

# The Genetic Equidistance Result of Molecular Evolution is Independent of Mutation Rates

Shi Huang

The Burnham Institute for Medical Research, 10901 North  
Torrey Pines Roads, La Jolla, CA 92037

Corresponding author: Shi Huang, E-mail: shuangtheman@yahoo.com

Received December 02, 2008; Accepted December 22, 2008; Published December 26, 2008

**Citation:** Shi Huang (2008) The Genetic Equidistance Result of Molecular Evolution is Independent of Mutation Rates. J Comput Sci Syst Biol 1: 092-102.

**Copyright:** © 2008 Shi Huang. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

## Abstract

The well-established genetic equidistance result shows that sister species are approximately equidistant to a simpler outgroup as measured by DNA or protein dissimilarity. The equidistance result is the most direct evidence, and remains the only evidence, for the constant mutation rate interpretation of this result, known as the molecular clock. However, data independent of the equidistance result have steadily accumulated in recent years that often violate a constant mutation rate. Many have automatically inferred non-equidistance whenever a non-constant mutation rate was observed, based on the unproven assumption that the equidistance result is an outcome of constant mutation rate. Here it is shown that the equidistance result remains valid even when different species can be independently shown to have different mutation rates. A random sampling of 50 proteins shows that nearly all proteins display the equidistance result despite the fact that many proteins have non-constant mutation rates. Therefore, the genetic equidistance result does not necessarily mean a constant mutation rate. Observations of different mutation rates do not invalidate the genetic equidistance result. New ideas are needed to explain the genetic equidistance result that must grant different mutation rates to different species and must be independently testable.

**Key words:** Genetic equidistance result; evolution; molecular clock; Neo-Darwinism

## Introduction

The Neo-Darwinian theory of evolution is the dominant mainstream theory for evolution and widely taught to biologists and the public at large. It suggests that evolution is a process of natural selection of randomly occurring fitter mutations. Macroevolution involves the same process as microevolution or population genetics and is simply prolonged microevolution. A major prediction of this theory is that macroevolution would take longer time and thus accumulate more molecular mutations or changes than microevolution. This prediction can be tested by analyzing molecular similarity among species, which was first done in the early 1960s (Doolittle and Blombaeck, 1964; Margoliash, 1963; Zuckerkandl and Pauling, 1962). Closely related species (in phenotypes or genealogy) should show more molecular

similarity than distantly related species. However, while this prediction can be demonstrated in some cases (e.g., human is closer to chimpanzees than to monkeys in both phenotypes/genealogy and molecules), it has also been falsified in many other cases. For example, the molecular distance between two subpopulations of medaka fish that had diverged for ~ 4 million years is 3-fold greater than that between humans and chimpanzees that are thought to have diverged for 5-7 million years (Kasahara et al., 2007). The molecular distance between two different fungi can be just as great as that between fungi and humans, which is completely unexpected from Neo-Darwinism and would indeed be shocking to anyone with a Neo-Darwinian mindset.

Such exceptions are obviously inconvenient to the widely publicized theory and hence rarely made known outside the small circle of molecular evolution specialists. One important consequence of these exceptions is that they make it impossible to trust the molecular phylogenies constructed by the present methods of molecular analysis. These methods assume, despite numerous factual exceptions or contradictions, that closer molecular similarity always means closer evolutionary distance. As a result, major conflicts between molecular dating and fossil dating are common. Given the frequent factual contradictions, it is almost certain that the theoretical basis for the interpretation of the major facts in molecular evolution is not completely correct.

In mathematics or physics, one exception is sufficient to doom any theory. The science of biology or any scientific discipline for that matter should not be held to a lower standard. When one allows exceptions, one has effectively rendered the theory non-testable and non-scientific. Such a theory would be no different from a false theory that happens to explain a fraction of nature while being contradicted by the rest. The only way to distinguish a true theory from a false or incomplete one is to see if it has not a single factual exception within its domain of application or relevance.

A most remarkable result of molecular changes during macroevolution is the near linear correlation between genetic distance as measured by DNA/protein sequence dissimilarity and time of species divergence as inferred from fossil records. This result is not predicted by Neo-Darwinism. It has been commonly interpreted to mean a constant mutation rate, which in turn directly provoked the molecular clock hypothesis. However, this hypothesis must negate the idea of selection, the cornerstone of Neo-Darwinism. While the Neo-Darwinian selection theory has spectacularly failed the molecular test, its ad hoc substitute for the domain of molecular evolution, the molecular clock hypothesis, is also imperfect and widely known to have countless contradictions. It is also obviously incoherent or schizophrenic to have two vastly different and non-connected theories of evolution, one for phenotype evolution based on the idea of selection and the other for molecular evolution based on the negation of the idea of selection. It is also intuitively absurd given the proven truth that phenotypes and genotypes are inseparably connected. Thus, the two theories cannot both be correct for macroevolution. I show here that the molecular clock hypothesis is merely an ad hoc restatement of a factual observation, the genetic equidistance result. It is a tautology and does not qualify as a scientific theory with true explanatory power.

