RC24689 (W0811-049) November 6, 2008
Other

IBM Research Report

Relationships between Molecular Clock Deviations, Interactions, and Selection in Human mtDNA Haplogroups

Daniel E. Platt
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Relationships between molecular clock deviations, interactions, and selection in human mtDNA haplogroups.

Daniel E. Platt

Abstract

Interactions between sites, selection and population effects, and other effects such as methylation of CpG groups all violate molecular clock assumptions, yet the molecular clock works surprisingly well. This study explores the relationship between these effects and the molecular clock, and seeks to understand the implications of these results in the context of recent studies of human mtDNA involving migration, environmental selection pressure, and other similar efforts. We present a maximum likelihood Poisson regression to the human mtDNA phylogenetic tree modeling a molecular clock, together with a similar computation of nonsynonymous to synonymous substitutions on each node, and their deviation from expectation determined from the entire phylogenetic tree. We show that the observed deviating nodes shows significant overlap between molecular clock counts and synonymous to nonsynonymous substitutions, primarily in leaf nodes and deep within the L-clades. This establishes connections between a mtDNA molecular clock model and prior studies also reporting deviations of nonsynonymous to synonymous substitution ratios between northern and temperate climes, but, at the finer level of analysis, we show that some of the groups of clades lumped by environment in

previous studies are actually heterogeneous in their deviations, tending not to support environmental selection. This indicates that simple models of variation in rates fails to capture underlying systematic interactions between sites as well as other events that promote mutations, and that Poisson regression can be applied to identify clades marked by such deviations.

Introduction

The molecular clock hypothesis represents an ideal against which deviations both reveal information and inject difficulties. Genetic forces that may promote deviations from the molecular clock[1] are a topic of increasing interest in recent research on the human geographic expansion.[2-9] The difficulties associated with deviations from the molecular clock on estimation of times of most recent common ancestors,[10; 11] on phylogeny construction (particularly across taxa), and with limited available data, have prompted the development of tools that allows local variation in the molecular clock at the cost of introducing more parameters and variability into the problem.[12]   However, the causes of violations of the molecular clock are also of significant interest, particularly because of what such violations may reveal about the biological selection and population processes that promoted these divergences, which has prompted several good review articles and many studies.[13-17]

Appendix A revisits the standard development of Markov processes in describing substitution processes, including interactions between sites, treatment of these variations as uncontrolled random variables, and variations due to purification by selection and interactions between sites. Instantaneous substitution rate matrices that emerge from that description form the basis of substitutions as a Poisson process in terms of actual substitution events. Selection and population effects, ranging in scale from heteroplasmy to population, inject variations in the relative timescales of selection and fixation, possibly producing differences in the rates observed in different parts of the phylogenetic tree. Other issues affecting deviations include the impact of correlated mutations which invalidates the additivity of subsitution rates at different sites. The effect of adding rates of correlated sites together is to inject an effective uncontrolled variation in the rates of the combined sites. This combines with other sources of variation, such as CpG effects. The induced variations plus variation among the various sites has been previously modeled with fair success by replacing the rates at each site by identical and independent Γ-distributions,[18; 19] essentially introducing only three parameters: α and β governing rate and shape, and the number of sites N. The sum of such variable rates over the roughly 15,000+ mtDNA non hypervariable sites that are typically observed would not be expected to show significant variation due to the central limit theorem, unless the variations in *rates* are correlated, or unless some process singularly promotes the rate of substitutions, such as CpG effects. Correlations in the *variations* of large numbers of rates may be expected in the presence of purifying selection, for example, which may result in the effective removal of significant numbers of substitutions from the population in older branches compared to leaves. Exploration of specific examples of known deviations from clock-like behavior against a Poisson-clock, as embodied by a maximum likelihood Poisson regression may help shed light on the relationship of those observed deviations and the underlying processes that produce them. Details of the computation by maximum likelihood Poisson regression as adapted here are presented in Appendix B.

Efforts to identify environmental selection pressure have focused on variations in climate[2; 4] as well as regionally localized haplogroups.[7-9] Enrichment of k_a / k_s ratios in younger haplogroups has been identified by several studies,[3; 11] arguing against climate-driven selection pressure.[3] While synonymous substitutions are seen to be subject to selection pressure for a number of possible reasons,[20] it is clear that the impact of selection on synonymous vs. nonsynonymous substitutions is different. A measure of a change in the ratio k_a / k_s therefore represents a test of differential selection pressure from one part of the phylogenetic tree to another.

