3. Synonyms after a single point mutation.

Amino Acid	Coding Codons							Number of synonyms
Ala	GCA GCC GCG GCU	GCA: 3	GCC: 3	GCG: 3	GCU: 3			12
Arg	AGA AGG CGA CGC CGG CGU	AGA: 2	AGG: 2	CGA: 4	CGC: 3	CGG: 4	CGU: 3	18

Asx	AAC AAU	AAC: 1	AAU: 1					2
Asp	GAC GAU	GAC: 1	GAU: 1					2
Cys	UGC UGU	UGC: 1	UGU: 1					2
Gln	CAA CAG	CAA: 1	CAG: 1					2
Glu	GAA GAG	GAA: 1	GAG: 1					2
Gly	GGA GGC GGG GGU	GGA: 3	GGC: 3	GGG: 3	GGU: 3			12
His	CAC CAU	CAC: 1	CAU: 1					2
Ile	AUA AUC AUU	AUA: 2	AUC: 2	AUU: 2				6
Leu	CUA CUC CUG CUU UUA UUG	CUA: 4	CUC: 3	CUG: 4	CUU: 3	UUA: 2	UUG: 2	18
Lys	AAA AAG	AAA: 1	AAG: 1					2
Met	AUG	AUG: 0						0
Phe	UUC UUU	UUC: 1	UUU: 1					2
Pro	CCA CCC CCG CCU	CCA: 3	CCC: 3	CCG: 3	CCU: 3			12
Ser	AGC AGU UCA UCC UCG UCU	AGC: 1	AGU: 1	UCA: 3	UCC: 3	UCG: 3	UCU: 3	14
Thr	ACA ACC ACG ACU	ACA: 3	ACC: 3	ACG: 3	ACU: 3			12
Trp	UGG	UGG: 0						0
Tyr	UAC UAU	UAC: 1	UAU: 1					2
Val	GUA GUC GUG GUU	GUA: 3	GUC: 3	GUG: 3	GUU: 3			12
Stop	UAA UAG UGA	UAA: 2	UAG: 1	UGA: 1				4
						Synonyms:		138
						138 / (64*9) =		0,24

Table 4. The Population of mRNA Molecules in a Typical Mammalian Cell.[47]

	Copies per Cell of Each mRNA Sequence	Number of Different mRNA Sequences in Each Class	Total Number of mRNA Molecules in Each Class
Abundant class	12,000	4	48,00
Intermediate class	300	500	150,000
Scarce class	15	11,000	165,000

Most mRNA species are present at a low level (5 to 15 molecules per cell) (47)

Table 5. Number of proteins for which domains are not shared by C. elegans and S. cerevisiae (from [84]) *

Domain	Description	Yeast	Worm
Domains found only in the worm			
PTB	Phosphotyrosine binding domain	0	11
NHR	Transcription factors with ligand and DNA binding Zn-finger domains	0	270
EGF	Calcium-binding, seen in epidermal growth factor, etc.	0	135
Degenerins	Amiloride-sensitive Na^+ channels	0	28
T-box	DNA-binding domain of transcription factors	0	21
FMRFamides	Neuropeptides	0	20
Cadherin	Calcium-dependent cell adhesion module	0	18

Paired box	DNA-binding domain with 2 helix-turn-helix (HTH) units	0	18
SMAD	Transcription factors	0	8
Insulin-like peptides	Peptide hormones	0	7
Laminin NT	N-terminal globular domain of the extracellular matrix protein laminin	0	5
Domains found only in yeast			
APSES	A fungal-specific DNA-binding domain, seen in Swi4p	6	0
C6	A fungal-specific binuclear Zn-binding cluster	54	0

* Number of proteins containing the given domain in on organism but the other.

Appendix

Binding sites are defined on Schneider's web site[1] as: 'The place a protein (or macromolecular complex) binds on a nucleic acid. A classic example is the set of binding sites for the bacteriophage Lambda Repressor (cl) protein.'

