Inverse relationship between genetic diversity and epigenetic complexity

Shi Huang，Ph.D.

The Burnham Institute for Medical Research

10901 North Torrey Pines Roads

La Jolla, CA 92037

shuang@burnham.org

Tel: 1-858-646-3120

Fax: 1-858-646-3192

Key words: Genetic equidistance result, evolution, molecular clock, maximum genetic diversity hypothesis, epigenetic complexity, Neo-Darwinian hypothesis

Running title: The maximum genetic diversity hypothesis

Abstract

Early studies of molecular evolution revealed a correlation between genetic distance and time of species divergence. This observation provoked the molecular clock hypothesis and in turn the 'Neutral Theory', which however remains an incomplete explanation since it predicts a constant mutation rate per generation whereas empirical evidence suggests a constant rate per year. Data inconsistent with the molecular clock hypothesis have steadily accumulated in recent years that show no correlation between genetic distance and time of divergence. It has therefore become a challenge to find a testable idea that can reconcile the seemingly conflicting data sets. Here, an inverse relationship between genetic diversity and epigenetic complexity was deduced from a simple intuition in building complex systems. Genetic diversity, i.e., genetic distance or dissimilarity in DNA or protein sequences between individuals or species, is restricted by the complexity of epigenetic programs. This inverse relationship logically deduces the maximum genetic diversity hypothesis, which suggests that macroevolution from simple to complex organisms involves a punctuational increase in epigenetic complexity that in turn causes a punctuational loss in genetic diversity. The hypothesis explains a diverse set of biological phenomena, including both for and against the correlation between genetic distance and time of divergence.

> The only real valuable thing is intuition……The whole thing of science is nothing more than a refinement of everyday thinking.
>
> - Albert Einstein

Introduction:

It is remarkable that the human mind is able to comprehend nature. The scientific understanding of nature is largely based on mathematics. Since mathematics is premised on axioms or self-evident intuitions, it can be easily inferred that intuition is the ultimate foundation of science. The relationship between intuition and a natural phenomenon is sometimes indirect or follows the hierarchy from intuition to mathematics, to physics, to chemistry, and to biology. But it can also be direct, for example, Newton's three laws of motion were originally postulated as 'axioms'. Intuition may directly impact the science of biology without going through the bridge of mathematics, or chemistry, or physics, although such an intuition-based law of biology has yet to be uncovered. An intuition-based theory is true on its own logical coherence (like a mathematical proof) and does not in principle need validation from empirical data. In contrast, no amount of experimental data could prove a provisional theory that is based on empirical observations.

The molecular clock hypothesis is an essential part of the modern evolution theory. The hypothesis was triggered by the empirical observation of a correlation between genetic distance as measured by DNA or protein sequence dissimilarity and time of species divergence as inferred from fossil records. Two kinds of sequence alignment can be made using the same set of sequence data. The first aligns a recently evolved organism such as a mammal against those that evolved earlier such as amphibians and fishes. The second aligns an outgroup organism such as fishes against those sister species that appeared later such as amphibians

and mammals. The first alignment indicates a linear correlation between genetic distance and time of divergence, implying indirectly a constant mutation rate among different species. The second alignment shows the genetic equidistance result where sister species are approximately equidistant to the outgroup. This directly triggered the idea of constant mutation rate among different species. Since both alignments use the same sequence data set, either alone is sufficient to reveal any information on genetic distance. But the data that most directly and obviously triggered the interpretation of constant mutation rate is the genetic equidistance result.

The molecular clock hypothesis was first informally proposed in 1962 based largely on data from the first alignment (Zuckerkandl and Pauling, 1962). Margoliash in 1963 performed both alignments and made a formal statement of the molecular clock after noticing the genetic equidistance result (Margoliash, 1963). “It appears that the number of residue differences between cytochrome c of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged. If this is correct, the cytochrome c of all mammals should be equally different from the cytochrome c of all birds. Since fish diverges from the main stem of vertebrate evolution earlier than ether birds or mammals, the cytochrome c of both mammals and birds should be equally different from the cytochrome c of fish. Similarly, all vertebrate cytochrome c should be equally different from the yeast protein.”

The molecular clock hypothesis asserts that the rate of amino acid or nucleotide substitution is approximately constant per year over evolutionary time and among different species. Two different species are thought to gradually accumulate mutations over time since their most recent common ancestor. Their genetic distance in ancient times is thought to be smaller than their distance today. None of these assertions are based on intuitions or could be considered as self-evident. Nor do they have direct experimental support. They are all ad hoc interpretations of the genetic equidistance result.

The empirical observation of an apparently constant mutation rate has provoked the 'Neutral Theory'. But this theory is now widely acknowledged to be an incomplete explanation. For example, Ayala noted: "The theoretical foundation originally proposed for the clock, namely the neutrality theory of molecular evolution, is untenable. The vagaries of molecular rates of evolution have contributed much to invalidating the theory."(Ayala, 1999). Pulquerio and Nichols noted: "The 'Neutral Theory' is not a complete explanation, however. For example, it predicts a constant substitution rate per generation, whereas empirical evidence suggests something closer to a constant rate per year." (Pulquerio and Nichols, 2007).

The constant mutation rate interpretation of the genetic equidistance result represents an over-interpretation of the actual result, since the result shows merely the outcome of evolution and says nothing about the past mutation process. In fact, the equidistance result has been found to be independent of mutation rate variations (Huang, 2008a). Violation of rate constancy does not mean violation of the equidistance result and the equidistance result does not necessarily mean rate constancy. The constant mutation rate interpretation of the equidistance result is a non-testable tautology and is not a real scientific explanation of the equidistance result (Huang, 2008a).

The common practice of relative rate tests that often interprets small deviations from an exact equidistance as being statistically significant is in fact flawed as it does not consider sampling variations (Huang, 2008a). It also overlooks the striking fact that the deviations are rarely large. If the real phenomenon here is non-equidistance with equidistance being coincidental, one would expect to see much larger variations in distance. Thus, the data shows that the real phenomenon here is equidistance while the small deviations from exact equidistance are coincidental and non-significant sampling variations.

Although there clearly exists a correlation between genetic distance and time of divergence, such correlation is not universal and is often violated as more data became known in recent years. Numerous studies based on extant organisms have questioned the constancy

of mutation rate (Ayala, 1999; Ho and Larson, 2006; Pulquerio and Nichols, 2007). A study of DNA and protein sequences of ancient fossils challenged a fundamental premise of the modern evolution theory (Huang, 2008b). It shows that genetic distance had not always increased with time in the past history of life on Earth. Another study showed that the genetic distance among flowering plants is much greater than that among mammals, even though flowering plants have evolved for similar amount of time as mammals (Huang, 2008a). The genetic distance between two subpopulations of medaka fish that had diverged for ~ 4 million years is 3-fold greater than that between two different primate species (humans and chimpanzees) that had diverged for 5-7 million years (Kasahara et al., 2007). The genetic distance measured on genealogical timescales (< 1 million years) is often an order of magnitude greater than that on geological timescales (> 1 million years) (Ho and Larson, 2006), suggesting that genetic distance measured in evolutionary time is independent of actual mutation rate measured in real time.

The molecular clock hypothesis was originally an ad hoc idea triggered by the genetic equidistance result and remains unsupported by any other independent facts despite the effort of the past 45 years. While it may explain the correlation between genetic distance and time of divergence, it clearly cannot explain the frequent factual violations of the correlation. A new and more complete idea is needed that must be able to reconcile the seemingly conflicting data sets. Here, a simple intuition in building complex systems was used to derive a novel principle of biology, the inverse relationship between genetic diversity and epigenetic complexity. This principle or its logical deduction, the maximum genetic diversity hypothesis, was found to explain a large set of biological data, including both for and against the correlation between genetic distance and time of divergence.

An intuition in building complex systems/machines

It is a self-evident intuition that simpler systems/machines can tolerate more variations/choices in building blocks. The more complex the system, the more restriction would

be placed on the choice of building blocks. A one-story house can be build by all varieties of bricks but only the stronger ones among them can qualify for a 100-story building because the weaker ones cannot withstand the weight of a 100-story building. The number of choices of different materials for constructing a toy bicycle is much greater than that for a space shuttle.

Inverse relationship between genetic diversity and epigenetic complexity

The building blocks for biological organisms are DNAs. The complexity of organisms is reflected by the ways a set of DNAs is used to make a cell or an organism with multiple distinct cell types. The more the cell types, the more the number of ways of using the same set of DNAs, and the more complex the organism. Phenotypes are determined by the primary sequence of DNAs or genotypes as well as by the ways by which DNAs are used or expressed, often termed epigenotypes or epigenetic programs. Each cell type represents a distinct epigenetic program of the same genotype. Cell types with distinct functions differ only in epigenotypes but not in genotypes (a small number of special cell types such as antibody producing cells are exceptions).

From the self-evident intuition of building complex machines, it is easy to deduce an equivalent principle in constructing biological organisms. Thus, simpler organisms with low epigenetic complexity can tolerate more variations in DNAs or have higher genetic diversity. Genetic diversity is defined here as genetic distance or dissimilarity in DNA or protein sequences between different individuals or species. Simple organisms are built more by the primary function of a gene rather than by a specific expression pattern of the gene. A gene may only have one expression pattern in simple organisms and many variants of the gene may be able to fit within that one expression pattern. In contrast, when an organism is built by multiple distinct gene expression patterns or cell types, the variation in gene sequence would be necessarily restricted.

The reason is easy to understand. If cell type A is determined by expression pattern X and cell type B by pattern Y of the same gene, a mutational variant of the gene must be compatible with both expression pattern X and pattern Y. Such multilevel compatibility reduces the number of variants of the gene that can meet the multiple requirements. If ten mutational variants can fit with expression pattern X, then may be only three of the ten would fit with both patterns X and Y. The more expression patterns or cell types or functional pathways/networks a gene is involved with, the more restriction would be placed on the number of variants of the gene. Genetic diversity is restricted by epigenetic complexity and vice versa. It is impossible to build complex epigenetic programs if the DNAs are constantly changing. To compensate for the loss in the range of genetic diversity, complex organisms use different epigenetic programming of the same gene set, in addition to mutation, to adapt to environments and to evolve new phenotypes. Fish and human share nearly identical gene sets and the evolution from fish to human is in a large part a process of epigenetic programming, analogous to writing distinct books with the same set of vocabulary.