In the early days of molecular evolution studies, genetic distance was simply represented by percent nonidentity in a given protein sequence. 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 simpler or less complex species that evolved earlier such as amphibians and fishes. The second aligns a simpler outgroup organism such as fishes against those more complex sister species that appeared later such as amphibians and mammals.

The first alignment indicates a near linear correlation between genetic distance and time of divergence, implying indirectly a constant mutation rate among different species. For example, human is closer to mouse, less to bird, still less to frog, and least to fish. The second alignment shows the genetic equidistance result where sister species are approximately equidistant to the simpler outgroup. For example, human, mouse, bird, and frog are all equidistant to fish in any given protein dissimilarity. Since all of the sister species are also equidistant in time to the outgroup fish, this directly triggered the idea of constant or similar mutation rate among different species, no matter how different they may be. Since both alignments use the same sequence data set, certain information may be revealed by either alone. But the data that most directly and obviously support the interpretation of a 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 (Zuckerlandl and Pauling, 1962). Margoliash in 1963 performed both alignments and made a formal statement of the molecular clock after noticing the genetic equidistance result (Kumar, 2005; 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 either 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 comparisons that produced the equidistance result, as Margoliash stated (Margoliash, 1963), "disregard the relation of amino acid substitutions observed to the actual number of effective mutational events which occurred." So, the equidistance result and the molecular clock hypothesis

were originally established by percent nonidentity in protein sequences. The actual number of mutational events in the past evolutionary process is irrelevant to the equidistance result, and is impossible to discern anyway if the percent nonidentity in fact represents the maximum that has long been reached before present time.

While the concept of a maximum distance is intuitively obvious especially for long evolutionary time, it has rarely even surfaced as an issue of concern in the molecular evolution field. All existing mathematical methods of relating percent nonidentity to the actual number of mutational events, such as the Poisson correction distance, make the unspoken assumption that the observed percent nonidentity today is a result of a linear and gradual increase in distance in the past and will continue to increase in the future. But such assumption is simply just that and has zero factual support. Given the uncertainty of such assumptions, it is much more prudent to base conclusions on the primary data, percent nonidentity, rather than on some mathematical transformations of the primary data where the assumptions for such mathematical models are groundless and more likely to be false than true. Regardless, however, the equidistance result will not be affected by these mathematical transformations.

The genetic equidistance result has been independently confirmed for numerous proteins and numerous species. This result is the most remarkable result of molecular evolution since it was completely unexpected from classical Neo-Darwinian theory. However, what has become popular known today is not the result itself but the molecular clock interpretation of it (Avice, 1994; Li, 1997; Nei and Kumar, 2000). Even the original discoverer of this result, E. Margoliash, has subsequently avoided highlighting the result. In a 1967 paper, Fitch and Margoliash compared the cytochrome c of 20 species (Fitch and Margoliash, 1967). Table 3 of the paper clearly showed the genetic equidistance result, for example, the yeast *Saccharomyces* has 57 mutational differences from the yeast *Neurospora*, 57 from monkey, 56 from human, and 58 from kangaroo. But Fitch and Margoliash did not comment on the obvious equidistance and instead concluded the opposite. "Indeed, from any phylogenetic ancestor, today's descendants are equidistant with respect to time but not, as computations show, equidistant genetically. Thus the method indicates those lines in which the gene has undergone the more rapid changes. For example, from the point at which the primates separate from the other mammals, there are, on the average, 7.5 mutations in the descent of the former and 5.8 in that of the latter, indicating that the change in the cytochrome c gene has been much more rapid in the descent of the primates

than in that of the other mammals."

Here, Fitch and Margoliash considered equidistance to mean exact identity in distance. But the equidistance result shows minor variations around a mean and should be considered an approximate result. Indeed, its interpretation by the clock idea is widely known to be approximately constant. The eagerness to interpret small variations of the equidistance result as significant differences in mutation rates probably reflected a compromise to accommodate the mindset of classical evolution biologists who view the idea of a constant mutation rate "unthinkable". (Nei and Kumar, 2000). While anyone with a high school education would easily see the contradiction between facts and theories if the equidistance fact is taught alongside the Neo-Darwinian theory, few could see the much more subtle contradiction between the two different theories, especially given that both theories routinely take exceptions for granted. This is perhaps why the equidistance result was ignored while its restatement posing as the molecular clock theory was promoted instead. And the molecular clock hypothesis was never presented as an ad hoc interpretation of the equidistance result after the 1963 Margoliash paper, as if the hypothesis were derived from logical reasoning based on some biological principle.