'For example, a set of ribosome binding sequences can be aligned at the translational initiation point.'[73] Repressors, polymerases, ribosomes and other macromolecules can identify and bind to specific nucleic acid sequences.[139]

A weight matrix is built by aligning at the binding site a collection of nucleotide sequences, using only examples with proven biological effects (this can also be done with amino acids on proteins). The frequency, or proportion, of each base, A,C,G or T, at each nucleotide is determined from the experimental data.

Example of weight matrix calculations as done with real experimental data[140]: A) the jth sequence matrix, s(b,l,j)

base b							
	C -3	**A -2**	**G -1**	**G 0**	**T 1**	**C 2**	**T 3**
A	**0**	**1**	**0**	**0**	**0**	**0**	**0**
C	**1**	**0**	**0**	**0**	**0**	**1**	**0**
G	**0**	**0**	**1**	**1**	**0**	**0**	**0**
T	**0**	**0**	**0**	**0**	**1**	**0**	**1**

B) individual information weight matrix, Riw(b,l)

Base b							
	-3	-2	-1	0	1	2	3
A	+0.42	+1.25	-1.41	-∞	-5.81	+1.12	+1.51
C	+0.58	-0.78	-2.40	-7.81	-5.49	-3.68	-1.56
G	-0.58	-1.04	+1.64	+1.99	-6.23	+0.72	-1.06
T	-1.02	-0.87	-1.67	-5.81	+1.98	-3.38	-1.59
	C	A	G	G	T	C	T

The sequence 5' CAGGTCTGCA 3' represented in matrix format. There is only one “1” in each column, marking the base at that position. The remainder of the column is filled with “0”s. B) The individual weight matrix for human donor splice junctions derived from data given in[141]. The weights of the matrix in B that are selected by the sequence in A are enclosed in boxes. Fig. 1. Matrix representation of a sequence and a sequence recognizer.[140]

In (17), average amount of information, R, is defined as the uncertainty of the receiver before receiving symbols minus the uncertainty after reception:

$R = H_{before} - H_{after}$ (bits per symbol)

‘For protein binding on a nucleic acid, the before state is the recognizer unbound or nonspecifically bound and the after state is it being specifically bound.’[88]

Shannon defines information as equivalent to the entropy, H, of the probability space containing the events, i:

$$-\sum_i p(i)\log_2 p(i)$$

Yockey has recently provided an excellent discussion[124]

References

[1] Schneider, T. D. (2000) *Nucleic Acids Res*., **28**, 2794-2799. [RETURN TO TEXT]

[2] http://www.lecb.ncifcrf.gov/~toms/delila/ev.html [RETURN TO TEXT]

[3] Ref. 142 p. 407 [RETURN TO TEXT]

[4] Rubin, G. et al. (2000), *Science*, **287**, 2204-2215. [RETURN TO TEXT]

[5] Loeb, L. A., Essigmann, J. M., Kazazi, F., Zhang, J., Rose, K. D. and Mullins, J. I. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 1492-1497. [RETURN TO TEXT]

[6] Ref. 5 p. 1496 [RETURN TO TEXT]

[7] Eigen, M. (1993) *Gene* **135**, 37-47. [RETURN TO TEXT]

[8] Holland, J. J., Domino, E., de la Torre, J. C. and Steinhauer, D. A. (1990) *J. Virol*. **64**, 3960-3962. [RETURN TO TEXT]

[9] Ref. 5 p. 1492, referencing: Coffin, J. M. (1995) *Science* **267**, 482-489. [RETURN TO TEXT]

[10] Ref. 5 p. 1493. [RETURN TO TEXT]

[11] Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996) *Science* 271, 1582-1588. [RETURN TO TEXT]

[12] Munich Information Center for Protein Sequences (MIPS). On-line: http://www.mips.biochem.mpg.de/proj/yeast/tables/inventy.html [RETURN TO TEXT]

[13] Chervitz, S. A., Aravind, L., Sherlock, G., Ball, C. A., Koonin, E. V., Dwight, S. S., Harris, M. A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J. M. and Botstein, D. (1998) *Science*, 282, 2022-2028. [RETURN TO TEXT]