Complex organisms and epigenetic programs

Epigenetic programs are not only inherited during mitotic cell division but are also transmitted through the germline to the next generation (Cropley et al., 2006; Hitchins et al., 2007). They control both expression levels of genes and the specific combination of co-expressed genes within a specific cell type. The epigenetic programs are here broadly defined, including both the primary epigenetic proteins as well as those secondary or tertiary proteins that could regulate the primary proteins. The number of human genes is only about 1.6 fold more than that of a fruit fly and about the same as the mouse or fish. However, the number of certain enzymes responsible for epigenetic gene organization, the PRDM subfamily of histone methyltransferases, increases dramatically during metazoan evolution: 0 in bacteria, yeasts, and plants; 2 in worms, 3 in insects; 7 in sea urchins, 15 in fishes, 16 in rodents, and 17 in

primates (Fumasoni et al., 2007; Huang, 2002). This faster pace of expansion of certain epigenetic enzymes, relative to the pace for the genome, in complex metazoan indicates a correlation between complex epigenetic programs and complex organisms.

Complex organisms are here defined as those that have complex epigenetic programs. Whether an organism is more complex than another organism can be roughly estimated based on a comparison of the number of genes involved in epigenetic programs. This is informative to differentiate unicellular organisms: yeasts have more epigenetic enzymes than bacteria and are therefore more complex; yeasts have several histone acetylases and SET domain histone methyltransferases while bacteria have none. Based on the number of the PRDM family of epigenetic enzymes, it is also easy to conclude that vertebrates are more complex in epigenetic programs than invertebrates or that primates are more complex than rodents or fishes.

When the numbers of epigenetic enzymes are similar for some multicellular organisms, then the number of tissue or cell types is a good measure of epigenetic complexity since each tissue or cell type is representative of a distinct epigenetic program or gene expression pattern. The more tissue types an organism has, the more the number of distinct epigenetic programs and hence the more complex the epigenetic program. The exact number of tissue types for any complex organism remains unknown, largely because there are many more neuronal cell types than we can presently recognize (Stevens, 1998). But this may not prevent one from drawing the conclusion that organisms that appeared early in evolution generally have less number of cell types than their descendant but distinctly different organisms that appeared later.

The number of neuronal cell types likely represents a major proportion of the total number of cell types in a complex animal. Also, epigenetic programs may control the complex interaction and organization of these neuronal cell types that manifest as intelligent brain functions. Thus, organisms with complex and intelligent brains are likely to contain more cell types or more complex interaction and organization of neuronal cell types. It is therefore easy to infer that the first primate has more cell types or complex organizations than the first mammal

which has more cell types or complex organizations than the first vertebrate. Also, animals that go through complex and prolonged developmental process contain more complex epigenetic programs since the development from a fertilized egg to an adult organism is largely an epigenetic process. The same tissue type often exhibits different expression patterns or epigenetic programs at different stages of development.

Organisms with the most complex and advanced brain (but not necessarily the largest in volume) are necessarily more complex in epigenetic programs or have more varieties of neuronal cell types and more complex interactions. Humans obviously have more distinct cell types and more complex neuronal interactions, thanks to our complex brain, than any other species that ever lived and are necessarily the most complex and diversified in epigenetic programs. Human brain shows dramatically more methylated DNAs than chimpanzees (Enard et al., 2004).

Epigenetic restriction of genetic diversity

Research on epigenetic programs is still at its infancy. Based on the limited knowledge of today, we can still envision several ways by which epigenetic programs may restrict genetic diversity. First, most genes are needed for the proper functioning of multiple fetal and adult tissues. A germline mutation in these genes needs to be compatible with multiple tissue types. Thus, the number of viable mutant variants is limited by the number of tissue types with which the gene is involved.

Second, some genes are only expressed in one tissue type, such as hemoglobin in red blood cells. These genes however still exhibit different expression patterns at different time points during development. The gene expression pattern of fetal red blood cells is different from adult red blood cells. So these genes still need to be compatible with several different developmental gene expression patterns. Furthermore, they need to be repressed in most cell types during development and during normal adult life. They need to be packaged into a

chromatin state that silences gene expression. Some mutant variants may interfere with such chromatin mediated repression and would be negatively selected.

Third, some genes are expressed in only one cell type but the function of the gene is needed for most cell types of an organism. The function of hemoglobin is needed for the oxygen supply of every cell type. Also, many house keeping genes such as actin are needed for most cell types. Such general function of a protein like hemoglobin and actin may be fine-tuned for the need of multiple tissues. A house keeping gene may also exhibit new functions or connections with new networks in complex organisms that are absent in simple organisms, such as the apoptosis function of cytochrome c. Also, for a complex organism to evolve a new cell type, it is necessary to keep the house keeping genes unchanged so that new cell types can evolve with the least amount of unnecessary disruption to existing cell types. It may not matter much as to which specific version of a house keeping gene is used but it is important to stick with one once it is selected by an organism.

Fourth, the coding region of every gene in complex organisms encodes not only amino acids but also epigenetic information such as the nucleosome code (Segal et al., 2006). A nucleosome code allows the nucleosome to locate in the right position in the genome. A silent mutation may nevertheless affect the nucleosome code and alters the chromatin packaging state of the gene, which may affect either gene repression or activation.

Fifth, complex organisms can eliminate reproductive cells carrying severe mutations (Fan et al., 2008). Also, embryos of complex organisms may die before birth if they did not develop properly due to mutations.

Sixth, epigenetic enzymes execute a senescence response to oncogenic mutations, thus nullifying the harmful effects of such mutations (Braig et al., 2005).

Finally, the non-coding and non-expressed regions of the genome are nevertheless packaged into chromatin and encode the nucleosome code and other information necessary for gene expression and organization, and are therefore not free from epigenetic restrictions. Many

epigenetic proteins interact with the genome in a sequence specific fashion such as the PRDM family that contains DNA-binding zinc-finger motifs (Huang, 2002). Even when an epigenetic enzyme has no intrinsic DNA binding property, it nevertheless interacts with a DNA binding transcription factor and therefore requires a specific DNA motif to function as either coactivators or corepressors (Rosenfeld et al., 2006).

The maximum genetic diversity (MGD) hypothesis

The inverse relationship between genetic diversity and epigenetic complexity is logically and self-evidently true on its own, just like the original intuition that triggered it. It in turn logically deduces what may be termed the maximum genetic diversity (MGD) hypothesis. The hypothesis has three themes. First, empirical facts of evolution show both macroevolution and microevolution (Figure 1). Macroevolution involves major advances in epigenetic complexity. The overall direction towards higher complexity however does not necessarily exclude occasionally going in the opposite direction. An organism is more complex if it has a higher degree of epigenetic complexity as indicated by its number of cell types or its number of epigenetic enzymes. Unlike macroevolution, microevolution is a gradual process of accumulating mutations due to either drift or selection as described by a watered down version of the molecular clock hypothesis or the 'Neutral Theory' and the Neo-Darwinian selection hypothesis. It may also involve some low degree of stochastic epigenetic reprogramming without a significant net change in epigenetic complexity.

Second, complex organisms are constructed more by epigenetic programs relative to simple organisms and are in turn inherently less tolerant of mutations. The maximum genetic diversity allowed for a complex organism is smaller than that allowed for a simple organism. The notion that genetic distance is roughly a function of time and mutation rates only applies to diverging organisms of similar complexity over short time scales prior to reaching the maximum cap. Most of the shared residues between two species are due to shared functions and

epigenetic complexity. A small fraction of the shared residues may be due to common adaptation to a common environmental selection that may vary from time to time (Figure 2). For distinctly different kinds of organisms, their genetic distance is independent of mutation rates and time but is determined by the maximum genetic diversity of the simpler organism. The gradual increase in epigenetic complexity with time during macroevolution of distinct organisms results in the linear correlation between maximum genetic distance and time of species divergence. Such a correlation holds only for macroevolution and is not related to mutation rates. It is fundamentally different from the correlation between genetic distance (prior to reaching maximum) and time of divergence during microevolution in short time scales. Actual mutation rates are usually fast enough for maximum genetic distance to be reachable in evolutionary time.

Finally, while both micro- and macro-evolution involve gradual accumulation of mutations and minor variations in epigenetic complexity, macroevolution from simple to complex organisms is associated with a punctuational increase in epigenetic complexity and in turn a punctuational loss in genetic diversity (Figure 2 and 3). From a common ancestor, the genetic distance between two splitting descendants may gradually increase with time until reaching a maximum level. This maximum genetic distance will stay roughly unchanged with time thereafter (Figure 3). Mutations still occur but only affect saturated sites or sites that suffer repeated hits. For microevolution, no major changes in epigenetic complexity will take place in either of the two splitting species. For macroevolution, one of the two splitting organisms will undergo a sudden increase in epigenetic complexity. This may take place soon after the two splitting organisms have reached their maximum genetic distance. The sudden increase in epigenetic complexity may be a response to the inadequacy of mutation alone in adapting to new environmental challenges. This punctuational jump in epigenetic complexity forces the genetic diversity of the new species to be lower than its sister species that remains largely unchanged in epigenetic complexity. This in turn causes the genetic distance between the new

species and its simpler sister species to be strictly determined by the maximum genetic diversity of the sister species.

The maximum genetic diversity hypothesis explains numerous facts

A large number of well established but puzzling observations can now be easily explained by the MGD hypothesis and a selected few are shown in the following to further illustrate the hypothesis. In addition, a few novel facts have been uncovered that would represent confirmations of the predictions of the hypothesis. None of these observations are needed to invoke the hypothesis in the first place, since the hypothesis was deduced from intuition. Therefore, all of them can be considered as independent lines of evidence in support of the hypothesis.

1. *Relationship between genetic diversity and time of origin.* It is well established that genetic diversity within a biological kind of old lineage is greater than that within a biological kind of young lineage (Figure 4A). The genetic diversity of bacteria is greater than eukaryotes (Ciccarelli et al., 2006). The fact that simple organisms with inherently high-level tolerance of genetic diversity evolved earlier in history generates the apparent correlation between the time of origin and genetic diversity (Figure 4A). But an equally valid relationship is between the time of origin and the epigenetic complexity of the organism (Figure 4B). If epigenetic complexity sets up a maximum cap on genetic diversity and if simple organisms appeared earlier than complex organisms, then the apparent correlation between time of origin and genetic diversity can be explained as an epiphenomenon of epigenetic complexity that is largely independent of mutation rates, generation times, and population size.

2. *The MGD hypothesis predicts the genetic equidistance result.* The maximum diversity allowed for an organism X is the same as the maximum genetic distance between X and all descendants of X. The equidistance from X shared by all different descendants of X is strictly determined by the epigenetic constraints imposed on X but is not linked to the more severe

epigenetic constraints imposed on the descendants of X. This notion is illustrated by a hypothetical case as shown in Table 1. If fish is allowed a maximum diversity of 60% difference in a hypothetical protein sequence of 10 amino acids as shown in Table 1, then fish 1 would differ from a maximum diverged fish 2 in 6 of the 10 amino acid positions. All evolutionary descendants of fish, whether a different subspecies of fish or an amphibian or a human, would all have the same maximum genetic distance with an extant fish that is equivalent to the maximum diversity of 60% allowed for fish (Table 1).