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 that will continue to increase in the future. None of these assertions are self-evident. Nor do they have direct experimental support. They are all ad hoc interpretations of the genetic equidistance result.

Unlike the genetic equidistance result, most other independent results show that different species have different mutation rates or clock rates (Avice, 1994; Goodman et al., 1974; Jukes and Holmquist, 1972; Laird et al., 1969; Langley and Fitch, 1974; Li, 1997; Nei and Kumar, 2000). A recent study of DNA and protein sequences of ancient fossils (Neanderthals, dinosaurs, and mastodons) challenged a fundamental premise of the molecular clock hypothesis (Huang, 2008a). It shows that genetic distance had not always increased with time in the past history of life on Earth. Neanderthals are more distant than modern humans are to the outgroup chimpanzees in non-neutral DNA sequences, contrary to expectations from the molecular clock interpretation of the equidistance result (Huang, 2008a). This

result of Neanderthals has been independently confirmed using protein sequences (Green et al., 2008). So, how can the molecular clock hypothesis be both correct (consistent with the genetic equidistance result) and wrong (inconsistent with results of variable clocks and ancient fossils).

The constant mutation rate idea has often been violated when it was given an independent meaning (from the equidistance result) that is testable (Avice, 1994; Ayala, 1999; Goodman et al., 1974; Green et al., 2008; Ho and Larson, 2006; Huang, 2008a; Jukes and Holmquist, 1972; Laird et al., 1969; Langley and Fitch, 1974; Li, 1997; Nei and Kumar, 2000; Pulquerio and Nichols, 2007). But it is non-testable or non-scientific when it has no independent meaning or merely means a restatement of the empirical result of equidistance. It is correct only in the trivial sense of tautology. It is true as a factual restatement of the equidistance result. But it has not been independently proven true as a scientific explanation of the equidistance result.

The tautology fallacy of the constant mutation rate interpretation can be illustrated by a simple example. Two turtles and a rabbit are running a 1-mile race. No one watches the race and one is only informed of the race result by a video camera aimed at the finish line. The result of the race is that the turtles and rabbit arrive at the finish line at approximately the same time in 1 hour. To explain this fact, one can deduce from the fact the same speed hypothesis. One can also deduce from the fact many other hypotheses such as 'God did it'. To determine which hypothesis is correct, one must perform independent tests of the predictions of each hypothesis. For it to be a true explanation and not a tautology, the same speed hypothesis or any other hypothesis must be backed up by independent evidence. Of course, any independent tests of running speed would reveal that the two turtles have similar speeds while the rabbit is much faster. After performing such independent tests, one can conclude that the same speed hypothesis is likely a true explanation for the two turtles but cannot be true for the rabbit. The hypothesis is a real explanation for the two turtles but is merely a tautology for the rabbit.

The molecular clock interpretation of the equidistance result is the equivalent of the same speed hypothesis for the turtle and rabbit race. The automatic rephrasing of the equidistance result as the 'constant mutation rate' has hindered a direct understanding of the result. All past efforts on this empirical observation have focused instead on explaining the constant mutation rate as if it were an empirical fact of the past mutation process. Various selectionist ideas as well as non-selectionist ideas have been proposed to account for the constant mutation rate (Clarke, 1970; Kimura,

1968; Kimura and Ohta, 1971; King and Jukes, 1964; Richmond, 1970; Van Valen, 1974). The 'Neutral Theory' has come out as the favorite. 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). Thus, despite numerous efforts in the past 45 years, the constant mutation rate remains unexplained by any fundamental principle of biology. However, no one has even attempted to explain the real original empirical fact, the genetic equidistance result, without presupposing a constant mutation rate.

The constant mutation rate idea has often been mistakenly treated as the same thing as the equidistance result. The common practice of interpreting minor variations from exact equidistance as significant has in part caused the vast majority of biologists to be unaware of the equidistance result. Whenever the constant mutation rate idea is violated, many would automatically infer that there would be no equidistance. It is commonly thought that if there is no constant mutation rate, there is no equidistance result. And if there is equidistance, then there would be constant mutation rate. The currently common practice of relative rate test is used to select genes that would show equidistance and hence constant mutation rate. Such genes would next be used for building phylogenetic trees, while genes with non-equidistance would be excluded. Here I show that the equidistance result remains valid regardless of independent results showing violations of the constant mutation rate. The genetic equidistance result is extremely robust and universal.