[14] http://bioslave.uio.no/~andrewsl/seqanal_model/multicellular.html [RETURN TO TEXT]

[15] Vreeland, R. H., Rosenzweig, W. D., and Powers, D. W. (2000) *Nature* 407, 897-900. [RETURN TO TEXT]

[16] Ref. 139 p. 425 [RETURN TO TEXT]

[17] Schneider, T. D. (1991) *J. Theor. Biol.*, 148, 83-123. [RETURN TO TEXT]

[18] Ref. 142 chapter 9 [RETURN TO TEXT]

[19] Bowie, J. U., and Sauer, R. T. (1989) *Proceedings of the National Academy of Sciences USA*, 86, 2152-2156 [RETURN TO TEXT]

[20] Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990) *Science*, 247, 1306-1310. [RETURN TO TEXT]

[21] Reidhaar-Olson, J. F., & Sauer, R. T. (1990) *Proteins: Structure, Function, and Genetics*, 7, 306-316. (the 10^{-65} figure assumes stable folded shapes will also retain their functionality). [RETURN TO TEXT]

[22] Behe, M.J., http://www.leaderu.com/orgs/fte/darwinism/chapter6.htm. [RETURN TO TEXT]

[23] Stephen, M. in Ref. 69 p. 25: Sauer's results[19] [20] [21] with the lambda and arc repressors indicate less than 1/4 the amino acids on average can be substituted and retain protein functionality. Note that $(1/4)^{150} < 10^{90}$. [RETURN TO TEXT]

[24] Nielsen, S. J., Präestegaard, M., Jørgensen, H. F., and Clar, B. F. C. (1998) *Biochem. J.* 333, 511–517. On-line: http://www.biochemj.org/bj/333/0511/bj3330511.htm#TOP [RETURN TO TEXT]

[25] Ref. 139 p. 424 [RETURN TO TEXT]

[26] Frank, J . and Agrawal, R. K. (2000) *Nature*, 406, 318-322. [RETURN TO TEXT]

[27] Schneider, T. D. (1994) *Nanotechnology*, 5, 1-18. [RETURN TO TEXT]

[28] Ref. 142 p. 244 [RETURN TO TEXT]

[29] Ref. 142 p. 386 [RETURN TO TEXT]

[30] Patterson, C. (1978) Evolution, British Museum (Natural History) Cornell University Press. (Referenced in (64) p. 214) [RETURN TO TEXT]

[31] Simpson, G. G (1953) The Major Features of Evolution, New York: Columbia University Press. Cited in (66) p. 102. [RETURN TO TEXT]

[32] Fisher, R. A. (1958) The Genetical Theory of Natural Selection, Oxford. Second revised edition, New York: Dover. Cited in (66) p. 102. [RETURN TO TEXT]

[33] Weindel, K.,'Konstitution von Nucleinsäuren: Hinweise auf funktionelle Optimierung' (2000) *Stud. Int. J.* 7, 36-38. [RETURN TO TEXT]

[34] Lodish, H., Berk, A., Zipursky, A.L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000) Molecular Cell Biology, W. H. Freeman and Company, 4th edition, New York, New York (p. 365). [RETURN TO TEXT]

[35] Chauvin, Rémy (1997) Le darwinisme ou la fin d'un mythe Éditions du Rocher, Monaco. (p. 198) [RETURN TO TEXT]

[36] Ref. 35 p. 203 [RETURN TO TEXT]

[37] Ban, N., Nissen, P., Hansen, J., Moore, P. B., Steiz, T. A. (2000) *Science* 289, 905-930. [RETURN TO TEXT]

[38] Voit, R., Schäfer, K., Grummt, I., (1997) *Mol Cell Biol* (1997) 17(8), 4230-4237. Download: http://mcb.asm.org/cgi/reprint/17/8/4230?view=reprint&pmid=9234680 See also: http://www.dkfz-heidelberg.de/polymeraseI/p53.htm [RETURN TO TEXT]

[39] Lucht, J. M., Dersch, P., Kempf, B., and Bremer, E., (1994) *J Biol Chem*, 269, 6578-6578. [RETURN TO TEXT]

