Lecture 2 From Synapse to Circuit

The most interesting, familiar and nevertheless remarkable example of selforganisation is life itself. Here the selforganising principle is Darwinian evolution by natural selection, and the key mechanism is selfreplication. Because selfreplication is inevitably imperfect, progeny exhibit heritable variations from their parents. These variations may affect survival and reproduction (i.e. net replication rates), and if environmental resources are limited, competition between variants leads to “survival of the fittest”, and gradual but inevitable improvements in replicative ability.

But there is an obvious paradox. Evolution cannot happen without variation, but if progeny are not identical to their parents, replication has not occurred. Resolution of this paradox, together with insights into a variety of related issues, can be gained by considering a model of molecular evolution developed by Manfred Eigen.

In life, the replicating entities are molecules, polynucleotides. The 2 key features of polynucleotides that allow Darwinian evolution are (1) these molecules are linear sequences of several different monomers, which can occur in essential infinite possible arrangements. (2) Crick-Watson base pairing allows fairly exact sequence copying.

Polynucleotide copying requires favorable conditions which must have occurred at least once in the universe since the big bang, here on earth. . There has to be an appropriate supply of monomers, these monomers have to polymerized on the parent template, and the new sequence must unzip from the old. At the dawn of life, in the so called “RNA World”, these molecules were ribonucleic acids. Crick-Watson base pairing allows RNA molecules to fold up on themselves into a variety of sequence-specific complex shapes. Sometimes these shapes have catalytic properties, so they function as ribozymes. Ribozymal activity in principle can foster self-copying. The RNA world quickly gave rise to DNA-based life. DNA sequences encode specific proteins which are even more effective at catalysing selfreplication.

The selfreplication of a particular sequence S (for example, 001100…) of length n from a supply of monomers M can be represented as a chemical reaction:

$nM + S \text{-}^{k}\text{--}\ 2S$

In words, n monomers (either 0 or 1) come together with a template sequence S to generate a new identical sequence plus the original template sequence with “rate constant” (probability per unit time) k.

From standard chemical kinetics we can write the following equation for this reaction (writing S for the concentration of S, M for the monomer concentration):

$dS/dt = kM^nS.$

In words, the rate of appearance of S is proportional to its concentration. The concentration of monomers is fixed (i.e. selfreplication does not deplete the monomer pool) so we can combine k with M^n to form a new constant ϕ. We call ϕ the “fitness” or

“Malthusian parameter”. It represents how fast the sequence selfreplicates. (Remember that Darwin was inspired by Malthus’ work on the tendency of human populations to grow exponentially, and “fitness” in Darwinian theory is not acquired at the gym). As we saw in lecture 1, this can only be true if S increases exponentially:-

$S = S_0 \exp \phi t$

S_0 = the initial concentration of the sequence. This relationship is shown in the graph. The concentration of S rapidly increases, like rabbits in a limitless field of succulent grass.

Graph here

Of course, the rabbits soon run out of grass , and settle down to a fixed plateau or celing population number which is set by the rate at which new grass grows. We can represent this mathematically by adding a factor that is 1 when the population is small, and 0 when the population hits the ceiling S_m (S_m is called the “carrying capacity” in ecological theory):

$dS/dt = \phi S\ (S_m - S)/\ S_m$

If we write $S/\ S_m = x$ (i.e. we define the population number as a fraction of the maximum possible population number), this becomes

$dx/dt = \phi\ x(1-x)$

which is known as the logistic equation. Its solution is

$x(t) = 1/[1+(1/x_0-1)\exp-\phi t]$

This equation is plotted in the next **graph**. It shows, as expected, that the number of sequences increases exponentially at first, but then slows down, and reaches a steady plateau.

Now let us consider 2 different sequences S1 and S2 (eg 001100 and 001101) made from the same supply of monomers but at different rates ϕ_1 and ϕ_2. Assuming $\phi_1 > \phi_2$ (i.e. that S1 replicated faster than S2), we will get the following behavior:

(Graph)

This happens because as monomers are progressively depleted, the rate of synthesis of both sequences slows BY THE SAME AMOUNT.