## Results

### The Genetic Equidistance Result is Independent of Variation in Mutation Rates in Different Species

It can be easily shown that different species have different mutation rates. A typical violation of the constant mutation rate can be illustrated by the Lsd1 protein. The time of divergence for two different bony fishes such as pufferfish (*T. nigroviridis*) and zebra fish (*D. rerio*) is ~ 140-200 MyBP (million years before present) as inferred from fossil records (Powers, 1991), or from slow evolving proteins such

as cytochrome c (unpublished observation). However, the genetic distance between the two fishes (13% dissimilarity in protein sequence) in Lsd1 is greater than that between chickens and mice (6% dissimilarity) which diverged ~ 310 MyBP, much earlier than the two fishes. This indicates that the mutation rate in Lsd1 is higher in fishes than in birds and mammals. This result holds regardless whether the mutation rate is calculated using percent nonidentity or other methods such as the Poisson correction distance or the gamma distance (data not shown).

However, Lsd1 shows the equidistance result where sea urchins are approximately equidistant to all vertebrates (31% dissimilarity to fishes, 30% to frogs, 27% to chickens, 28% to mice). So violation of a constant mutation rate does not mean violation of the genetic equidistance result. For a protein such as cytochrome c, the fishes have comparable mutation rate as birds and mammals and it is well known that most vertebrates are equidistant to a simpler outgroup in this protein (Fitch and Margoliash, 1967; Margoliash, 1963). The equidistance result therefore holds for both types of proteins that either has a constant mutation rate or has not. It is independent of mutation rate variations.

One of the best known genes that show vastly different mutation rates in different species is the SOD gene (Ayala, 1986; Ayala, 1997). However the equidistance result still holds for this gene as shown by Table 4 of the 1986 paper by Ayala, where yeast is approximately equidistant (69-63 changes) to human, rat, horse, cow, fish, and fly (Ayala, 1986). So while SOD can be shown to have different mutation rates, the same data set can also be used to justify a perfectly constant clock for SOD, if the constant clock interpretation of the equidistance result is granted. However, Ayala was apparently unaware of this other side of his data that shows the equidistance result, and went on to conclude that SOD has a variable molecular clock.

Ever since the 1967 paper by Fitch and Margoliash (Fitch and Margoliash, 1967), the genetic equidistance result has been consistently ignored by the molecular evolution field whenever a gene can also be simultaneously shown to have variable mutation rates. This suggests that the field did not really believe its own interpretation of the equidistance result and preferred to ignore the interpretation whenever it was contradicted by other observations. Perhaps because of the lack of a convincing interpretation, the equidistance result, the most universal and conspicuous fact of molecular evolution that should have been taught to all biologists and the public, has been made essentially unknown to almost all biologists including most evolution biologists. (For example, there is no indication that Ayala knows the result when you read his papers where the equidistance result was plain apparent but was never mentioned.) I independently rediscovered the equidistance result in 2006 when I did a homology comparison of my favorite gene RIZ1. I was shocked by it since my Neo-Darwinian mindset would never have expected it. I soon realized that no one has a sensible explanation for it yet.

I also found that flowering plants have higher mutation rates than mammals and yet flowering plants and mammals are still equidistant to the simpler outgroup protists. Biology textbooks commonly teach that flowering plants and mammals coevolved. Based on the fossil record, the first flowering plants evolved at about the same time as the earliest mammals during the early Cretaceous period, about 125 MyBP. I randomly selected 5 proteins from the apple tree (*M. domestica*) and determined the sequence identity in these 5 proteins between the apple tree and the flowering plant *A. thaliana* (Table 1). The time of divergence between these two flowering plants is not precisely known but must be less than 125 million years. I also determined the sequence identity in these 5 proteins between two highly diverged mammals (human and cattle or *B. taurus*), between

	Percent identity					MyBP
	ADH	EF1a	eIF2b	Pin1	PP1	
<i>M. domestica</i> v.s. <i>A. thaliana</i>	83	95	73	79	92	125
<i>H. sapiens</i> v.s. <i>B. Taurus</i>	94	100	96	98	100	125
<i>H. sapiens</i> v.s. <i>G. gallus</i>	87	98	94	N.A.	100	310
<i>H. sapiens</i> v.s. <i>X. tropicalis</i>	86	96	86	86	99	360
<i>H. sapiens</i> v.s. <i>D. rerio</i>	80	94	77	79	97	450

**Table 1: Genetic distance within flowering plants is greater than that within mammals after similar amount of time of evolution.** Five proteins from the apple tree (*M. domestica*) were randomly selected for determining the genetic distance between the apple tree and the flowering plant *A. thaliana*, between human and cattle (*B. taurus*), between human and bird (*G. gallus*), between human and amphibians (*X. tropicalis*), and between human and fish (*D. rerio*).