If amphibian is allowed a maximum diversity of 40% difference, which is lower than fish because amphibian is more complex, all evolutionary descendants of amphibian, whether a different subspecies of amphibian or a mouse or a human, would all have the same genetic distance from an extant amphibian that is equivalent to the maximum diversity of 40% allowed for amphibian (Table 1). But the epigenetic constraint on amphibian has no effect on the distance between amphibian and fish, which is strictly a result of the epigenetic constraint on fish. The epigenetic constraint on amphibian only affects or determines the equidistance to amphibian shared by all different descendants of amphibian. All fish descendants that do not look like fish can be viewed as maximum diverged fishes and should show approximately the same maximum distance with an extant fish that is the same as the maximum diversity allowed for fishes.

It is well known that sequence regions conserved in simple organisms are often also conserved in complex organisms. Sequence regions not conserved in complex organisms are also often not conserved in simple organisms. This explains the fact as illustrated in Table 1 that a comparison of fish (with a hypothetical maximum diversity of 60%) and human (with a hypothetical maximum diversity of 10%) should result in a dissimilarity of 60% equaling the maximum diversity of fish, rather than 70%.

This notion that the maximum genetic diversity of a simple kind of organism determines and is about the same as the maximum distance between the simple organism and the later

appearing complex organisms can be illustrated by the example of cytochrome c. The maximum diversity in this protein sequence is about 70% difference within bacteria, for example, between *Bordetella parapertussis* and *Paracoccus Versutus*. The maximum distance between bacteria and mammals is about 65% difference, such as between *Bordetella parapertussis* and *Pan troglodytes* (chimpanzees). Within fungi, the maximum diversity is about 40% difference, for example, between *Aspergillus oryzae* and *Yarrowia lipolytica.* The maximum distance between fungi and mammals is about 43% difference, such as between *Aspergillus oryzae* and *Pan troglodytes*. Within arthropods, the maximum diversity is about 24% difference, for example, between *Drosophila melanogaster* and *Tigriopus californicus.* The maximum distance between arthropods and mammals is about 25% difference, such as between *Drosophila melanogaster* and *Pan troglodytes*.

3. *The relationship between time and genetic distance in microevolution is different from that in macroevolution.* Most genes (about 90%) have been found to behave consistently as good clocks in macroevolution, and show the same pattern as originally found for cytochrome c (Fitch and Margoliash, 1967; Margoliash, 1963): human is more related to primates, less to rodents, still less to birds, still less to frogs, and still less to fish (e.g., see Table 2). However, despite their consistent pattern in macroevolution, many genes give erratic or contradictory results when the timing of split in microevolution is measured. For example, pufferfish (*Takifugu rubripes*) and zebrafish (*Danio rerio*) are believed to have diverged not more than 140-200 MyBP (million years before present) based on the first fossil evidence of teleostei in the early Cretaceous period (Powers, 1991). If the situation between the two fishes is similar to what one originally found for cytochrome c in macroevolution, one would expect 90% of all genes to show more identity between the fishes than between human and bird since the time of divergence for human and bird is much earlier (310 MyBP).

In a survey of 40 randomly picked genes, I found 36 (90%) that show the expected macroevolution pattern where human is more related to bird, less to frog, and still less to fish. In

contrast, only 19 (48%) show more identity between the two fishes than between human and bird. Depending on which gene is used as clock, the time of divergence between the two fishes would vary from 91 to 420 million years (Table 2). In fact, I employed the molecular clock method to derive an average time of divergence using these 40 genes by calibrating against the fossil divergence time between human and bird (310 MyBP). However, I obtained an obviously incorrect time (417+/-172 MyBP) that is more than two fold greater than the actual time as indicated by the fossil record. As a positive control to show that my method is similar to those of others, I derived a mean time of divergence between human and amphibians and found it to be similar to that obtained by others (Kumar and Hedges, 1998).

Apparently, some of the subspecies split or microevolution is not equivalent to the changes in macroevolution, but the Neo-Darwinian hypothesis treats them the same. In contrast, the MGD hypothesis considers them to be very different in evolutionary dynamics. So, clocks derived from macroevolution should not be expected to work also for microevolution. Genetic distance between two distinct species of macroevolution always reflects the maximum genetic distance. However, genetic distance between two similar species that have diverged more recently would gradually increase as a function of time before they reach the maximum (Figure 2). Different genes would diverge according to different mutation rates. If the time is not enough for all genes to reach the maximum diversity level, some genes may reach a diversity level closer to the maximum than some other genes. The genes in fish are allowed a maximum diversity level greater than genes in birds and humans. So if some genes reached a diversity level closer to the maximum, they would put the time of split between the two fishes earlier than that between birds and humans. But some other genes may only reach a certain diversity level much lower than the maximum because of slower rate of mutations and insufficient time. These genes would put the time of split between the fishes later than that between birds and humans.

4. *Radiation of mammals and the Cambrian explosion.* The two main areas of disagreement between molecular clocks and the animal fossil record concern the radiation of

mammal orders around the Cretaceous-Tertiary boundary (65 MyBP) and animal phyla at the Cambrian explosion 520 MyBP (Hedges, 2002). In each case, molecular clocks show much deeper divergence. The MGD hypothesis suggests that the rates of change in genetic distance for macroevolution are determined by epigenetic complexity. They tend to be slower than the actual mutation rates. If these slower rates are used to date microevolution in the horizontal direction, we would expect to see a deeper time of divergence than the actual time, as we have already seen above for the two fishes. Some mammalian radiation events involve varieties within a kind, such as the split between mouse and rat, and may represent microevolution in the horizontal direction. They may accumulate mutations faster than the slower rate observed for macroevolution. This would cause the time of split between mouse and rat to be older than the actual time: 23-41 million years from the molecular clock estimate versus 10-12 million years from the fossil record (Hedges, 2002). Some mammalian radiation events may involve a major change in kinds and represent macroevolution, such as the split between primates and rodents. They may involve a higher than average rate of change in epigenetic complexity and in turn in maximum genetic distance. In this case, if the average rate of change in epigenetic complexity is applied, it would give a deeper time of divergence than the actual time.

The rate of change in epigenetic programs between phyla may be much greater than that between different species within one phyla. For example, vertebrates have a much greater number of PRDM epigenetic enzymes than arthropods (Huang, 2002). But the number of PRDM genes among different species of vertebrates is similar. The rate of change in epigenetic programs in macroevolution within the vertebrate phyla may be slower than that between phyla or between arthropods and vertebrates. So when the slow rate estimated from speciation events within one phyla, that of vertebrate, is used to calibrate the time of phyla divergence between arthropods and vertebrates, the time would be estimated to be deeper than the actual time (1000 MyBP versus 520 MyBP) (Hedges, 2002).

5. *Simpler organisms show higher genetic diversity than complex organisms after evolving for the same amount of time.* The MGD hypothesis predicts that simple organisms should show higher genetic diversity than complex organisms after the same amount of time of evolution. Indeed, flowering plants have much greater genetic diversity than mammals even though they have both coevolved for similar amount of time (Huang, 2008a). Flowering plants are less complex in epigenetic programs and have zero PRDM family of epigenetic enzymes while mammals have 16 to 17. It is also obvious that flowering plants have less number of cell types than mammals. The genetic diversity of flowering plants after less than 125 million years of evolution is about equivalent to that reached by vertebrates after 450 million years of evolution. So the hypothesis explains equally well both data for and against the correlation between genetic diversity and time of divergence.

6. *Direct evidence of maximum genetic diversity.* The MGD hypothesis predicts that the genetic distance between some ancient species of similar kind or epigenetic complexity may have reached a maximum cap long before present. I tested this prediction for the fungi kingdom. The baker's yeast *Saccharomyces cerevisiae* belongs to the *Ascomycota* phylum, the *Saccharomycotina* subphylum, the *Saccharomycetes* class, the *Saccharomycetales* order, the *Saccharomycetaceae* family, and the *Saccharomyces* genus. A large number of observations have established the well-known top-down direction of evolution where the major pulse of divergence of phyla occurs before subphyla or classes, classes before that of order, orders before that of families, and families before that of genera. If many fungi may share similar epigenetic complexity, the MGD hypothesis predicts that, if time is long enough for genetic distance to reach the cap, the maximum genetic distance between two fungi genera of the same family should be similar to that between two fungi families, or orders, or phyla. In contrast, the molecular clock hypothesis predicts that the genetic distance between two fungi genera of the same family should be smaller than that between families, still smaller than that between orders, still smaller than that between classes or subphyla, and still smaller than that between phyla.

I randomly picked three proteins for analysis, Pin1, CytC (cytochrome c), and CMD (Calmodulin). As shown in Table 3, the protein sequence identity in Pin1 between two distant genera (*S. cerevisiae/Saccharomyces* and *D. hansenii/Debaryomyces)* of the same family is 44%, which is about the same as that between two families of the same order (39% between *S. cerevisiae/Saccharomycetaceae and Y. lipolytica/Dipodascaceae*), or about the same as that between two subphyla of the same phylum (42% between *S. cerevisiae/ Saccharomycotina* and *G. zeae PH-1/Pezizomycotina),* or about the same as that between two phyla of the same kingdom *(41% between S. cerevisiae/Ascomycota* and *C. Neo-formans/Basidiomycota*). For the protein CytC, the identity between two distant genera (78% identity) seems to be larger than that between two distant families (73% identity) which is still larger than that between two distant subphyla (67% identity) which is still larger than that between two distant phyla (60% identity) (Table 3). This pattern is consistent with the top down direction of evolution and suggests that the time may not yet be long enough for the genetic distance in CytC among the presently sequenced fungi taxa to reach the maximum cap, consistent with the known slow mutation rate of the CytC protein. For the protein CMD, genetic distance between taxa above the family level appears to have reached a maximum at 56-60% identity. These data show that there is a maximum cap on genetic distance at some faster mutating loci like Pin1 and CMD between two species of similar kind in the fungi kingdom. The cap may be gradually reached by gradual accumulation of mutations within a certain amount of time.