Another review has also identified evidence of selection pressure, without such supportable evidence of climate-driven selection.[6] The review of substitution rates and Markov processes suggests some remarkable constraints on the type of processes that could produce differential rates of evolution. Specifically, these include correlated substitutions, as well as the necessity of correlations between the variations of substitution rates across sites. In this second group, interactions between selection and drift stands out as a possible source of rate variations dependent on locations within the phylogenetic tree.

This paper identifies a number of haplogroups that show some deviation from the best-fit molecular clock, and explores other measures of differential evolution, including nonsynonymous to synonymous substitution ratios, k_a / k_s, determined exhaustively for the entire phylogenetic tree. The method for determination of deviations from the molecular clock presented here echoes Sarich and Wilson’s approach, which compared differences between Poisson-distributed variables, determined by simulation to be normally distributed for Poisson counts $N \geq 20$, yielding χ^2 distributed variables.[21] A later study compared observed counts with those of counts placed by simulation on a phylogenetic tree, essentially implementing a Poisson distribution to measure probabilities branch-by-branch.[22] Ancestral states were inferred using a modified Sarich-Wilson algorithm,[23] consistent with the relative rates test of the molecular clock.

While likelihood ratio tests[24] ultimately grounded on the Sarich-Wilson algorithm,[23] and have found application in exploring the impact of selection, [3; 7; 8] the study presented here actually seeks to identify differences in processes in different lineages of the phylogenetic tree, rather than identifying deviations from a Poisson molecular clock. In this study, we seek to explore in greater detail what influences cause deviations from the Poisson molecular clock, and how this information can shed light on selection processes and interactions between sites' substitution that cause deviations from such a molecular clock. Therefore, we are computing a maximum likelihood best fit to a Poisson molecular clock, and exploring the deviations from such a clock for the effects of interactions and selection.

Materials and Methods

Alignments and SNP Counts

Data collected by Herrnstadt et al[7; 25] were employed in this study. This set of data has been thoroughly tested and validated. Following a reduced-median network analysis,[26; 27] it was pointed out by Bandelt[28] that there are a number of common sources of error, some of which were specifically identified in the Herrnstadt et al neighbor-joining study. Herrnstadt et al responded by acknowledging Bandelt's contribution,[25] identified and corrected those errors, as well as others, and made the data available. Since then, the data, 560 complete mtDNA sequences excluding D-Loop sites are available courtesy of MitoKor at http://mito546.securesites.net/science/560mtdnasrevision.php. 56 of these are L clades, which were republished, together with 37 new L-clade sequences, all with control region sites. This new data was published with the supplementary material online.[7] Two chimpanzees and bonobo sequences from http://www.genpat.uu.se/mtDB/[29] were employed to yield some outgroup information required to determine ancestral states of immediate children of the root.

These sequences were pair-wise aligned with the Revised Cambridge Reference Sequence (rCRS)[30] by applying the linear global alignment algorithm "stretcher"[31] implemented in Emboss[32] (http://emboss.sourceforge.net), and by ClustalW[33] and all SNPs identified across all of the sequences were indexed.

SNPs included deletions, transversions, transitions, and deletions relative to rCRS. However, insertions relative to rCRS were not included since associations between multiple insertions present ambiguities in comparisons between sequences bearing such SNPs. Deviations from the rCRS that represent insertions or deletions represent special problems in comparing two sequences[34; 35] with each other because alignments within multiple-site insertions or deletions are not consistently indexed in their alignment with each other. The most consistently identifiable mutations to process are nucleotide changes from the rCRS, excluding deletions or insertions. This implies blind spots that could be important to groups outside of the rCRS H haplogroup, as well as the possibility of loosing reliable resolution of mutations in regions involving insertions relative to rCRS. If the SNP is due to base calling or alignment errors, then the number of sequences in which such a SNP may appear would be expected to be low. For each alignment, the number of sequences supported by each SNP was determined. The relationship between support and putative error is explored.

These sequences' SNPs identified in the alignment were indexed according to rRNA, tRNA and coding segments, as well as D-LOOP and HVS-I and –II membership using the information in Anderson et al.[36] At each site, an index of all possible substitutions (A, C, G, T) against the RCRS, using the tRNA tables[36] to determine peptides for each specific possible nucleotide substitution in the coding segments.