[40] Wenzel, A., Schwab, M. , *Eur J Cancer* (1995) 31A(4), 516-519. On-line abstract: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7576956&dopt= Abstract [RETURN TO TEXT]

[41] Gawantka, V., Delius, H., Hirschfeld, K., Blumenstock, C., and Niehrs, C. (1995) *Embo J*, 14, 6268-6279. [RETURN TO TEXT]

[42] Hateboer, G., Gennissen, A., Ramos, Y. F., Kerkhoven, R. M., Sonntag-Buck, V., Stunnenberg, H. G., and Bernards, R. (1995) *Embo J*, 14, 3159-3169. [RETURN TO TEXT]

[43] Heix, J., Vente, A., Voit, R., Budde, A., Michaelidis, T. M., and Grummt, I. (1998) *Embo J*, 17, 7373-7381. [RETURN TO TEXT]

[44] Schneider, T. D. (1991) *J. Theor. Biol.*, 148, 125-137. [RETURN TO TEXT]

[45] Erb, K. J., Ruger, B., von Brevern, M., Ryffel, B., Schimpl, A., and Rivett, K. (1997) *J Exp Med*, 185, 329-339. [RETURN TO TEXT]

[46] Eck, S. L., The Prospects for Gene Therapy, University of Pennsylvania. On-line: http://www.hosppract.com/genetics/9910/mmceck.htm [RETURN TO TEXT]

[47] Ref. 142 p. 437 [RETURN TO TEXT]

[48] Barth, M., Marschall, C., Muffler, A., Fischer, D., and Hengge-Aronis, R. S (1995) *J Bacteriol*, 177, 345534-345564. [RETURN TO TEXT]

[49] Flouriot, G., Griffin, C., Kenealy, M., Sonntag-Buck, V., and Gannon, F., (1998) *Mol Endocrinol*, 12, 1939-1954. [RETURN TO TEXT]

[50] Alexandraki, D., Katsoulou, C., Thireos, G. and Tzermia, M.: http://www.imbb.forth.gr/groups/yeast/projects/YGS.html [RETURN TO TEXT]

[51] Wilson, R. K. (1999) *Trends in Genetics*, 15(2), 51-58. [RETURN TO TEXT]

[52] A C. elegans Database (ACeDB) [RETURN TO TEXT]

[53] Yockey, H. P. (1977) J. Theor. Biol., 67, 345-376; Yockey, H. P. (1981) *J. Theor. Biol.*, 91, 13-31. [RETURN TO TEXT]

[54] Weiss, O., Jiménez-Montaño, M. A. and Herzel, H. (2000) *J. Theor. Biol.* 206, 379-386. See references provided therein. [RETURN TO TEXT]

[55] Yockey, H. P. (1981) *J. Theor. Biol.*, 80, 21-26. [RETURN TO TEXT]

[56] Yockey, H. P. (1977) *J. Theor. Biol.*, 67, 377-398. [RETURN TO TEXT]

[57] Ridley, Mark (1996) Evolution, Blackwell Science Inc, 2nd edition, Boston, USA p. 189. (Alcohol dehydrogenase, critical for cell metabolism, contains 255 residues with no neutral amino acids, and one allele based on a single amino acid difference). [RETURN TO TEXT]

[58] Kisters-Woike, B., Vangierdegom, C., Muller-Hill, B. (2000) *Trends in Biochemical Sciences* 25(9) 419-421. [RETURN TO TEXT]

[59] Mitchell, Pamela J. and Tjian, Roberts (1989) *Science* 245:371-378. [RETURN TO TEXT]

[60] Lindah, T. (1993) *Nature*, 362, 709-715. [RETURN TO TEXT]

[61] Eiseley, L. (1979). Darwin and the Mysterious Mr. X (New York: E.P. Dutton, p. 55 Pointed out in: Humber, P. G., Impact No. 283, http://www.icr.org/pubs/imp/imp-283.htm [RETURN TO TEXT]

[62] Truman, R. (1998) *Creation Ex Nihilo Technical Journal*, 12(3), 358–361. On-line at: http://www.AnswersInGenesis.org/docs/4057.asp [RETURN TO TEXT]