I also found direct evidence of maximum cap in fishes. Zebrafish (*D. rerio*) and pufferfish (*T. nigroviridis*) diverged not more than 140-200 MyBP ago as mentioned above. If they diverged by the gradual model and if time is long enough for at least some genes to reach the maximum genetic distance, the MGD hypothesis predicts that some genes would show a genetic distance between the two fishes that is similar to the maximum genetic diversity allowed for fishes. The maximum genetic diversity of fishes is of course roughly the same as the genetic distance between fishes and a distinct fish descendant such as a mammal. I examined a large

number of chromatin modifying enzymes and found 13 out of 32 with a distance between the fishes to be the same or slightly greater than the distance between a fish and a mammal (Table 4). The SET domain family of histone lysine methyltransferases (KMTs) is specifically more enriched with genes that evolved fast with 6 out of 9 genes analyzed reaching maximum cap in the fishes. This feature of the KMT family is significantly different from a slowly evolving family such as ribosomal proteins with only 2 of 12 proteins analyzed reaching maximum cap in the fishes ($P < 0.05$, Fisher's exact test, two tailed). Not a single gene was found to have significantly greater distance between the two fishes than between fish and mammal, indicating clearly the existence of a cap on genetic distance.

7. *Actual mutation rate in real time is faster than that calculated from phylogenetic analysis*. It is well known that mutation rate from pedigree analysis on genealogical timescales is often an order of magnitude or more greater than mutation rate from phylogenetic analysis over geological time (Ho and Larson, 2006). Thus, phylogenetic diversity or distance over geological time is uncoupled from actual mutation rate observed on genealogical timescales. It suggests that actual mutation rates are often fast enough for most organisms to reach a maximum cap in genetic distance over geological timescale. Indeed, if actual mutation rates are slower than those from phylogenetic analysis, it would falsify the MGD hypothesis.

The phylogenetic diversity or distance reflects the maximum diversity allowed for an organism. Some of the variants at a particular time period accumulated as a result of random mutations may not persist long over geological time and may have to be replaced by another set of variants at a later time period (Figure 2). Maximum genetic distance between two species would stay constant over time while the same genetic distance may be maintained by different sets of variants at different times (Figure 2). A set of variants best suited for life at one time may not be the best at a different time and would have to be replaced.

8. *Stasis and punctuation in the fossil record.* The MGD hypothesis suggests that morphological phenotypes for complex organisms are better correlated with epigenotypes.

Advances in epigenotypes in macroevolution occur largely via punctuation (Figure 2). Such punctuation events are followed by stasis in epigenotypes in microevolution. Thus the hypothesis predicts both stasis and punctuation at the level of epigenotypes and in turn at morphological levels. Consistently, the fossil record shows both stasis and punctuation at morphological levels (Gould and Eldredge, 1993).

9. *Cancer as a disease of both genetics and epigenetics.* The MGD hypothesis predicts that high epigenetic complexity has a way of limiting the incidence of mutations. A relaxation in epigenetic control may be expected to allow more mutations to occur. Indeed, human cancer provides a good illustration of this prediction. Mutations are common in cancer. Epigenetic programs are often deregulated in cancer and methylation deficiency is a hallmark of cancer (Feinberg and Tycko, 2004; Huang, 2002). Loss of epigenetic control as indicated by loss of DNA methylation occurs during aging and precedes mutations in cancer (Suzuki et al., 2006). A rate-limiting step in carcinogenesis by major environmental factors such as nutrient-imbalanced diet is the deregulation of an epigenetic enzyme RIZ1/PRDM2 (Zhou et al., 2008). In addition, the hypothesis predicts that high genetic diversity or too many mutations would interfere with epigenetic programming. Indeed, too many mutations, either germ line or somatic, are well known to cause cancer, which is essentially a disease where the normal epigenetic programs have been replaced by a cancer specific program. Thus, the hypothesis unifies cancer genetics and epigenetics and explains why cancer appears to be a disease of both genetic mutations and epigenetic anomalies.

10. *Copy number variations of the genome.* Advances in epigenetic complexity may involve changes that affect large regions of the genome, such as amplification or deletion of long stretches of DNA. Thus, such copy number changes may be expected to be a common behavior of the genome just like point mutations are. Indeed, copy number variations are observed to be common in the human genome (Redon et al., 2006). Within a specific level of epigenetic complexity, a certain range of neutral and random copy number changes are allowed

that may affect slightly epigenetic programs, just like a certain range of random point mutations are allowed. Relaxation of epigenetic programs is expected to allow more abnormal copy number changes to occur. Indeed, cancer is commonly caused by loss of epigenetic control and often exhibits aneuploidy and amplifications or deletions of long stretch of DNA.

11. *Genetic diseases.* The prevalence of genetic or familial diseases in humans indicates plainly that a large portion of genetic diversity, i.e. those represented by those disease mutations, cannot become a part of the normal range of genetic diversity among humans. Most genetic diseases affect only a tiny population of humans. Just imagine how much more diversified the human race would be if all those rare disease mutations would become fixed in the whole population. If mutations in the retinoblastoma gene, a cell cycle regulator important for many different cell types, do not cause cancer in the retina of children, the diversity in the retinoblastoma gene locus would be greatly expanded. The fact of rare disease mutations in humans is sufficient to prove the hypothesis that there is an upper limit to the amount of genetic diversity in an organism. The fact that those rare disease mutations are mostly tissue specific is consistent with the notion that the upper limit is set up by the complexity of epigenetic programs. If humans lack the retina cell type or the retina specific epigenetic program, most of the mutations in the retinoblastoma gene would have been tolerated as normal variations and the genetic diversity of humans would have been in turn expanded. Also, numerous disease alleles in humans correspond to normal alleles in rhesus macaques (Gibbs et al., 2007). Thus, many alleles or mutant variants that can be tolerated in a less complex organism in fact cause diseases in humans.

12. *Anomalies of the genetic equidistance result*. A small number of genes show anomalies and are routinely excluded from phylogenetic analysis based on the molecular clock hypothesis. An example is the mitochondrial protein ND6. My analysis showed that all vertebrates ND6 proteins are equidistant to the outgroup sea urchin but fishes ND6 proteins are closer to frogs than to mammals. The molecular clock hypothesis has no explanation for such a

gene that shows both equidistance as well as non-equidistance. However, the MGD hypothesis easily explains it. Some of the shared sequences are due to common environmental selections (Figure 2). Fishes may have more in common with frogs than with mammals in their adaptation strategies for the ND6 protein.

13. *Inverse correlation between genome size and genetic diversity*. Large size genomes (measured here as number of genes) require more complex epigenetic regulation than small genomes and are expected to show less genetic diversity. Indeed, there is a strong inverse correlation between genome size and genetic diversity (Ciccarelli et al., 2006). Genetic diversity is more responsive to changes in genome size in bacteria than in eukaryotes, indicating that genetic diversity is restricted more by epigenetic complexity than by genome size in eukaryotes.

In microbes, there is an inverse relationship between genome size and mutation rate per base pair per replication (Drake et al., 1998). In four metazoans analyzed, the mutation rate per base pair per replication is lowest for humans, higher for mice, and still higher for drosophila or worm. These data are expected from the MGD hypothesis.

14. *No bacterium lineage could be identified as the closest relative of eukaryotes*. Based on the overall trend in evolution from simple to complex organisms and the earliest fossil evidence of life on Earth, it is almost certain that bacteria were the ancestors of the eukaryotes. However, the MGD hypothesis predicts that no single bacterium lineage could be identified among bacteria as the closest relative of eukaryotes. Such a lineage, if indeed exists, would have long reached maximum diversity and would show equidistance to eukaryotes as other bacteria. In contrast, if there is no maximum cap on diversity or if time is not long enough yet and if the Neo-Darwinian hypothesis is true, one should be able to identify the bacterium lineage that is closer to eukaryotes than most other bacteria. But extensive studies show that no such bacterium lineage can be identified. Recent data show that the identification of archaea as closer to eukaryotes is only true for some class of genes such as those involved in translation

(Pennisi, 1998). For many other genes, archaea are in fact more distant to eukaryotes than eubacteria. The overall pattern of genetic similarity suggests that common selection and coincidence may account for most of the sequence identities between eukaryotes and bacteria.

The closer relationship between a bacterium species and eukaryotes in some genes but not others has been commonly interpreted to mean horizontal gene transfer, even though there is little independent evidence for it. It is more likely however that the closer relationship are fortuitous due simply to the fact that bacteria have much greater genetic diversity and some gene variants of bacteria would by chance resemble an eukaryotic version. If one compares a gene from a mammalian species against orthologous genes of all species of bacteria in the Genbank, one would find that the degree of similarity would vary to a great extant (e.g., for GLUD1, the identity between human and all bacteria ranges from 30% to 50%). In contrast, if one compares a gene from an individual bacterium species against all vertebrate species in the Genbank, one would find that the degree of similarity falls within a very narrow range (e.g., for GLUD1, the identity between the bacterium *Pedobacter sp. BAL39* and all vertebrates ranges from 47% to 53%). Vertebrates have lower genetic diversity and there is much less probability for a variant of vertebrates to be much more closely related by chance than other variants to an individual variant of bacteria.

There are data against the idea of horizontal gene transfer. If a gene was transferred from a prokaryotic lineage into the vertebrate lineage, this likely occurred within the past 400 to 500 million years, after most of the major prokaryotic phyla were established. Therefore, any transferred gene should be more closely related to its donor lineage than to any other prokaryotic lineage, which would be detectable in phylogenetic trees. However, it was found that most of the genes shared between vertebrates and bacteria did not show patterns consistent with bacterial to vertebrate gene transfer (Salzberg et al., 2001).

15. *Ubiquitously expressed genes have lower genetic diversity than tissue specific genes.* The MGD hypothesis predicts that ubiquitously expressed genes have lower diversity

than tissue specific genes due to selection against mutations that cannot fit with multiple cell types. Indeed, an analysis of 2400 genes between human and rodent found that ubiquitously expressed proteins have average genetic distances between human and rodent that were threefold lower than those of tissue-specific genes (Duret and Mouchiroud, 2000).

16. *Stability of epigenetic programs.* It is well known that artificial selection or breeding of animals can only generate varieties of the same type but never of a different type. This fact plainly indicates that genetic variation within an organism is not without a limit. The epigenetic program that allows a genome to manifest a dog phenotype also prevents the same genome from randomly drifting into something that is not allowed by the epigenetic program. Indeed, random drifting is far more likely to give rise to cancer rather than a novel functional organ. If genotypes can be rather unstable or easily influenced by random mutations, the epigenotypes are relatively much more stable. Indeed, when cultured for up to *ten years*, hundreds of cell divisions later, *Drosophila* wing disc cells can still give rise to adult wing structures (Hadorn, 1967). The stability of epigenotypes is also indicated by the stasis and extinction phenomena in the fossil record. If environment becomes unsuitable for survival, a species would more often than not go extinction rather than change itself in its basic epigenetic programs. In today's world, all we observe is extinction of species rather than drastic transformation of species.