	Percent identity, <i>S. lemnae</i> vs.	
	<i>H. sapiens</i>	<i>A. thaliana</i>
EF1A	74	73
IDH	71	68
H3	85	85
DPOLA	37	35
Actin	66.4	66.8
Hsp70	72	73
H4	85	83
ERF1	57	55
Tubulin B	91	90
Tubulin A	84	90
Calmodulin	90	89

**Table 2: Humans and plants (*A. thaliana*) are equidistant to protists (*S. lemnae*).**

All available protein sequences of *S. lemnae* from Genbank were analyzed by BLASTP against human and plant proteins. The informative proteins are listed here.

human and bird (*G. gallus*), between human and amphibians (*X. tropicalis*), and between human and fish (*D. rerio*). As shown in Table 1, the sequence identity between the two flowering plants is much less than that between the two mammals and is equivalent to that between human and fish. So, the flowering plants have reached a genetic distance that is much higher than that reached by the mammals after about the same amount of time of evolution. The genetic distance 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.

Yet, despite the faster rate of genetic divergence in flowering plants, they and mammals are equidistant to the outgroup protists. For example, for the EF1a gene, the alveolata protist (*S. lemnae*) is 74% identical to humans and 73% identical to *A. thaliana* (Table 2). A random sampling of all available proteins of the protist *S. lemnae* at the Genbank revealed 11 informative proteins that all showed approximate equidistance to humans and plants (Table 2). Among these, 7 showed more similarity between protist and human than between protist and plant while 3 showed less ( $P > 0.05$ ). Again, violation of a constant mutation rate does not mean violation of the genetic equidistance result.

### Most Proteins Show the Genetic Equidistance Result

Many proteins are found to violate the molecular clock in experiments examining the genetic distance between similar species such as two different fishes. For example, pufferfish (*T. rubripes*) and zebrafish (*D. rerio*) are believed to have diverged not more than 140-200 MyBP based on the first fossil evidence of teleostei in the early Cretaceous period (Powers, 1991). One would expect most 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 proteins, I found only 19 (48%) with more identity between the two fishes than between human and bird. So about half of all genes in fishes have faster mutation rate than the molecular clock deduced from macroevolution of vertebrates. It is now common practice to exclude these genes in calculating divergence time for microevolution (Kumar and Hedges, 1998).

The fact that about half of all genes have different mutation rates in different species offers another way to resolve whether the genetic equidistance result is independent of the measurable variations in mutation rates in different species. If most genes can be shown to display the equidistance result despite the independent fact that half of them have different mutation rates in different species, then we can conclude that the equidistance result is independent of rate variations.

I randomly selected 50 proteins from frogs (*X. laevis*) and compared each to chickens (*G. gallus*) and humans (*H. sapiens*). Among these proteins, 11 (22%) showed exact equal distance (to frogs) of chickens and humans, 28 (56%) showed greater distance between humans and frogs than between chicken and frogs, and 11 (22%) showed less ( $P > 0.05$ ). For most of these proteins (46/50 or 92%), the difference between chicken and human in their percent identities to frogs is less than 4% (Table 3), indicating approximate equidistance. For 4 other proteins (4/50 or 8%), the difference between chicken and human in their percent identities to frogs is 7% to 8%. However, all 4 proteins showed approximate equidistance when sea urchins were used as the outgroup (Table 3). Thus, the seeming non-equidistance to frogs in these 4 proteins may not represent a significant violation of the equidistance result. Since all of the 50 randomly selected proteins showed the equidistance result,

	<u>Percent identity</u>			
	<u>F-C</u>	<u>F-H</u>	<u>F-SE</u>	<u>H-SE</u>
HGF/AAB34354	70	67		
mN-cadherin/ AAB37685	74	73		
TrkB/ NP_001079579	95	94		
APP747/ AAB24853	86	86		
Rb1/ AAB23173	61	60		
ADPRT/ P31669	79	77		
B-catenin/ NP_001084045	97	97		
Chordin/ NP_001081778	59	51	34	34
c-Mos/ NP_001081563	57	53		
PR/ NP_001079100	64	68		
Chk1/ NP_001082039	75	76		
P53/ NP_001081567	52	51		
Hsp70/ P02827	95	97		
Prion/ NP_001082180	40	36		
Noggin/ NP_001079113	86	78	28	27
Ralb/ NP_001084154	94	94		
Desmin/ NP_001080177	81	78		
AAI60739	39	42		
AAI60733	39	36		
AAI60732	72	69		
AAI60718	64	63		
AAI60762	64	62		
AAI60704	34	36		
P85PI3K/ AAL68953	81	80		
DNA ligase IV/ AAL56554	71	72		
Stat5/ AAK94906	86	86		
P2X4/ AAG45104	70	70		
HIRA/ CAC41093	77	79		
DNase1/BAB20384	42	44		
Lis1/ AAK52334	96	96		
P450/ BAA37080	63	62		
Ptc1/AAK15463	81	79		
Paxillin/AAG40874	75	75		
Cdc21/NP_001081448	86	85		
ND4/AAA66467	60	59		
ND1/AAA66458	71	64	62	59
Rest/ AAF06720	64	61		
PTEN/AAD46165	89	88		
EphB2/ AAB94603	87	87		
RpS6/ AAD01647	95	96		
Cdc46/BAA09949	87	86		
Mcm2/ BAA09948	87	85		
DNMT1/ BAA11458	76	77		
CCND1/ CAA61664	84	77	45	45
EIF4E/ NP_001084017	92	91		
AAI60783	45	44		
Sdad1/ AAH70991	80	80		
COUP-TF2/ NP_001095229	77	75		
Src/ NP_001079114	87	87		
Serpin/NP_001079103	38	38		