[63] Truman, R. and Read, B. (1999) *Creation Ex Nihilo Technical Journal*, 13(2), 73-75. [RETURN TO TEXT]

[64] Remine, Walter James (1993) The Biotic Message: Evolution versus Message Theory, Saint Paul Science, Saint Paul, Minnesota. (p. 209) [RETURN TO TEXT]

[65] Ref. 66 p. 100 [RETURN TO TEXT]

[66] Spetner, Lee (1998) NOT BY CHANCE! Shattering the Modern Theory of Evolution, The Judaica Press, Inc. Chapter 4. [RETURN TO TEXT]

[67] Truman, R., http://www.trueorigin.org/dawkinfo.htm; 08-June-1999 [RETURN TO TEXT]

[68] Gitt, Werner (1994) Am Anfang war die Information, Hänssler-Verlag, Neuhausen/Stuttgart, Germany. (currently being reedited and re-translated into English). [RETURN TO TEXT]

[69] Meyer, S. (April 1, 2000). 'DNA and Other Designs', *First Things*, 102, 30-38. On-line: http://www.arn.org/docs/meyer/sm_dnaotherdesigns.htm http://www.firstthings.com/ftissues/ft0004/articles/meyer.html [RETURN TO TEXT]

[70] Anderson, T. J., Klugmann, M., Thomson, C. E., Schneider, A., Readhead, C., Nave, K.A., and Griffiths, I. R., (1999) *Ann N Y Acad Sci*, 883, 234-46. [RETURN TO TEXT]

[71] Sereda, M., Griffiths, I., Puhlhofer, A., Stewart, H., Rossner, M. J., Zimmerman, F., Magyar, J. P., Schneider, A., Hund, E., Meinck, H. M., Suter, U., and Nave, K. A. (1996) *Neuron*, 16, 1049-1060. [RETURN TO TEXT]

[72] Angaut-Petit, D., Toth, P., Rogero, O., Faille, L., Tejedor, F.J., Ferrus, A. (1998) *Eur J Neurosci* 10(2), 423-434. On-line: http://fly.ebi.ac.uk:7081/.bin/fbidq.html?FBrf0104397 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&tool=FlyBase &list_uids=98420134 [RETURN TO TEXT]

[73] Schneider, T. D. and Stephens, R. M. (1990) *Nucleic Acids Res.*, 18, 6097-6100. [RETURN TO TEXT]

[74] Schneider, T. D. (1997) *J. Theor. Biol.*, 189, 427-441. [RETURN TO TEXT]

[75] Kimura, M. (1983) The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge (p. 46, 143, 238) (Referenced in (64) p.227) [RETURN TO TEXT]

[76] Kondrashov, A. S. (1988) *Nature*, 336, Dec. 1, 435-440. (Referenced in (64) p. 228). [RETURN TO TEXT]

[77] Fersht, A. R. (1981) *Proceedings of the Royal Society* (London), B 212, 351-379. (Referenced in (66) p. 92). [RETURN TO TEXT]

[78] Drake, J. W. (1991) *Annual Reviews of Genetics*, 25, 125-146. (Referenced in (66) p. 92). [RETURN TO TEXT]

[79] Ref. 66 p. 92 [RETURN TO TEXT]

[80] Grosse, F., Krauss, G., Knill-Jones, J. W., and Fersht, A. R. (1984) *Advances in Experimental Medicine and Biology*, 179, 535-540. (Referenced in (66) p. 92). [RETURN TO TEXT]

[81] Schneider, T. D. (1997) *Nucleic Acids Res.*, 25, 4408-4415. [RETURN TO TEXT]

[82] Heix, J., and Grummt, I. (1995) *Curr Opin Genet Dev*, 5, 652-656. [RETURN TO TEXT]

[83] Schneider, T. D. (1996) *Methods Enzymol.*, 274, 445-455. [RETURN TO TEXT]

[84] Schultzaberger, R. K. and Schneider, T. D. (1999) *Nucleic Acids Res.*, 27, 882-887. [RETURN TO TEXT]