A specific epigenetic program allows a certain degree of variation in genotypes and in turn a certain range of adaptive capability in response to environmental changes. When the environmental changes exceed the adaptive capability allowed within a specific epigenetic program, the organism would simply go extinct rather than change. Change is not without a limit. Change is not the only feature of evolution. Equally important and obvious as change is the opposite of change. If constant random change within a limited range in genotypes is a hallmark of evolution, then long period of stasis and stability in epigenotypes followed by short period of punctuational advance in epigenotypes is an equally important hallmark of evolution. Indeed, the genetic code is the optimal code for error minimization or for minimizing the effects

of random changes; it is the most stable of all possible codes and is optimal for stability rather than for random changes (Freeland et al., 2000).

17. *Low genetic diversity of chromosome X.* Comparisons of genomes have shown a lower rate of sequence divergence on chromosome X than the autosomes for many species (Patterson et al., 2006). Chromosome X undergoes X inactivation in females, which is an epigenetic event. Thus, genes located on X encounter more epigenetic restrictions than genes on autosomes and are therefore expected from the hypothesis to be less tolerant of mutations. This explanation is far more reasonable than the suggestion of interbreeding between humans and chimpanzees (Patterson et al., 2006).

18. *Human has the lowest genetic diversity.* The genetic diversity within chimpanzees is two to three times greater than within humans, even though both species are thought to have evolved for the same amount of time since their most recent common ancestor (Becquet et al., 2007). The striking fact that human shows the lowest DNA diversity among all species has commonly been explained by the bottleneck hypothesis: most human populations are thought to go extinct at one time in history except one small population that survived to produce the six billion people living today. But there is no independent evidence of such near extinction event and there is little hope of ever uncovering such evidence. There are also lines of evidence against the bottleneck hypothesis (Li and Sadler, 1991; Xiong et al., 1991). In contrast, the homogeneity of human DNA can be easily explained by the MGD hypothesis. The organism with the most complex and diverse epigenetic programs or the most number of cell types is necessarily supposed to have the lowest diversity in DNA.

Neanderthals appeared earlier than modern humans and have slightly larger brains. It is unclear whether Neanderthals may be less intelligent or have less complex brain than modern humans, which may explain their mysterious extinction. The MGD hypothesis predicts that Neanderthals are less complex in epigenetic programs and have a less complex brain than modern humans because Neanderthals appear to exhibit more DNA diversity (Krings et al.,

2000; Orlando et al., 2006). This prediction is obviously consistent with the fact that it is modern humans rather than the Neanderthals that dominate the Earth today. It is also consistent with the evolution trend that less complex organisms appeared earlier in history.

19. *Mammals are more distant to snakes than to birds.* Mammals and reptiles (including birds) were separated ~310 MyBP. Thus, all the reptiles (including birds) should be equidistant to a mammal if the constant mutation rate idea is true. But the MGD hypothesis predicts that simpler reptiles such as snakes, which lost limbs, should have higher genetic diversity and hence be more distant to a mammal than complex reptiles such as birds. A random sampling of five proteins indeed shows that snakes are more distant to humans than birds are (Hemoglobin alpha chain, albumin, cytochrome b, PDGF, and ND6).

20. *More recently evolved complex brachiopods are closer to mammals.* The inarticulate brachiopod genus Lingula (order Linguilida) is the oldest, relatively evolutionarily unchanged animal known. The oldest Lingula fossils are found in Lower Cambrian rocks dating to roughly 550 MyBP. Terebratulids are modern articulate brachiopods and appeared later in evolution around ~430 MyBP. The molecular clock hypothesis predicts that mammals should be equidistant to Lingula and terebratulids. But the MGD hypothesis predicts that mammals should be closer to terebratulids given that they evolved later and should have lower genetic diversity. Indeed, a random sampling of several proteins showed that mammals are closer to terebratulids than to Lingula (Cox1, Cox2, Cox3, ND1, and COB). Also, terebratulids are closer to mammals than to a fellow brachiopod Lingula.

In contrast to the brachiopods, complex plants (flowering plants) that appeared later in evolution and simpler plants (mosses) that appeared earlier are about equidistant to mammals in several randomly analyzed genes (EF1a, Adh1a, EIF2b, Pin1, PP1, RPC1, and Cox1). The identity between flowering plants and mosses are much greater than between mammals and mosses, in contrast to brachopods where the distance between mammals and Lingula is similar

to that between terebratulids and Lingula. Thus, plants have evolved plant-specific conserved domains since separating from mammals but before divergence of mosses and flowering plants.

21. *The genetic diversity of tuatara*. The tuatara of New Zealand is a living fossil reptile and has very slow metabolic and growth rates, long generation times and slow rates of reproduction. Contrary to expectations from Neo-Darwinian theory, tuatara has high 'mutation rates', significantly higher than those of mammals measured in real time by the same method of using mitochondrial D-loop DNA sequences from fossils of ~10,000 years old (Hay et al., 2008). However, this result is to be expected from the MGD hypothesis since reptiles should have higher maximum genetic diversity than mammals. If ~10,000 years is sufficient for reptiles and mammals to reach maximum cap in genetic diversity in the D-loop region, then the reptiles would show higher genetic diversity, resulting in the appearance of a higher 'mutation rate'. But in reality, tuatara can have slower mutation rate but still show higher genetic diversity than a mammal if time is long enough for tuatara to reach the maximum cap or to be close to the cap.

22. *Evidence from fossil DNA and protein sequences.* Finally, the MGD hypothesis explains the recent results that ancient fossil specimens are more distant to an outgroup than extant sister species are and that ancient fossil specimens have greater genetic distance than extant sister species (Huang, 2008b). The Neo-Darwinian gradual mutation hypothesis predicts that ancient specimens of extinct species cannot be more distant to an outgroup than extant sister species are (Figure 5A). Also, two distinct ancient specimens from different era cannot be more distant than their extant sister species are. But the MGD hypothesis predicts the exact opposite (Figure 5B). The recent analysis of fossil DNA and protein sequences fully conforms to the predictions of the MGD hypothesis.

Implications for molecular phylogeny

Molecular phylogeny analysis aims to classify the time of divergence between morphologically similar species that either do not have fossil records or cannot be clearly

distinguished by fossil records. A mutation rate is usually calibrated using fossil records of vertebrate macroevolution. The most commonly used calibration date is the divergence time of 310 million years between birds and mammals. However, as discussed here, the 'mutation rate' deduced from macroevolution is not the real mutation rate as it is known in real time measurements. It therefore cannot be used to time microevolution of species that have diverged only recently and have not yet reached a maximum cap in genetic diversity. Such microevolution reflects the real mutation rate and should be timed using a mutation rate that is measured by real time analysis such as pedigree analysis. For example, to date the divergence of pufferfish and zebrafish, one should only use genes that have not reached a maximum diversity. Thus, cytochrome c may be used but most KMTs should not be used. However, the mutation rate of cytochrome c should be deduced not from divergence of birds and mammals but from pedigree analysis of living pufferfish and zebrafish.

Based on mutation rates derived from pedigree analysis of the mitochondrial D-loop region, the human race is estimated to be only ~ 6500 years old (Parsons et al., 1997). However, what the result means is uncertain since the maximum genetic diversity of the mitochondrial D-loop region is unknown for humans. If the genetic diversity has not yet reached a cap within 6500 years, then we may conclude that the human race is indeed 6500 years old. On the other hand, if the cap has been reached in 6500 years, then we will not be able to discern the real age of human race using the mitochondrial D-loop DNA. The age could be much older while the diversity of the D-loop DNA would no longer increase with time after reaching the cap. Given that the oldest fossil of modern humans is much older than 6500 years (about 30,000 to 65,000 years old), it is likely that the maximum diversity of the mitochondrial D-loop can be reached in ~ 6500 years. Thus, if the fossil record is true, the maximum distance in the D-loop that we observe today within the human race would in fact represent the maximum allowable for the human organism.

Conclusions:

The inverse relationship between genetic diversity and epigenetic complexity is self-evident. It does not need independent validation of empirical facts, just like the intuition that a complex system is more selective in building materials than a simple system. Nonetheless, it or its necessary logical deduction, the maximum genetic diversity hypothesis, has found support in numerous facts and has yet to meet a factual contradiction, as would be expected for any self-evident logical truth. It explains more facts than does the molecular clock hypothesis and represents the first testable or scientific explanation of the genetic equidistance result. Many data that were simply ignored before can now be understood. Most of the existing literature of molecular evolution would need to be re-interpreted in light of the new hypothesis. The molecular clock hypothesis and the Neutral Theory cannot account for macroevolution and are only relevant to some microevolution events over short timescales. The inference of divergence time based on sequence identity is still practically useful in some cases. But a distinction must be made between divergence that has reached a maximum and divergence that has not.

Acknowledgments:

This work was supported by NIH (RO1 CA 105347). I thank Drs. Phil Skell and Klara Briknarova for critical reading of the manuscript.

References:

Ayala, F. J. (1999). Molecular clock mirages. BioEssays *21*, 71-75.

Becquet, C., Patterson, N., Stone, A. C., Przeworski, M., and Reich, D. (2007). Genetic structure of chimpanzee populations. PLoS Genet *3*, e66.

Braig, M., Lee, S., Loddenkemper, C., et al. (2005). Oncogene-induced senescence as an initial barrier in lymphoma development. Nature *436*, 660-665.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Cropley, J. E., Suter, C. M., Beckman, K. B., and Martin, D. I. (2006). Germ-line epigenetic modification of the murine A vy allele by nutritional supplementation. Proc Natl Acad Sci *103*, 17308-17312.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. Genetics *148*, 1667-1686.

Duret, L., and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol *17*, 68-74.

Enard, W., Fassbender, A., Model, F., Adorjan, P., Paabo, S., and Olek, A. (2004). Differences in DNA methylation patterns between humans and chimpanzees. Curr Biol *14*, R148-149.

Fan, W., Waymire, K. G., Narula, N., Li, P., Rocher, C., Coskun, P. E., Vannan, M. A., Narula, J., Macgregor, G. R., and Wallace, D. C. (2008). A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. Science *319*, 958-962.

Feinberg, A. P., and Tycko, B. (2004). The history of cancer epigenetics. Nat Rev Cancer *4*, 143-153.

Fitch, W. M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Freeland, S. J., Knight, R. D., Landweber, L. F., and Hurst, L. D. (2000). Early fixation of an optimal genetic code. Mol Biol Evol *17*, 511-518.

Fumasoni, I., Meani, N., Rambaldi, D., Scafetta, G., Alcalay, M., and Ciccarelli, F. D. (2007). Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. BMC Evol Biol *7*, 187.

Gibbs, R. A., Rogers, J., Katze, M. G., Bumgarner, R., Weinstock, G. M., Mardis, E. R., Remington, K. A., Strausberg, R. L., Venter, J. C., Wilson, R. K.*, et al.* (2007). Evolutionary and biomedical insights from the rhesus macaque genome. Science *316*, 222-234.