Even if the starting concentration of the less fit sequence is much greater than that of the most fit, nevertheless it will always be eventually completely displaced by the other:

(Graph)

Thus, the fittest always wins and the losers go extinct! (In finite populations chance may favor losers, as we all know). This behavior is fundamental to evolution and arises because selfreplication gives rise to potentially exponential growth.

[**Advanced material .** Consider a more general version of the growth law: $dx/dt = \phi x^m$. We have just considered the selfreplication law, where m = 1. If instead m = 0, growth would be linear with time, not exponential. The equilibrium population would be composed of a mixture of S1 and S2, in a ratio given by their respective fitnesses. Since in real life differences in fitness are rather small, after many rounds of replication and mutation, all possible sequences would occur, and genomes would randomly drift rather than evolving. If m = 2 or more, then growth is “superexponential”, so that even less fit sequences would take over, merely if they have a high initial concentration; the rare sequences arising by mutation wouldn’t stand a chance, however well they self-replicate]

The picture remains essentially the same if we consider not just 2 sequences but all 2^n possible sequences made from the same pool of monomers. Of all these sequences, only the fittest will survive, all the others are consigned to the scrap heap. However, as n gets larger there are so many possible different sequences that even in a very large population it gets increasingly unlikely that any particular sequence will ever be found. (In a large vat one might have an Avogadro’s number of sequences all together, but that is still far less than the number of possible different sequences with n = 100.) So evolution might fail because of chance events. This situation is avoided if there are mutations.

Eigen considered what would happen if we include replication errors, or mutations. An example would be the incorrect synthesis of 001101 on a template 001100. Suppose we start with 100% 001100 and at one point in time an error occurs. If the erroneous sequence 001101 selfreplicates faster than the parent sequence, we might expect, from the above argument, that, even though it appears only once, it will win. This is broadly true, but as we will see not always. It is because errors are inevitable, and some errors are superior to their parent, that evolution is a selfimproving process, and why once started it can generate organisms of remarkable complexity and sophistication.

Eigen considered a fixed but large population of selfreplicating strings of fixed length n. For simplicity we will suppose that the monomers can only be 2 possible sorts, 0 and 1. There are therefore 2^n possible sequences. Eigen asked whether a specific starting sequence can indefinitely survive repeated rounds of inexact replication, or alternatively if the sequence gradually gets corrupted by the accumulation of errors. Let us consider first a simplified picture of the situation. Because the total population is fixed, the replication of the strings must be offset by disintegration (or “death”) of the strings

(otherwise the population would continue to grow). Call this death rate d. Suppose the probability that strings disintegrate or die, is d. Suppose that the error rate (or mutation rate) for copying single bases is e. A particular string has 3 possible fates. (1) It can die, with probability d. (2) it survives (with probability 1-d) and accurately replicates, which will occur with probability $(1\text{-}e)^n$ (since this is the probability that correct copying occurs at all n sites simultaneously) (3) it survives (probability (1-d) but is incorrectly copied (i.e. a mutation occurs), which occurs with probability $(1\text{-}(1\text{-}e)^n)$, since this is the probability that at least one mistake occurs.
Because these 3 possibilities give rise to 0, 2 or 1 copy of the string respectively, the number of exact strings produced is given by the sum of the terms 0d, $2(1\text{-}d)\ (1\text{-}e)^n$ and $(1\text{-}d)\ (1\text{-}(1\text{-}e)^n)$, which is $(1\text{-}d)([1\text{-}e]^n + 1)$. If the sequence is to survive, at least one exact string must be formed every time the strings replicate , so the survival criterion is given by

$(1\text{-}d)([1\text{-}e]^n + 1) = 1$

Rearranging and taking natural logs

$n = \ln d/(1\text{-}d)/\ln (1\text{-}e)$

Now if e is a small number, $\ln (1\text{-}e) \sim e$

So $n = \ln [d/(1\text{-}d)]/e$

Thus there is a reciprocal relation between the length of a selfreplicating sequence and the error rate for copying its individual bases. What does this result mean? It shows that if the error rate is too high, or the string is too long, then sequence information is lost. “Replication” continues but it is too inaccurate to preserve sequence information, and the sequences that are found become essentially random. Evolution has stopped.