**Table 3: The genetic equidistance result for 50 randomly selected proteins.** Fifty proteins were randomly selected from frogs/*X. laevis* (F) and compared with chicken/*G. gallus* (C) and human (H). Percent identities in protein sequence are shown. For 4 of these proteins that showed greater variations from exact equidistance, a comparison with sea urchin/*S. purpuratus* (SE) was made to confirm approximate equidistance. Protein names and accession numbers are shown.

whereas one expects only half of them since at least half is known to have non-constant mutation rates, the data suggest that the equidistance result is independent of the constancy of mutation rates ( $P < 0.0001$ ). It also suggests that nearly all vertebrate proteins show the genetic equidistance result.

## Discussion

It is commonly argued that the molecular clock may be a stochastic clock. It may not tick at a constant rate like a real clock. It may be sometimes slow and sometimes fast. But the average rate over long time is constant and predictable. Thus to explain the equidistance to sea urchin of zebra fish and mouse, when zebra fish can be shown to have faster mutation rates than mouse in the last 140-200 million years as discussed above, it is argued that the ancestor of zebra fish must have had slower mutation rate than the ancestor of mouse. Similarly, to explain the equidistance to protists of flowering plants and mammals, when flowering plants can be shown to have faster mutation rates than mammals in the last ~125 million years as discussed above, it is argued that the ancestor of flowering plants must have had slower mutation rate than the ancestor of mammals.

Such argument has several fatal flaws. First, it is not testable and hence not scientific. It cannot be expected to have independent factual support and is merely a tautology. It has no independent merit and cannot exist independent of the result it is trying to explain.

Second, it does not have a biological reason or mechanism. It is not a deduction of a fundamental biological principle.

Third, it is not logical. The constant mutation rate idea is obviously a sensible explanation for the equidistance to sea urchins for a million different individuals of zebra fishes that can be independently confirmed to have similar mutation rate. By logical inference, if the constant mutation rate idea is a true explanation for the equidistance of different organisms that can be independently confirmed to have the same mutation rate, then it already means that different organisms that can be independently confirmed to have different mutation rates would not be equidistant to an outgroup. The same idea therefore cannot also be the reason for the equidistance of different organisms that can be independently confirmed to have different mutation rates.

Finally, if the constant mutation rate represents a statistical average, it would be useless for predicting whether a specific individual species would have a constant or non-constant rate for any given time period. It would invalidate

the whole enterprise of molecular phylogeny as is currently practiced. For example, if we do not have the fossil record for flowering plants and relied solely on molecular analysis as shown in Table 2, we would have reached the absurd conclusion that apple tree and *A. thaliana* have diverged 450 MyBP. This example shows that similar errors due to non-constant mutation rate could invalidate many other molecular dating results, including the 5-7 million divergence time between humans and chimpanzees (Wilson and Sarich, 1969), which is in sharp conflict with the fossil estimation of ~18 million years (Lewin, 2005; Pilbeam, 1968; Schwartz, 1984; Schwartz, 2005; Simons, 1961; Simons and Pilbeam, 1965). If we do not accept the kind of molecular dating for the flowering plants, we also have no reason to trust the same kind of molecular dating for the human-chimpanzee split.

If we insist on restating the genetic equidistance result as constant mutation rate, we still have not explained the biological reason for the constant mutation rate, since no theories so far proposed to explain the constant mutation rate are complete explanations. From such restatement, we have learned nothing about the biology behind the equidistance result.

A proper way to establish that small variations in distance are not significant is to sample multiple individuals of each sister species. A single individual of species A may be either more or less distant to an outgroup than a single individual from sister species B. However, if large number of individuals were analyzed, the mean distance to the outgroup should not be significantly different between the two sister species. Also, the number of comparisons that show A to be more distant to the outgroup than B should be similar to the number of comparisons that show A to be less distant to the outgroup than B. This kind of analysis has shown that humans and chimpanzees are equidistant to gorillas (Huang, 2008a). In a study using mitochondrial DNAs from 30 randomly selected human individuals and 30 chimpanzee individuals, the number of comparisons that showed greater distance between humans and gorillas than between chimpanzees and gorillas (13) was similar to the number of comparisons that showed greater distance between chimpanzees and gorillas than between humans and gorillas (11), while 6 showed that human and chimpanzees are exactly equal distant to gorillas (Huang, 2008a).