Gould, S. J., and Eldredge, N. (1993). Punctuated equilibrium comes of age. Nature *366*, 223-227.

Hadorn, E. (1967). Dynamics of determination. Symp Dev Biol *25*, 83.

Hay, J. M., Subramanian, S., Millar, C. D., Mohandesan, E., and Lambert, D. M. (2008). Rapid molecular evolution in a living fossil. Trends Genet *24*, 106-109.

Hedges, S. B. (2002). The origin and evolution of model organisms. Nat Rev Genet *3*, 838-849.

Hitchins, M. P., Wong, J. J., Suthers, G., Suter, C. M., Martin, D. I., Hawkins, N. J., and Ward, R. L. (2007). Inheritance of a cancer-associated MLH1 germ-line epimutation. N Engl J Med *356*, 697-705.

Ho, S. Y. W., and Larson, G. (2006). Molecular clocks: when times are a-changin'. Trends Genet *22*, 79-83.

Huang, S. (2002). Histone methyltransferases, diet nutrients, and tumor suppressors. Nat Rev Cancer *2*, 469-476.

Huang, S. (2008a). The genetic equidistance result of molecular evolution is independent of mutation rates. Submitted, Preprint available at Nature Precedings <http://hdl.handle.net/10101/npre.2008.1733.1>

Huang, S. (2008b). Ancient fossil specimens of extinct species are genetically more distant to an outgroup than extant sister species are. In press, Riv. Biol. Preprint available at Nature Precedings <http://hdl.handle.net/10101/npre.2008.1676.1>

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y.*, et al.* (2007). The medaka draft genome and insights into vertebrate genome evolution. Nature *447*, 714-719.

Krings, M., Capelli, C., Tschentscher, F., Geisert, H., Meyer, S., von Haeseler, A., Grossschmidt, K., Possnert, G., Paunovic, M., and Paabo, S. (2000). A view of Neandertal genetic diversity. Nat Genet *26*, 144-146.

Kumar, S., and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution. Nature *392*, 917-920.

Li, W. H., and Sadler, L. A. (1991). Low nucleotide diversity in man. Genetics *129*, 513-523.

Margoliash, E. (1963). Primary structure and evolution of cytochrome c. Proc Natl Acad Sci *50*, 672-679.

Orlando, L., Darlu, P., Toussaint, M., Bonjean, D., Otte, M., and Hanni, C. (2006). Revisiting Neandertal diversity with a 100,000 year old mtDNA sequence. Curr Biol *16*, R400-402.

Parsons, T. J., Muniec, D. S., Sullivan, K., Woodyatt, N., Alliston-Greiner, R., Wilson, M. R., Berry, D. L., Holland, K. A., Weedn, V. W., Gill, P., and Holland, M. M. (1997). A high

observed substitution rate in the human mitochondrial DNA control region. Nat Genet *15*, 363-368.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S., and Reich, D. (2006). Genetic evidence for complex speciation of humans and chimpanzees. Nature *441*, 1103-1108.

Pennisi, E. (1998). Genome data shake tree of life. Science *280*, 672-674.

Powers, D. A. (1991). Evolutionary genetics of fish. Advances in Genetics *29*, 119-228.

Pulquerio, M. J., and Nichols, R. A. (2007). Dates from the molecular clock: how wrong can we be? Trends Ecol Evol *22*, 180-184.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W.*, et al.* (2006). Global variation in copy number in the human genome. Nature *444*, 444-454.

Rosenfeld, M. G., Lunyak, V. V., and Glass, C. K. (2006). Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. Genes Dev *20*, 1405-1428.

Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. (2001). Microbial genes in the human genome: lateral transfer or gene loss? Science *292*, 1903-1906.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J. P., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772-778.

Stevens, C. F. (1998). Neuronal diversity: too many cell types for comfort? Curr Biol *8*, R708-710.

Suzuki, K., Suzuki, I., Leodolter, A., Alonso, S., Horiuchi, S., Yamashita, K., and Perucho, M. (2006). Global DNA demethylation in gastrointestinal cancer is age dependent and precedes genomic damage. Cancer Cell *9*, 199-207.

Xiong, W. J., Li, W. H., Posner, I., Yamamura, T., Yamamoto, A., Gotto, A. M., Jr., and Chan, L. (1991). No severe bottleneck during human evolution: evidence from two apolipoprotein C-II deficiency alleles. Am J Hum Genet *48*, 383-389.

Zhou, W., Alonso, S., Takai, D., Lu, S.C., Yamamoto, F., Perucho, M., and Huang, S. (2008). Requirement of RIZ1 for cancer prevention by methyl-balanced diet. Submitted, Preprint available at Nature Precedings <http://hdl.handle.net/10101/npre.2008.1732.1>

Zuckerkandl, E., and Pauling, L. (1962). Molecular disease, evolution, and genetic heterogeneity. In *Horizons in Biochemistry* (Kasha, M. and B. Pullman, eds., New York: Academic Press).

Table 1. Genetic equidistance explained by the maximum genetic diversity hypothesis. A hypothetical protein sequence of 10 amino acids is listed for each organism. Conserved positions are represented by numbers. Positions that change from time to time are represented by X. The hypothetical maximum diversity allowed for fish is 60%, for amphibian 40%, and for human 10%. The maximum diversity of 60% for fish necessarily determines that all descendants of fish, whether amphibian or human, would all have the same maximum distance of 60% with fish that is identical to the maximum diversity allowed for fish.

Species	Sequence
Fish 1	0123xxxxxx
Fish 2	012326xxxx
Fish 3	012326xxxx
Amphibian 1	012334xxxx
Amphibian 2	01233424xx
Amphibian 3	01233424xx
Human 1	012334315x
Human 2	012334315x
Human 3	012334315x

Maximum diversity (percent difference)	
Fish 1 *vs.* fish 2	60
Amphibian 1 *vs.* amphibian 2	40
Human 1 *vs.* human 2	10

Maximum distance (percent difference)	
Human *vs.* amphibian	40
Human *vs.* fish	60
Amphibian *vs.* fish	60

Table 2. Molecular clocks give consistent timing for macroevolution but inconsistent timing for microevolution. Percent identities between species are listed for four randomly selected genes. All four genes behave as good clocks in macroevolution from fish (*D. rerio)* to frog *(X. laevis)* to bird *(G. gallus)* to mouse (*M. musculus)* to human (*H. sapiens)*, which is consistent with the timing based on the fossil record as indicated for each divergence. In contrast, they give wildly contradictory timing when used to time microevolution divergence between pufferfish and zebrafish. The estimated time varies from 420 to 91 million years depending on which of the four genes is used as clock. The mutation rate or clock rate of each gene was derived from plotting the number of amino acid changes between protein sequences against species age estimated from fossil evidence. MyBP, million years before present. N.A., gene sequence not available.

	Percent identity				MyBP
	Prdm2	BTK	CytC	GCA1A	
H. sapiens v.s. *D. rerio*	39	61	80	66	450
H. sapiens v.s. *X. laevis*	55	N.A.	85	75	360
H. sapiens v.s. *G. gallus*	71	85	87	81	310
H. sapiens v.s. *M. musculus*	91	98	91	91	91
F. rubripes v.s. D. rerio	46				420
		71			400
			89		200
				91	91

Table 3. Genetic distance among different species of fungi. Three proteins, Pin1, CytC, and CMD from the baker's yeast were used to BLAST against the fungi database of NCBI . Percent identities in protein sequence between species of different genus, families, subphyla, and phyla are listed.

	Percent identity		
	Pin1	CytC	Cmd
Between genera within the same family *Saccharomycetaceae*			
S. cerevisiae v.s. D. hansenii/Debaryomyces	44	78	63
S. cerevisiae v.s. E. gossypii/Eremothecium	63		95
S. cerevisiae v.s. K. lactis/Kluyveromyces	68	84	94
Between families within the same order *Saccharomycetales*			
S. cerevisiae vs Y. lipolytica/Dipodascaceae	39	73	56
S. cerevisiae v.s. C. albicans/mitosporic Saccharomycetaceae	42	84	60
Between subphyla within the same phylum *Ascomycota*			
S. cerevisiae v.s. G. zeae PH-1/Pezizomycotina	42	67	
S. cerevisiae v.s. S. pombe/Schizosaccharomycetes	45	70	56
Between phyla within the same kingdom *Fungi*			
S. cerevisiae vs. R. oryzae/Zygomycota	43		
S. cerevisiae vs. C. Neo-formans/Basidiomycota	41	66	59
S. cerevisiae vs. C. cinerea/Basidiomycota	75	60	
S. cerevisiae vs. U. maydis/Basidiomycota	70	60	
S. cerevisiae vs. B. emersonii/Chytridiomycota			*58*

Table 4. The genetic distance between two fishes in many chromatin modifying enzymes is similar to that between a fish and a mammal. The percent identity between zebrafish (*D. rerio*) and pufferfish (*T. nigroviridis*), human (*H. sapiens*), or mouse (*M. musculus*) is shown for a number of chromatin modifying epigenetic enzymes. Genes are considered as having reached maximum distance in fishes if the distance between the two fishes is equal or slightly greater than between a fish and a mammal.

	(% identity) *D. rerio vs*.		
	T. nigroviridis	*H. sapiens*	*M. musculus*
Genes reached maximum distance			
Suv39H1/KMT1A	61	63	62
Smyd2/KMT3C	70	75	70
SET7/9/KMT7	71	73	73
PRDM11	61		64
PRDM4	57	59	59
PRDM15	60	63	63
PRMT4	81	81	85
Lsd1/KDM1	87	92	89
Jarid1b/KDM5b	62	62	62
MYST1/KAT8	87	87	85
SIRT5	71	75	71
HDAC1	80	83	82
HDAC4	78	77	79
Genes not yet reached maximum distance			
Suv4-20H1/KMT5B	59	53	54
EZH2/KMT6	82	77	76
PRDM2/KMT8	48	41	43
PRMT6	67	54	55
PRMT7	69	62	61
PRMT5	79	78	78
PRMT8	90	88	88
Jmjd2b/KDM4b	60	52	51
HAT1/KAT1	77	70	70
PCAF/KAT2B	88	82	78
CBP/KAT3A	66	61	61
MYST2/KAT7	89	77	77
Clock/KAT13D	73	70	69
SIRT3	66	55	58
SIRT4	73	64	66
SIRT6	76	73	72
SIRT7	63	55	54
HDAC3	96	92	92
HDAC8	84	73	75

Figure legends:

Figure 1. Macroevolution and microevolution. The vertical direction is macroevolution and involves major changes in epigenetic complexity over time. The horizontal direction is microevolution and involves changes in varieties within a specific level of epigenetic complexity. The estimated number of species for each kind of organisms is indicated in parentheses. Time is not to scale and in the direction from past to future.