The above probability argument is oversimplified, because it does not explicitly include the various replication rates of the different sequences. Eigen made a more detailed analysis, using the idea of a “master sequence”. The master sequence is simply the sequence that is best at selfreplicating (for example, because it codes for better replicase activity). It is better than the other possible sequences by a factor s (the selectivity). Eigen deduced the following equation (a sketch of the derivation is given in the appendix to this lecture):

$N < (\ln s)/e$ (derived in the Appendix).

Again, this shows the reciprocal relation between the error rate and the maximum string length that is compatible with Darwinian evolution.

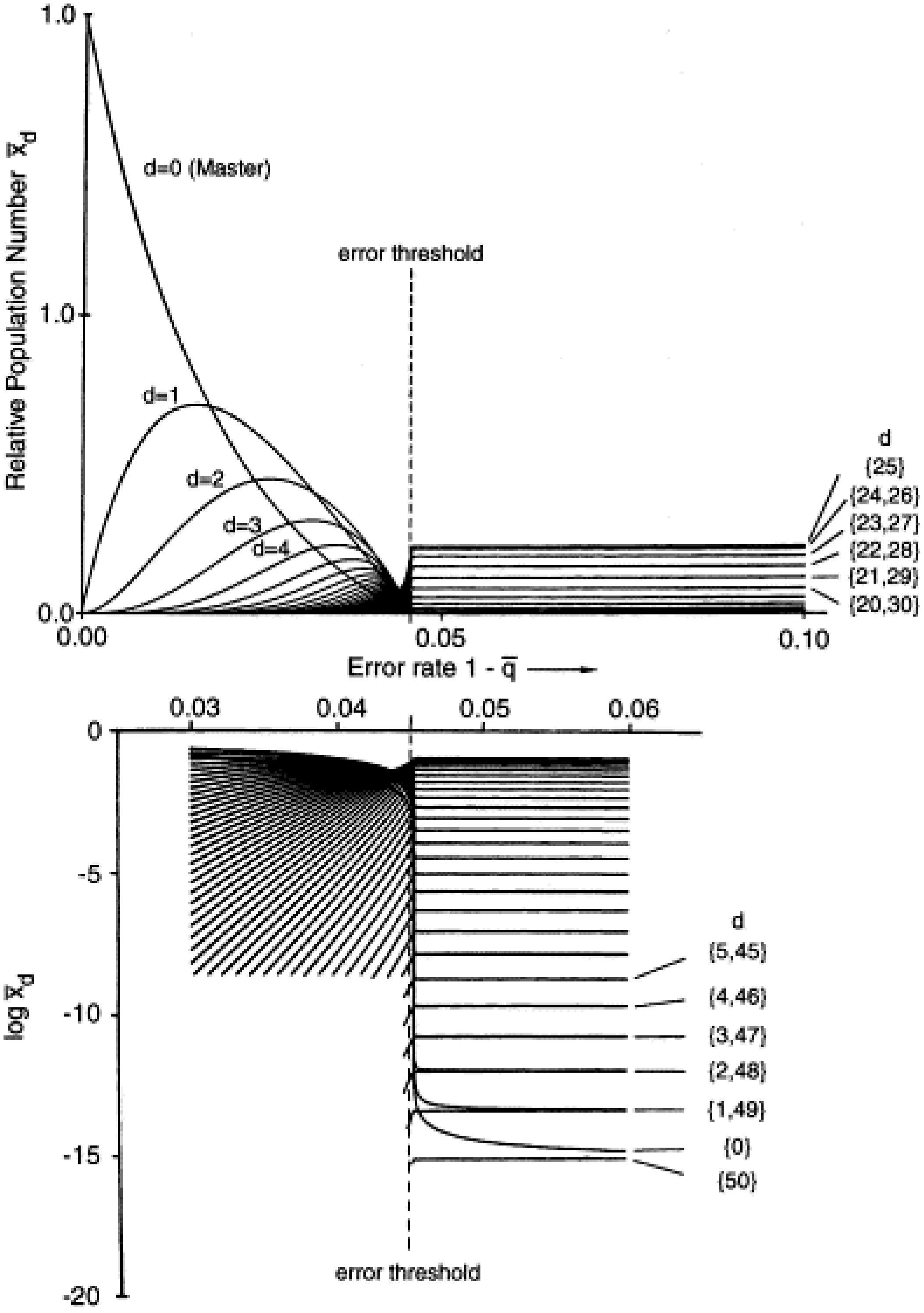
In the ferromagnet, magnetism results from 2 opposing forces, the ordering effect of the cooperative interaction between nearby spins, and the disordering effect of thermal agitation. In molecular evolution the ordering effect is preferential replication of the master copy, which tends to maintain large numbers of this sequence in the total population. The disordering process is mutation – inaccurate base matching. Some small fraction of the attempted exact replications of the master sequence turn out incorrect sequences. These incorrect sequences in turn will be (usually exactly) replicated, leading to occasional even less accurate copies. However, the gradual accumulation of these progressively less accurate copies is winnowed out by natural selection, which allows the master copy to replicate more rapidly. But this winnowing effect becomes less effective as the error rate increases. The "tipping point" between selection and mutation occurs at lower and lower error rates as the polynucleotide chain gets longer, because there are more and more ways for errors to occur (since the error rate is defined per base).