At this point in time, for most species, we do not yet have sequence information for multiple individuals of a species. Thus it is not yet possible to statistically establish that the small variations in equidistance in many cases are indeed non-significant. However, given the overwhelming data of

approximate equidistance, when expectation based on non-constant mutation rates would be much greater variations in distance, it is easy to infer that the real result here is equidistance (with minor variations from the mean) rather than non-equidistance with equidistance being coincidental. Indeed, if the equidistance result were not real, the constant clock idea would not have been invented in the first place (Kumar, 2005; Margoliash, 1963).

Some common practice such as the relative rate test has often interpreted small variations from exact equidistance to be significant (Avice, 1994; Li, 1997; Nei and Kumar, 2000). Many evolution biologists who perform such tests mistakenly consider the real phenomenon to be non-equidistance with equidistance being coincidental. But the relative rate test may not be appropriate in most cases because it does not consider sampling variations. It also does not consider the large differences in functional constraints on mutations in different kinds of species of different epigenetic complexity or organismal complexity (Huang, 2008b; Yang et al., 2003). Furthermore, it presupposes the truth of the gradual mutation model of speciation when it remains an open question whether genetic distance had always increased with time in the past history of life on Earth. The recent analysis of fossil organisms in fact shows that genetic distance had not always increased with time in the past (Green et al., 2008; Huang, 2008a).

To consider small differences in distance as being significant also makes it impossible to reconcile it with other contradicting facts. For example, the albumin protein of a specific bird individual is 47% identical to that of a specific human and 44% identical to that of a specific rat. Some evolution biologists have viewed such small differences to be statistically significant after performing the relative rate test (Nei and Kumar, 2000). This however contradicts the fact that a frog (*X. tropicalis*) albumin gene is 38% identical to human and 40% to rat. It is impossible for the rat lineage to have a faster mutation rate than humans when birds are the outgroup but a slower mutation rate than humans when frogs are the outgroup. If the faster mutation rate than humans with birds as the outgroup is real, the rate with frogs as the outgroup can only be faster and cannot possibly be slower or equal, since rats and humans do not have separate ancestors prior to the frog to bird transition. Therefore, the facts can only be explained by considering such small differences as insignificant variations of the equidistance result. Rats and humans are equidistant to birds as well as to frogs. All different mammals are equidistant to birds in the range of 43-47% identity in the albumin gene.

The genetic equidistance result merely shows the out-

come of evolution and says nothing about the actual mutation process during the past history of evolution. In contrast, the common interpretation or restatement of this result, i.e., the constant mutation rate or molecular clock, is all about the mutation process. So there is a clear distinction in meaning between the equidistance result and its common interpretation known as the 'constant mutation rate'. The equidistance result does not necessarily entail a constant mutation rate or any other ideas about the mutation process, while the constant mutation rate idea covers the equidistance result and much more and represents an over-interpretation of the actual result.

## Conclusions

The genetic equidistance result is arguably the most remarkable result of molecular evolution since it was completely unexpected from classical Neo-Darwinian evolution theory. This result and the biology behind the result have unfortunately remained obscure despite the past 45 years of research. The equidistance result could trigger many interpretations but the idea of constant mutation rate has become the most popular. However, there is no independent evidence for it other than the equidistance result that originally provoked it. It is merely a tautology. The observation of frequent violations of the constant mutation rate has misled many to automatically assume that there is no equidistance result in many cases. The study here establishes the fact that the equidistance result is extremely robust and universal that is independent of variation in mutation rates. The equidistance result shows the outcome of evolution but does not directly reveal any information about the actual mutation process in the past history of life on Earth. New ideas are needed to explain the equidistance result that must grant different mutation rates to different species and must be independently testable.

## Acknowledgments

This work was supported by the NIH (RO1 CA 105347). It has gone through a long and repeated submission and peer review process, and I thank the numerous reviewers for their comments.

## Methods

Protein sequences from a specific taxon were retrieved from the NCBI protein database. For example, to retrieve all protist *S. lemnae* protein or cDNA sequences, I did Search for *Lemane* on the NCBI home page (using the word *S. lemnae* to search the Protein database).

The exact nature of the genes (function type, reason for

study, and time or order of appearance in the Genbank) is independent of the equidistance result. Thus, while the availability of a gene sequence in the Genbank has specific reasons and hence is not strictly random, none of the reasons is in anyway linked to the equidistance result. Their availability in the Genbank is therefore effectively random as far as the equidistance result is concerned. The selection of 50 genes from the frog (*X. laevis*) protein database was by first retrieving a list of all frog proteins by doing a key word search using *laevis*, followed by selecting the first 50 informative proteins based on their numerical order on the list.