Figure 2. Genetic distance between two splitting organisms at various times during macroevolution. A 10 amino acid peptide with amino acids represented by numbers is shown to illustrate the dissimilarity or genetic distance between the species at various times during evolution. X represents amino acid positions that may change from time to time. A fraction of these X residues may be shared in different organisms due to common external environments that may differ from time to time. The ancestor organism A0 gives rise to two descendant lineages that gradually accumulate genetic distance until reaching a maximum at time T1. At this time, a punctuational jump in epigenetic complexity occurs in one of the lineages generating B1. The descendant organism A1 remains phenotypically similar to the ancestor A0. The lineage leading to B1 is phenotypically similar to A1 prior to the punctuational jump at time T1. The epigenetic jump in B1 reduces the genetic diversity of B1, as indicated by the reduction in the number of X positions.

Figure 3. The Neo-Darwinian hypothesis versus the maximum genetic diversity hypothesis. (A) The Neo-Darwinian model of microevolution and macroevolution. Genetic distance increases with time with no maximum cap. Fish and amphibians are used as examples. The transition from fish to amphibian is indicated by the dashed line. The starting point of the dashed line represents the time when amphibian epigenotype or phenotype first

became obviously distinct from that of fish. **(B)** Model of microevolution and macroevolution based on the MGD hypothesis.

Figure 4. Inverse relationship between genetic diversity and epigenetic complexity. (**A)** Maximum genetic diversity within each type of organisms in cytochrome c correlates with the time since the first appearance of each type. The percent amino acid change in cytochrome c within each type of organism was obtained by BLAST against protein database at the National Center for Biotechnology Information. Similar data was first reported in the 1960s (Fitch and Margoliash, 1967; Margoliash, 1963). (**B)** High epigenetic complexity as measured by the number of cell types per organism inversely correlates with genetic diversity and the time since the first appearance of each organism. The first eukaryote is more complex than bacteria in having more cellular compartments and more epigenetic enzymes. The number of cell types is estimated based on the complexity of the nervous systems to be relatively the most in the first primate, less in the first mammals, and still less in the first vertebrate. The figure is meant to show this relative trend but does not intend to show the precise number of cell types.

Figure 5. Genetic distance between organisms at various times during macroevolution. **A.** Genetic distance according to the Neo-Darwinian gradual mutation hypothesis. **B.** Genetic distance according to the MGD hypothesis. A 10 amino acid peptide with amino acids represented by numbers is shown to illustrate the dissimilarity or genetic distance between the species at various times during evolution. X represents amino acid positions that may change from time to time. A fraction of these X residues may be shared in different organisms due to common external environments that may differ from time to time.

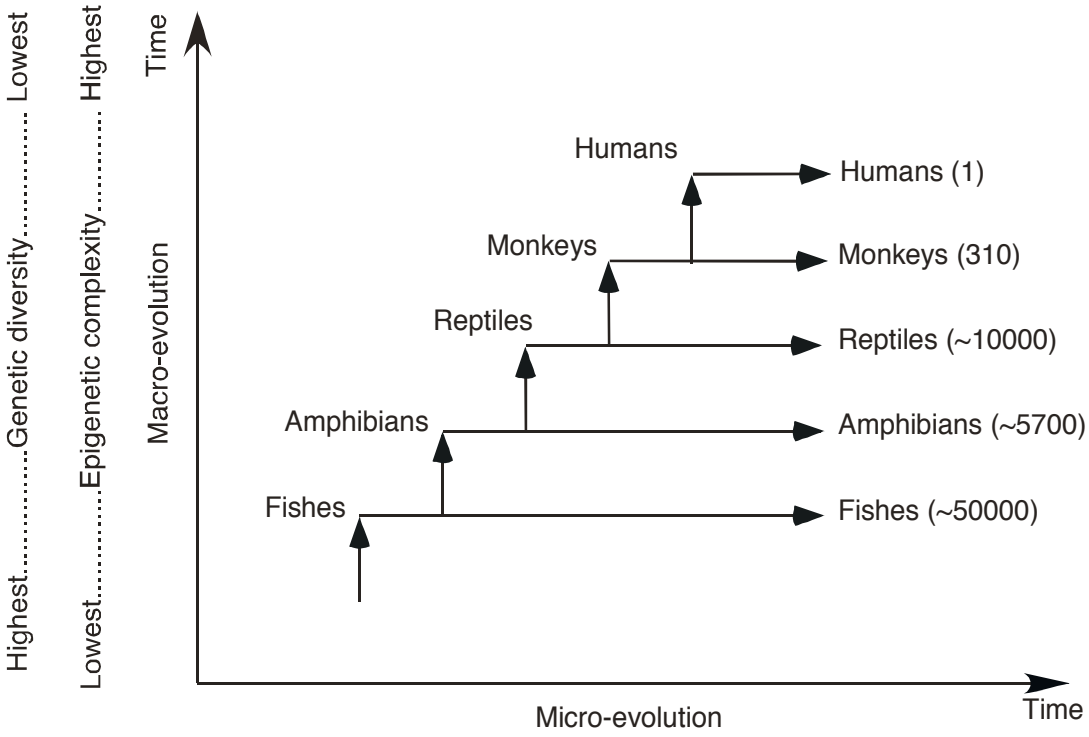
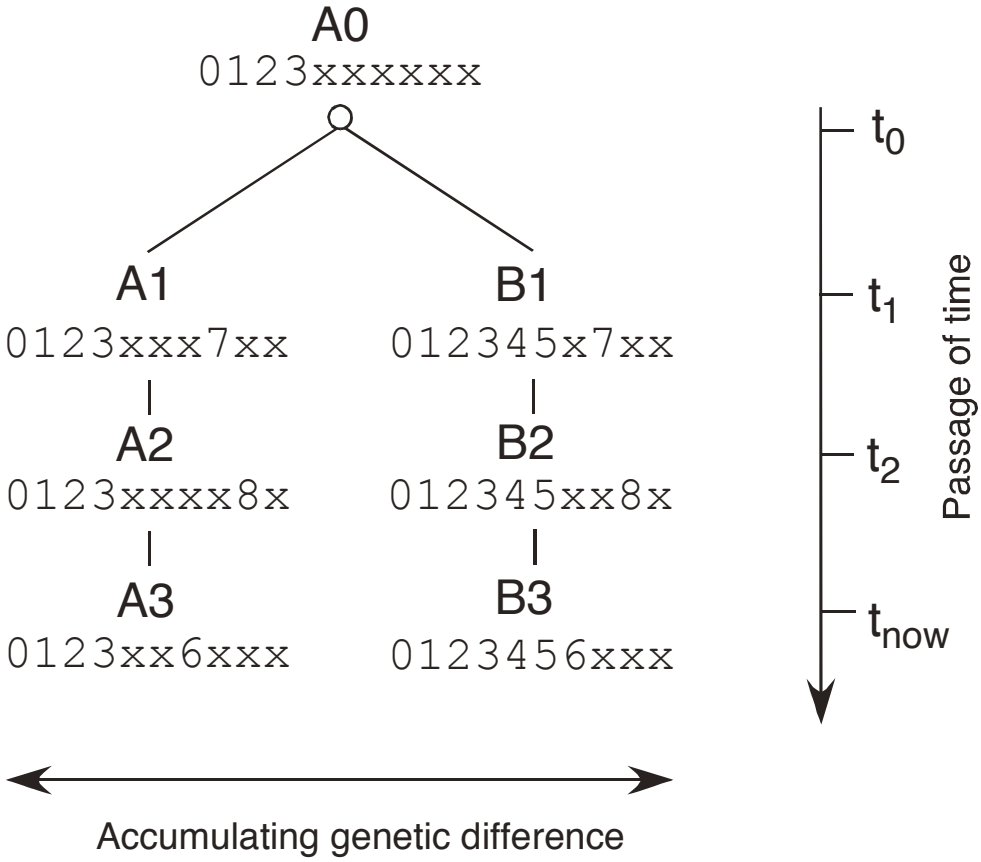


Figure 1

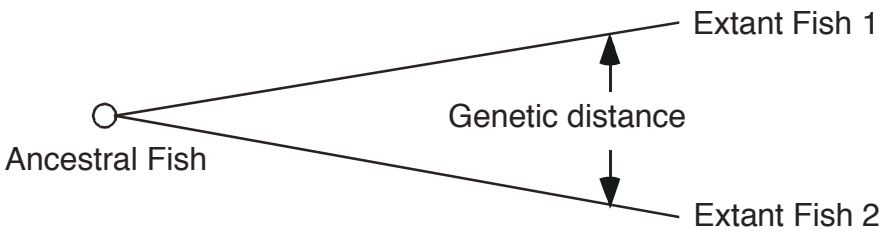


Distance (A1-B1) = Distance (A3-B3) = 50% dissimilarity
Distance (A1-B2) = 60% dissimilarity > Distance (A3-B3)
Distance (A2-B3) = 60% dissimilarity > Distance (A3-B3)
Distance (A3-B3) = Maximum distance within A3 (A3.1-A3.2)