Computer simulations (see Figure) show this in more detail.

Figure Legend

The graphs show calculations of the behavior of the Eigen model for various error (mutation) rates – the bottom graph is a blow-up of the top graph near the critical error rate, using a log ordinal scale.
A population of self-replicating and evolving polynucleotide molecules, all 50 bases long and either 1 or 0 is considered (eg purine or pyrimidine). The total number of molecules is kept constant (by diluting the mixture as it replicates, and discarding excess). One particular sequence, the "master copy" or "main copy" replicates ten times faster than all the other sequences. *d* is the number of bases which differ from the master sequence (so called "Hamming distance"). The ordinate shows the fraction of the total number of molecules belonging to a particular Hamming class (i.e with a particular value of *d*), x_d. Thus the curve labeled *d*=0 shows the fraction of the total that have the master sequence, the curve labeled d=1 shows the fraction that have sequences differing in only one base pair from the master etc. The abscissa shows the per-base error rate for replication (q is the "quality factor", 1-q is the error rate).

As expected, with zero error rate the master copy completely wins the competition, since it replicates faster than all the others ($x_0 = 1$). As the error rate increases, the fraction of the total that has the master sequence decreases, since as the master replicates, it sometimes makes errors, generating mutant copies. Multiple mistakes (d=2,3…) get increasingly more likely as the error rate increases. However, there is a sudden qualitative change in the curves at the critical error rate 0.045, the error threshold. Beyond the error threshold, sequences become essentially random. Because there are far more ways to make 25 mistakes than to make either no mistakes or 50 mistakes, the population is dominated by sequences with 25 or so differences from the master. In fact, since the choice of base (0 or 1) at each position is random, the outcome beyond the error threshold is essentially like tossing 25 coins, and laying the coins out in a row in the order they are tossed, and is given by a binomial distribution (see lecture on synaptic transmission). The critical error rate e_c agrees with the value predicted by the formula $e_c = \ln s/n$, since in this case n= 50 and s = 10.



The ordinate shows the error rate. The abscissa shows what fraction of the total population of molecules has 0, 1, 2 etc errors. At an error rate of 0, all the molecules have the correct, master, sequence. However, as the error rate increases, fewer and fewer of the molecules are correct, and copies that have 1 or more mutations become more frequent. Still, at error rates well below the threshold ln s/n, the population is composed of a mixture of sequences, all more or less closely related (or identical) to the master sequence. Eigen refers to this family of successful sequences as the “quasispecies”, but the success of the slightly wrong sequences is not due to their own replicative abilities but to the efficiency of the master, somewhat like an extended family of layabouts who sponge on one industrious member.

If the error rate increases still further, a rather dramatic thing happens : the quasispecies suddenly disappears (or, more exactly, it occurs no more frequently than would occur by chance). In the graph shown this occurs quite precipitously at an error rate of 4.5 %. The simulations were done for a chain length of 50, and the selective value s was 10 (meaning the master sequence replicated ten times faster than all the other sequences) so the predicted error threshold is ln10/50 = .045 , as observed.