[85] Johnson, F. B., Krasnow, M. A. (1992) *Genes Dev* 6(11), 2177-2189. On-line abstract: http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=1358759&dopt= Abstract [RETURN TO TEXT]

[86] Ref. 142 p. 417-418. [RETURN TO TEXT]

[87] Naar, A. M., Beaurang, P. A., Robinson, K. M., Oliner, J. D., Avizonis, D., Scheek, S., Zwicker, J., Kadonaga, J. T., and Tjian, R. (1998) *Genes Dev*, 12, 3020-3031. **[RETURN TO TEXT]**

[88] Schneider, T. D. (1999) *J. Theor. Biol.*, 195, 135-137. **[RETURN TO TEXT]**

[89] http://sdb.bio.purdue.edu/fly/gene/spalt.htm **[RETURN TO TEXT]**

[90] http://flybase.bio.indiana.edu/allied-data/lk/interactive-fly/torstoll/toll1.htm **[RETURN TO TEXT]**

[91] Methot, N., Basler, K. (1999) *Europ. Dros. Res. Conf.* 16, 128. http://astorg.u-strasbg.fr:7081/allied-data/lk/interactive-fly/dbzhnsky/slimb1.htm http://astorg.u-strasbg.fr:7081/.bin/fbidq.html?FBrf0110345 **[RETURN TO TEXT]**

[92] Methot, N., Basler, K., *Cell* (1999) 96(6), 819-831 On-line: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&tool=FlyBase &list_uids=99200393 **[RETURN TO TEXT]**

[93] http://www.mh-hannover.de/tagungen/dgfi99/abstract/f18.htm Workshop F. Signal Transduction and Gene Regulation. Abstract F.18 Department of Molecular Pathology, Institute of Pathology, University of Würzburg, Würzburg, Germany **[RETURN TO TEXT]**

[94] Bamshad, M., Watkins, W. S., Dixon, M. E., Le, T, Roeder, A. D., Kramer, B. E., Carey, J. C., and Jorde, L. B. (1999) *Pediatric Res* 45(3), 291-299. On-line: http://www.wwilkins.com/PDR/0031-39983-99p291.html **[RETURN TO TEXT]**

[95] http://www.med.cornell.edu/gradschool/fac/freedman.html Freedman, L. P. **[RETURN TO TEXT]**

[96] http://nowlisten.bmb.psu.edu/pugh/251/13.htm Pugh, F. **[RETURN TO TEXT]**

[97] Ref. (142) Chapter 9 **[RETURN TO TEXT]**

[98] Genuario, R. R., Perry, R. P. (1996) *J. Biol. Chem.* 27(1), 4388-4395. See also: http://www.fccc.edu/research/reports/report95/perry.html **[RETURN TO TEXT]**

[99] http://broadwing.med.unc.edu/wrkunits/5curr/genetics/faculty/crews.html **[RETURN TO TEXT]**

[100] http://www.nmr.chem.uu.nl/users/gert/chap1/chapt1.htm Folkers, G. E., PhD thesis **[RETURN TO TEXT]**

[101] Baniahmad, A., Ha, I., Reinberg, D., Tsai, S. , Tsai, M.-J. and O'Malley, B. W (1993) *Proc. Natl. Acad. Sci. USA* 90, 8832-8836. **[RETURN TO TEXT]**

[102] Cowell, I. G. (1994) *TIBS* 19, 38-42. **[RETURN TO TEXT]**

[103] Fondell, J. D., Roy, A. L., and Roeder, R. G. (1993) *Genes Dev* 7, 1400-1410. **[RETURN TO TEXT]**

[104] Sauer, F., Fondell, J. D., Okhuma, Y., Roeder, R. G., and Jäckle, H. (1995) *Nature* 375, 162-164. **[RETURN TO TEXT]**

[105] http://www.umanitoba.ca/faculties/science/microbiology/345glob.html (Lecture Group 2: Global Regulatory Patterns and Sigma Factors Sporulation in B. subtilis.) **[RETURN TO TEXT]**