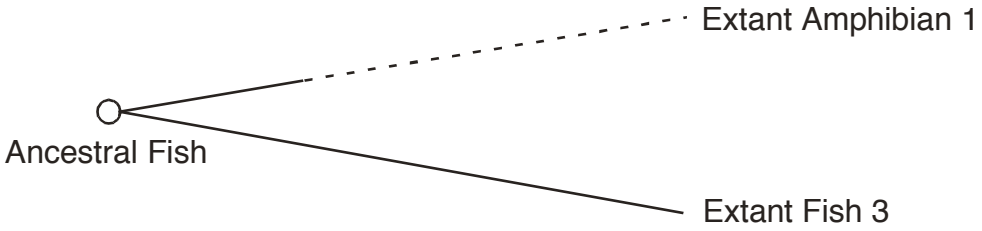
Figure 2

A

Neo-Darwinian microevolution model

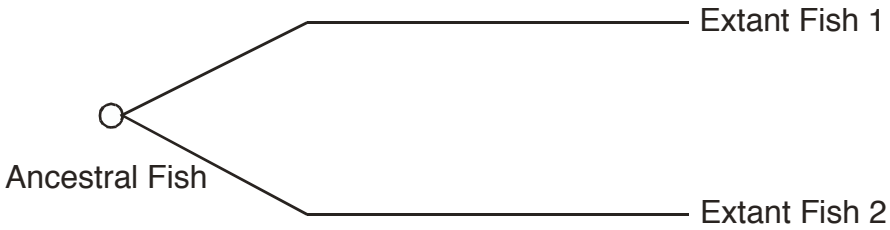


Neo-Darwinian macroevolution model



B

MGD microevolution model



MGD macroevolution model

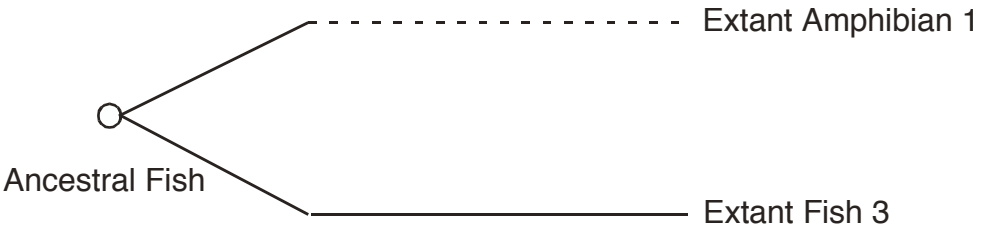
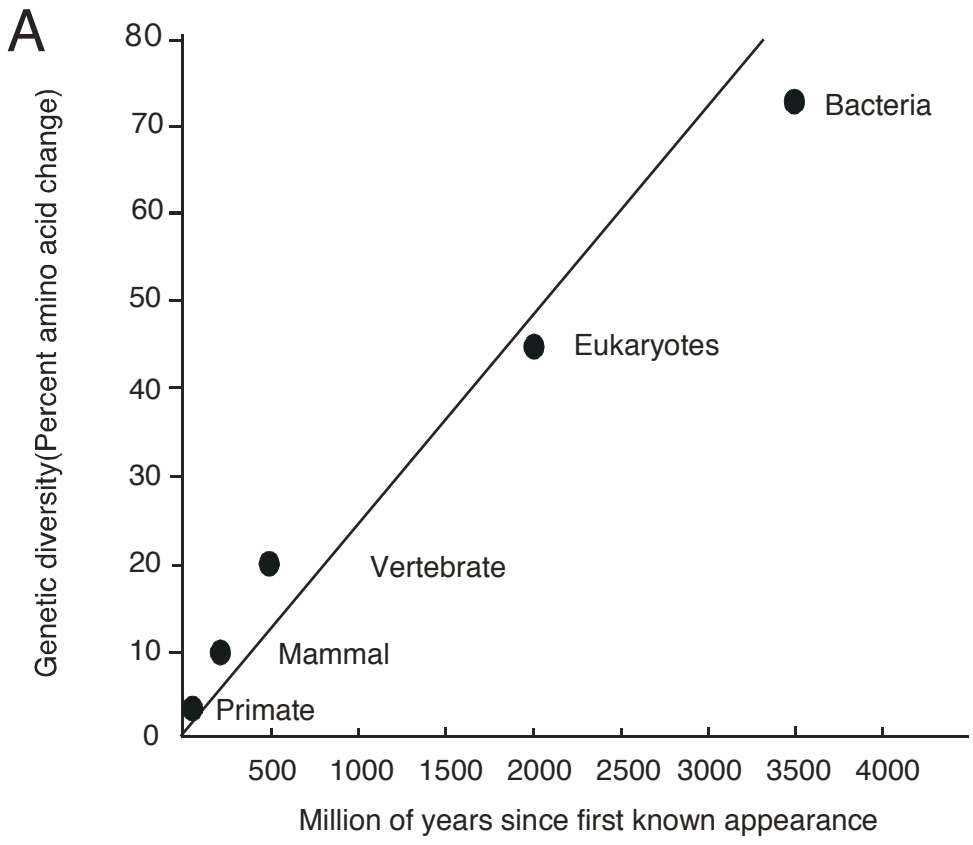
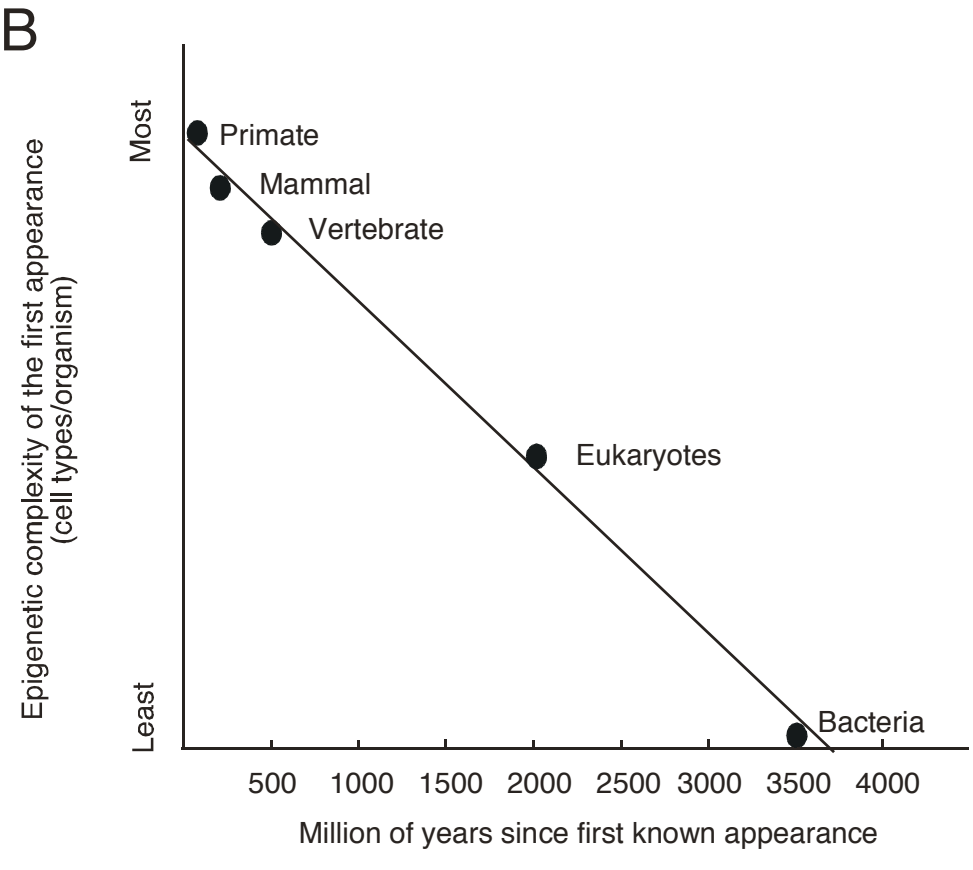


Figure 3

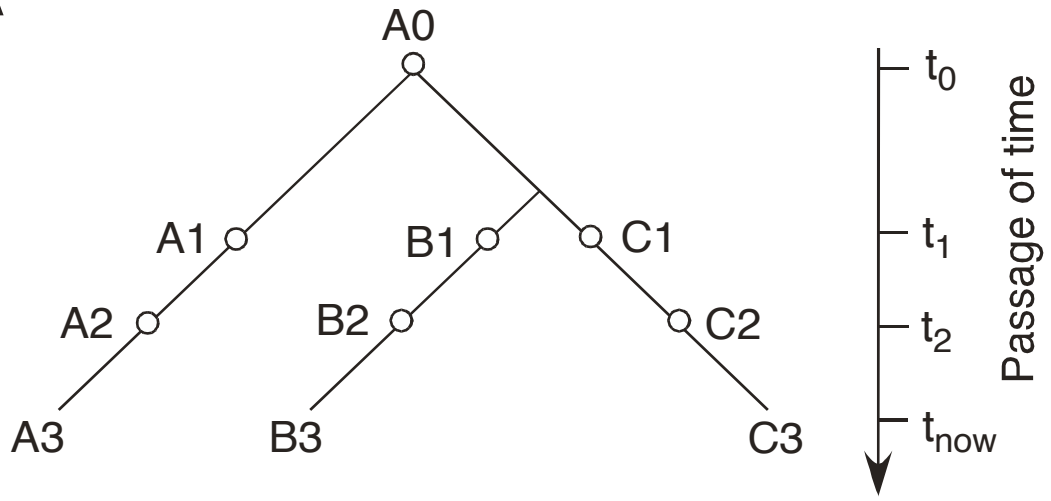




Least ..... Genetic diversity.....Most

Figure 4

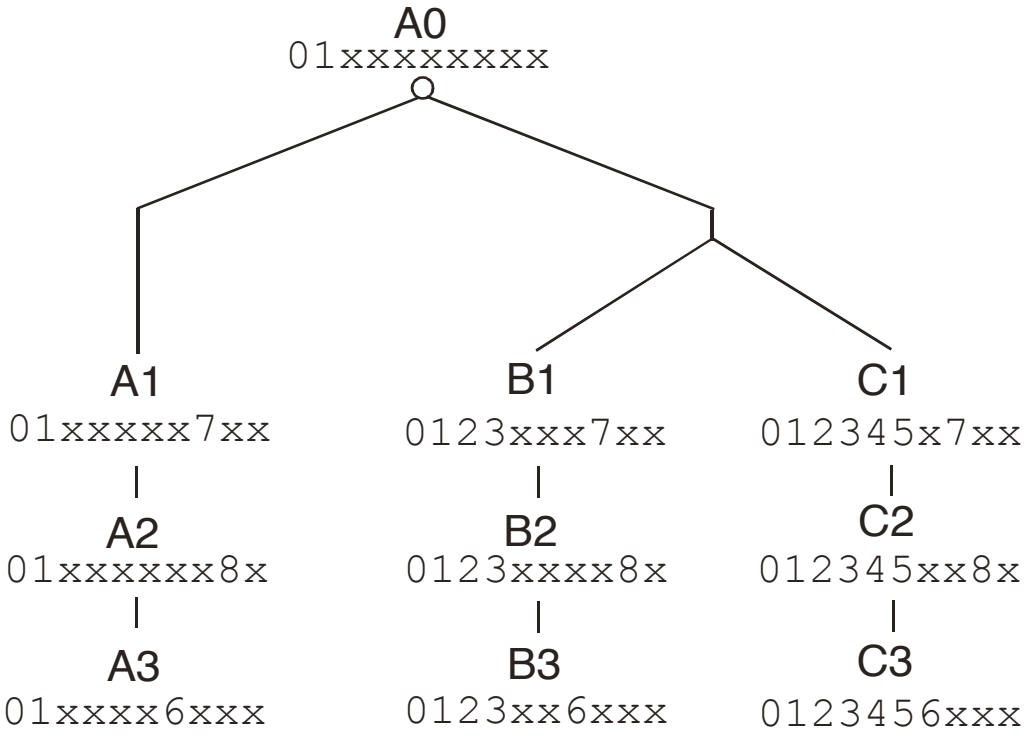
A



Distance (C1-B2) < Distance (C3-B3)

Distance (C1-A3) < Distance (B3-A3)

B



Distance (C1-B2) > Distance (C3-B3)

Distance (C1-A3) > Distance (B3-A3)

Figure 5