Because order (in this case, sequence information) is lost at the error threshold, the Eigen model has a phase transition. In fact, it can be shown that the Eigen model is closely related to the Ising model we considered in the last lecture.

This analysis illuminates our initial paradox: that Darwinian evolution requires both replication (accurate copying) and mutation (inaccurate copying). According to Eigen, Darwinian evolution is only possible if the ratio of the sequence length to the error rate is below some critical value. Since in Nature fitness differences are quite small, and the dependence on s is only weak (logarithmic), the string length must be less than the reciprocal error rate.

In real evolution the environment is gradually changing (partly because other organisms are also evolving), and the sequence that replicates most efficiently also changes. Clearly, if the master sequence is to mutate rapidly enough to keep up with these changes, its length must be shorter than that imposed by the Eigen criterion. On the other hand, the shorter the sequence, the smaller the number of possible selfreplication strategies available. If sequences are too long, evolution is not possible. If sequences are too short, evolution may also be impossible. The fact that evolution has occurred successfully for billions of years shows that despite drastic changes (such as immense meteor impacts) the earth is a relatively benign, stable place. We will see that such stability is also essential for learning to occur.

Implications of the Eigen theory.

1. Origins of life. In the RNA world, the maximum length of sequences that could evolve was very limited, because the per-base error rate for copying RNA is quite high – on the order of 1%. Thus the first, spontaneously-formed, selfreplicase must have been shorter than about 200 bases. Do such replicases exist? Recent work by David Bartel has shown that they do, so life *could* have started by itself. However, this sequence must itself have assembled by chance, and the chances of spontaneously assembling a 200 base RNA are only 1 in 4^{200}. Offsetting this are

(1) Avogadro’s number (2) the extent of the surface of the earth (3) the total number of earthlike planets in the universe [the latter being an example of the Anthropic Principle].

2. RNA and DNA. Shortly after its origin, life switched from RNA to DNA. The reason is that DNA can be replicated much more accurately, allowing the exploration of a far larger number of sequences (indeed, probably more than can be explored in the lifetime of the earth). However, the adoption of DNA as the information-storage medium carried the disadvantage that an elaborate system for decoding and protein synthesis had to evolve. Probably this was achieved via a series of intermediate RNA/DNA hybrid architectures, and indeed traces of this are found even in modern life-forms. Modern cells can replicate DNA with per-base error rates as low as 10^{-9}, allowing genomes as large as a gigabyte.

3. Sex. Polynucleotides replicate asexually. Sex allows 2 independent mutations to come together in the same genome. If together these confer a fitness advantage they will be selected. Individually they may have no effect on fitness, and the chance that the 2 mutations would appear solely by asexual replication (before the environment changes yet again) is very low. In sexual reproduction, this happens much faster. Essentially there is a cooperative exchange of information between genomes. However, there is the disadvantage that organisms switching from asexual to sexual reproduction immediately lower their fitness by a factor of 2 (since it takes 2 to tango). Understanding why the former advantage outweighs the latter disadvantage is a classic problem in evolution.

4. RNA viruses. An experimental test of Eigen’s theory was provided by RNA viruses (such as HIV). As predicted it is found that these viruses are always shorter than the limit set by Eigen’s criterion. However, they are not much shorter. Thus RNA viruses behave as “quasispecies”. The commonest sequence is not particularly numerous. Other sequences related to the commonest sequence by a few mutational differences are, colelectively, more numerous (though individually they may be less frequent). So the virus particles present in the host are a broad spectrum of related genomes. This gives the virus a great advantage in avoiding the immune response of the host.
This provides a novel therapeutic attack – a drug which increases the mutation rate can tip the viral population over the error threshold, so it mutates itself out of existence.

5. Complexity of life. Genomes are limited in length by the error threshold, but if they can grow larger they can explore a greater range of replicative strategies. Other things being equal, this creates a strong upward pressure on genome size. This means that as evolution proceeds there is a tendency for life to get more complex. Indeed, complexity has been defined as that which increases when systems selforganise. As populations evolve, they incorporate information about their environment into their genomes. This may show up as a size increase. It can also show up as a reduction in the fraction of the genome that is “noise” or “junk”.