[106] Uttenbogaard, M. N., Giebler, H. A., Reisman, D., and Nyborg, J. K. (1995) *J. Biol. Chem.* 270, 28503-28506. On-line abstract: http://zebra.biol.sc.edu/~reisman/ref3.html **[RETURN TO TEXT]**

[107] http://ibms.sinica.edu.tw/Ehtml/pi_e/yslin.html **[RETURN TO TEXT]**

[108] Roberts, S. G. E., Choy, B., Walker, S. S., Lin, Y.-S. and Green, M. R. (1995) *Current Biology* 5, 508-516. **[RETURN TO TEXT]**

[109] Hsu, Y.-S., Tang, F.-M., Liu, W.-L., Chuang, J.-Y., Lai, M.-Y. and Lin, Y.-S. (1995) *J. Biol. Chem.* 270, 6966-6974. **[RETURN TO TEXT]**

[110] http://bioscience.igh.cnrs.fr/1998/v3/d/hipskind/6.htm Hipskind, R. A. and Bilbe, G. (1998) *Frontiers in Bioscience* 3, 804-816, **[RETURN TO TEXT]**

[111] Horlein, A. J., Naar, A. M., Heinzel, T., Torchia, J., Gloss, B., Kurokawa, R., Ryan, A., Kamei, Y., Soderstrom, M., Glass, C. K. (1995) *Nature,* 377, 397-404. **[RETURN TO TEXT]**

[112] Kurokawa, R., Soderstrom, M., Horlein, A., Halachmi, S., Brown, M., Rosenfeld, M. G., and Glass, C. K., (1995). *Nature,* 377, 451-454. **[RETURN TO TEXT]**

[113] Nohturfft, A., DeBose-Boyd, R. A., Scheek, S., Goldstein, J. L., and Brown, M. S. (1999) *Proc Natl Acad Sci U S A,* 96, 11235-11240. **[RETURN TO TEXT]**

[114] Scheek, S., Brown, M. S., and Goldstein, J. L. (1997) *Proc Natl Acad Sci U S A,* 94, 11179- 11183. **[RETURN TO TEXT]**

[115] Griffin, C., Flouriot, G., Sonntag-Buck, V., and Gannon, F., (1999) *Mol Endocrinol,* 13, 1571-1587. **[RETURN TO TEXT]**

[116] Griffin, C., Flouriot, G., Sonntag-Buck, V., Nestor, P., and Gannon, F., (1998). *Endocrinology,* 139, 4614-4625. **[RETURN TO TEXT]**

[117] Srere, P. A. and Knull, H. R. (1998) *TIBS* 23, 319-320. **[RETURN TO TEXT]**

[118] Zheng, M., Doan, B., Schneider, T. D. and Storz, G. (1999) *J. Bacteriol.,* 181, 4639-4643. **[RETURN TO TEXT]**

[119] Anderson, T. J., Schneider, A., Barrie, J. A., Klugmann, M., McCulloch, M. C., Kirkham, D., Kyriakides, E., Nave, K. A., and Griffiths, I. R. (1998) *J Comp Neurol,* 394, 506-519. **[RETURN TO TEXT]**

[120] Niemann, S., Sereda, M. W., Rossner, M., Stewart, H., Suter, U., Meinck, H. M., Griffiths, I. R., and Nave, K. A., (1999) *Ann N Y Acad Sci,* 883, 254-261. **[RETURN TO TEXT]**

[121] Keating, P., Harrison, D., T.H. Morgan School of Biological Sciences, University of Kentucky: http://flybase.bio.indiana.edu/.data/docs/abstr/1998dros/f5647.html **[RETURN TO TEXT]**

[122] Suzuki, D., Miyata, T., Nangaku, M., Takano, H., Saotome, N., Toyoda, M., Mori, Y., Zhang, S- Y., Inagi, R., Endoh, M., Kurokawa, K. and Sakai, H. (1999) *Journal of the American Society of Nephrology*, 10(12), 2606-2613. On-line: http://www.lrpub.com/JASN/1046-667312-99p2606.html **[RETURN TO TEXT]**