Homology comparisons were performed using BLASTP on the NCBI server. Percent nonidentity in protein sequence was used to measure genetic distance as originally used in the 1960s when the genetic equidistance result was first discovered. The equidistance result would not be affected in any way when percent nonidentity was converted into Poisson or Gamma distance.

## References

1. Avise JC (1994) Molecular markers, natural history and evolution. Springer: New York, NY.
2. Ayala FJ (1986) On the virtues and pitfalls of the molecular evolutionary clock. *J Hered* 77: 226-235.
3. Ayala FJ (1997) Vagaries of the molecular clock. *Proc Natl Acad Sci USA* 94: 7776-7783.
4. Ayala FJ (1999) Molecular clock mirages. *BioEssays* 21: 71-75.
5. Clarke B (1970) Darwinian evolution of proteins. *Science* 168: 1009-1011.
6. Doolittle RF, Blombaeck B (1964) Amino-Acid Sequence Investigations Of Fibrinopeptides From Various Mammals: Evolutionary Implications. *Nature* 202: 147-152.
7. Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. *Science* 155: 279-284.
8. Goodman M, Moore GW, Barnabas J, Matsuda G (1974) The phylogeny of human globin genes investigated by the maximum parsimony method. *J Mol Evol* 3: 1-48.
9. Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, et al. (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* 134: 416-426.
10. Ho SYW, Larson G (2006) Molecular clocks: when times are a-changin'. *Trends Genet* 22: 79-83.
11. Huang S (2008a) Ancient fossil specimens are genetically more distant to an outgroup than extant sister species are. *Riv Biol* 101: 93-108.
12. Huang S (2008b) Histone methylation and the initiation of cancer, *Cancer Epigenetics*. CRC Press: New York.
13. Jukes TH, Holmquist R (1972) Evolutionary clock: nonconstancy of rate in different species. *Science* 177: 530-532.
14. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, et al. (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447: 714-719.
15. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624-626.
16. Kimura M, Ohta T (1971) On the rate of molecular evolution. *J Mol Evol* 1: 1-17.
17. King JL, Jukes TH (1964) Non-Darwinian evolution. *Science* 164: 788-798.
18. Kumar S (2005) Molecular clocks: four decades of evolution. *Nat Rev Genet* 6: 654-662.
19. Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* 392: 917-920.
20. Laird CD, McConaughy BL, McCarthy BJ (1969) Rate of fixation of nucleotide substitutions in evolution. *Nature* 224: 149-154.
21. Langley CH, Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. *J Mol Evol* 3: 161-177.
22. Lewin R (2005) *Human Evolution*, 5th edn. Blackwell Publishing Ltd: Malden, MA 02148, USA.
23. Li WH (1997) *Molecular evolution*. Sinauer Associates: Sunderland, MA.
24. Margoliash E (1963) Primary structure and evolution of cytochrome c. *Proc Natl Acad Sci* 50: 672-679.
25. Nei M, Kumar S (2000) *Molecular evolution and phylogenetics*. Oxford University Press: New York.
26. Pilbeam D (1968) The earliest hominids. *Nature* 219:

- 1335-1338.
27. Powers DA (1991) Evolutionary genetics of fish. *Advances in Genetics* 29: 119-228.
28. Pulquerio MJ, Nichols RA (2007) Dates from the molecular clock: how wrong can we be? *Trends Ecol Evol* 22: 180-184.
29. Richmond RC (1970) Non-Darwinian evolution: a critique. *Nature* 225: 1025-1028.
30. Schwartz JH (1984) The evolutionary relationships of man and orang-utans. *Nature* 308: 501-505.
31. Schwartz JH (2005) *The Red Ape, Orangutans and Human Origins*. Westview Press: Cambridge, MA 02142, USA.
32. Simons EL (1961) The phyletic position of *Ramapithecus*. *Postilla* 57: 1-9.
33. Simons EL, Pilbeam DR (1965) Preliminary revision of the *Dryopithecinae* (Pongidae, Anthropoidea). *Folia Primatol (Basel)* 3: 81-152.
34. van Valen L (1974) Molecular evolution as predicted by natural selection. *J Mol Evol* 3: 89-101.
35. Wilson AC, Sarich VM (1969) A molecular time scale for human evolution. *Proc Natl Acad Sci USA* 63: 1088-1093.
36. Yang J, Lusk R, Li WH (2003) Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci USA* 100: 15661-15665.
37. Zuckerkandl E, Pauling L (1962) *Molecular disease, evolution, and genetic heterogeneity*, *Horizons in Biochemistry*. Academic Press: New York.