References

Johnston et al, “RNA-catalysed RNA polymerization: accurate and general RNA-templated primer extension. Science. 2001 may 18. 292: 1319 –1325

Adami et al. Evolution of biological complexity. Proc Nat Acad. Sci. 97, 4463-4469

Adami,C. Artificial Life (book)

M. **Eigen Error** catastrophe and antiviral
strategy **PNAS**, October 15, 2002; 99(21): 13374 - 13376.

Appendix. The following discusses the derivation of the Eigen equation N < (ln s)/(1-e), which sets the maximum length of an evolving polynucleotide. It is not necessary to study or understand this appendix, which requires some familiarity with eigenvectors.

Review of Eigen’s Analysis

Eigen considered the (possibly erroneous) replication of polynucleotide strings of fixed length n. Individual binary strings (eg 001011001100..) occupy a corner of a 2^n dimensional hypercube. He considered a population of strings, i.e. a population of hypercubes each of which has a single marked corner. The total population number *m* is constrained to be constant by a dilution process, which offsets any overall increase or decrease due to replication or degradation by removing strings proportionate to their number in the population (“multiplicative normalization”, see below). In the absence of error each string increases at a rate proportional to its “fitness”, and the competitive constraint ensured that only the fittest sequence survives. Mathematically this arises because the numbers of each string can be represented as a concentration vector **x** which is multiplied by a diagonal fitness matrix **F**. Normalisation requires that $\mathbf{Fx} = \lambda\mathbf{x}$, so the population evolves to the leading eigenvector of **F** (λ is the eigenvalue associated with the eigenvector). Since **F** is diagonal, a single sequence wins, as discussed in the lecture (the eigenvectors of a diagonal matrix are1,0,0….etc).

If replication can be erroneous, we must consider all possible transitions between all possible sequences, represented by a transition matrix **T.** In the case that the probability that a base be incorrectly copied is independent of string position, and represented by *e*, this transition matrix can be represented as the product of the fitness matrix and a mutation matrix **M** whose elements are $(1\text{-}e)^{n-d}\, e^{d}$ (*d* is Hamming distance, i.e. the number of bases that are incorrect in a given string compared to the master). The steady state **x** is now given by the leading eigenvector of **MF.** To find this eigenvector Eigen introduced the zeroth-order perturbation approximation for λ_{max}, which in the case that all sequences but one (the fittest) have the same fitness, is s, the selective advantage of the “master copy”. (Similar but more complicated results follow for the general case where all fitnesses differ). In other words, since we know that the exact λ_{max} for the zero error case is s, it should still be approximately s provided the off-diagonal terms of **MF** are close to zero, as they will be for small mutation rates. This approximate eigenvalue is then used together with the exact transition matrix **MF** to find an approximation to the leading eigenvector. It should be noted that these approximations become better as n increases, and are almost exact for $n > 10$. If one knows the eigenvalues, the eigenvectors can be calculated using the equation $(\mathbf{MF} - s\mathbf{I})\mathbf{x} = \mathbf{0}$. This equation involves all the (unknown) stationary concentrations of each possible sequence. An error catastrophe would exist if despite the selective advantage of the master copy, all sequences were represented with equal probability. Setting all the concentrations equal, one finds Eigen's celebrated critical error rare $e_c = (\ln s)/n$. If $0 < e < e_c$, a "quasispecies" results: the population is represented by a mixture of marked strings, in which the master copy still outnumbers all other sequences, but sequences that are only small Hamming distances away from the master are quite likely. The two most important conclusions of this analysis are (1) evolution is favored in the range $0 < e < e_c$ (since if the fitness landscape changes to favor a sequence that is represented within the quasispecies, a shift to that sequence is rapid (2) evolution is not possible for $e > e_c$, so innovations (such as the RNA – DNA transition) that give smaller e_c allow larger n's, and a greater range of adaptation. These results all follow from the 2 basic assumptions of the model: (1) sequences exist as hypercube corners (2) copying errors allow transitions between corners; these transitions are confined to the edges of the hypercube (the interior of the hypercube corresponds to energetically unfavorable states, such as incorrect hydrogen bond formation).