[123] Yockey, H. P. (1974) *J. Theor. Biol,* 46, 369-406. **[RETURN TO TEXT]**

[124] Yockey, H. P. (2000) *Computers and Chemistry*, 24, 105-123. **[RETURN TO TEXT]**

[125] Kohler, M., Kornau, H. C., and Seeburg, P. H., (1994) *J Biol Chem,* 269, 17367-17370. **[RETURN TO TEXT]**

[126] Roeder, R. G. (1996) *Trends Biochem. Sci.* 21, 327–335. **[RETURN TO TEXT]**

[127] Orphanides, G., Lagrange, T. and Reinberg, D. (1996) *Genes Dev.* 10, 2657–2683 **[RETURN TO TEXT]**

[128] Dandekar, T., Beyer, K., Bork, P., Kenealy, M. R., Pantopoulos, K., Hentze, M., Sonntag-Buck, V., Flouriot, G., Gannon, F., and Schreiber, S., (1998) *Bioinformatics,* 14, 271-278. **[RETURN TO TEXT]**

[129] Dawkins, R. (1986). The Blind Watchmaker, Penguin Books, London. [RETURN TO TEXT]

[130] Ref. 66 p. 111 [RETURN TO TEXT]

[131] Morris, S. C. (2000) *Cell*, 100, 1-11. [RETURN TO TEXT]

[132] Chen, L., DeVries, A. L, and Cheng, C.-H. (1997) *Proc. Natl. Acad. Sci. USA* 94, 3817-3822. [RETURN TO TEXT]

[133] Beschin, A., Bilej, M., Brys, L., Torreele, E., Lucas, R., Magez, S., and De Baetselier, P. (1999) *Nature*, 400, 627-628. [RETURN TO TEXT]

[134] Lawn, R. M., Schwartz, K., and Patthy, L. (1997) *Proc. Natl. Acad. Sci. USA* 94, 11992-11997. [RETURN TO TEXT]

[135] Kawasaki,M. (1996) *Biol. Bull.* 191, 103-108. [RETURN TO TEXT]

[136] Guo,Y.-X., and Kawasaki, M. (1997) *J. Neurosci.* 17, 1761-1768. [RETURN TO TEXT]

[137] Hildebrand, J. G., and Shepherd, G. M. (1997) *Ann. Dev. Neurosci.* 20, 595-631. [RETURN TO TEXT]

[138] Sarfati, J. D. (1998) *CEN Tech.J.* 12(3), 263-266; online at: http://www.answersingenesis.org/index.asp?id=3991, http://www.sanger.ac.uk/Projects/C_elegans/ [RETURN TO TEXT]

[139] Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.* 188, 415-431. [RETURN TO TEXT]

[140] http://www-lecb.ncifcrf.gov/~toms/paper/ri/latex.htm [RETURN TO TEXT]

[141] Stephens, R. M. and Scheider, T. D. (1992) *J. Mol. Biol.*, 228, 1124-1136. [RETURN TO TEXT]

[142] Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., Watson, J. D. (1994) Molecular Biology of The Cell, 3rd ed. Garland Publishing, Taylor & Francis Group. [RETURN TO TEXT]

[143] Hengen, P. N., Bartan, S. L., Stewart, L. E., Schneider, T. D. (1997) *Nucleic Acids Research*, 25, 4994-5002. [RETURN TO TEXT]

[144] Hoyle, Fred, "Mathematics of Evolution" , Acorn Enterprises LLC, Memphis, Tennesse, USA. 1999. Use equation (1.6) on p. 11. [RETURN TO TEXT]

[145] Shapiro, R., Origins: A Skeptic's Guide to the Creation of Life on Earth, Bantam Books, USA., Feb., 1987. See p.157 - 160. [RETURN TO TEXT]

[146] Lodish, H., et. al., Molecular Cell Biology, 4th ed., W. H. Freeman and Company, USA, 2000. See p.187. [RETURN TO TEXT]

[147] Yockey, H.P., Information Theory and Molecular Biology, Cambridge University Press, Great Britain, 1992. [RETURN TO TEXT]