Haplogroup Assignment

These sequences were assigned to haplogroups as outlined by the Genographic Project public participation markers.[37] More detailed information from the L clades, together with some finer detail in some other haplogroups was spliced into this tree. Inclusion of more detailed L clades appropriate to this study presents some difficulties. Nomenclature among L clades is not consistent. For example, L0a as defined by Kivisild et al[3] is marked as L1a by Torroni et al.[9] The Torroni study used L1a only as an outgroup. This study will follow the phylogeny described by Kivisild, which is most consistent with the tree on the mtDB website (http://www.genpat.uu.se/mtDB/)[29] The L2 clades are more consistently marked in the literature. The Mitomap phylogeny avoids some of these issues by simply not labeling the branches except by marker. Other phylogenies consulted include those of Bandelt[38] and Macaulay.[39]

The phylogenetic tree shows polytomy. This is problematical because methods that can be used to identify mutations and mutation counts, such as Sarich and Wilson's approach,[23] require a bifurcated tree where each node has a well defined sibling and outgroup. The tree's polytomies were reduced to bifurcations[40] with branch orders selected to roughly reflect observed mutation counts. Polytomy places B, R9, J, U, T, and R* at indistinguishable times. In general, rearrangement of bifurcations can be expected to modify the substitutions and counts, yielding some ambiguity in the effect of re-ordering bifurcations since the sibling and outgroups can change. However, comparisons of several re-orderings of bifurcations do not tend to resolve bifurcation order by count. Rearranging bifurcations, specifically placing J and T adjacent to each other for example, does not produce significantly different results, and very low counts are observed at the inserted bifurcation nodes regardless of the ordering of the inserted bifurcated branches. Following common convention, the 'x' in front of a branch means “exclusive of” or “complement of.” The original phlogenies employed here signified sequences bearing no markers for any other specific subclade with a “*”. The tree presented here retained these notations, marking them in Figure 1.

Inferring Ancestral States

The algorithm employed in this study is an adaptation of Sarich and Wilson’s approach.[23] In that approach, it is noted that the distance between two sibling nodes A and B is comprised of the sum of the distances between A and its parent P and B and its parent P. Therefore $D_{AB} = D_{AP} + D_{BP}$. However, the parent state is unknown, so distances D_{AP} and D_{BP} are undetermined, and are, in fact, the quantities we would like to determine. However, D_{AB} can be measured. D_{AP} and D_{BP} may be determined by considering the sibling O of P as a nearest out-group, which share a parent Q. Then $D_{AO} = D_{AP} + D_{PQ} + D_{QO}$ and $D_{BO} = D_{BP} + D_{PQ} + D_{QO}$. Here again, D_{AO} and D_{BO} can be measured. From these relationships, it follows that $D_{AP} = (D_{AO} - D_{BO} + D_{AB})/2$, and $D_{BP} = (D_{BO} - D_{AO} + D_{AB})/2$. The intermediate distances involving Q do not contribute.

In this study, we note that the above argument may be applied locus by locus to determine the ancestral states of individual substitutions. In that case, differences between siblings at a specific locus indicate that a substitution that occurred on one or the other branch. Comparison with the out-group determines which branch experienced the mutation. There is a complication in that the sequences associated with each node include the union of each of the node’s offspring. Given the presence of homoplasy and noise, there may be some variability, where differences are noted between siblings only for some sequences.

Determination of the placement of mutations proceeded as follows. For all loci and nucleotide combinations, each node was tested for placement proceeding from root to leaf. If the mutation was placed in a node, tests for that substitution did not continue into the daughter nodes. A substitution was identified as a mutation within a node if it appeared in at least one sequence within the node, in less than 5% of sibling and out-group, and in more than 0.5% of both daughter nodes.

Mutations are determined for coding and noncoding regions. Hypervariable regions are excluded.

It has been noted that the Fitch algorithm is capable of identifying ambiguities in a maximum parsimony estimate of the number of mutations given the possibility of multiple mutations along a lineage.[41] In this sense, it is formally the best algorithm, except that the modified relative-rates approach easily allows for the consideration of multiple sequences per node employing thresholds to screen and/or allow for possible noise in the data set. Adaption of the Fitch algorithm requires many randomly selected sequences to obtain a similar combinatorial picture of the number and reliability of placement of mutations.[10] The differences between the Fitch algorithm and the Sarich and Wilson approach will show up in the presence of homoplasy. However, the probability of finding such in a data set given so few mutations along a human mtDNA lineage to the most recent common ancestor is small. Both relative rates and the Fitch algorithm are likely to suffer from noise. Therefore in this exploration of the human mitochondrial tree, the Sarich and Wilson approach was deemed adequate.

The number of mutations m_{ij} associated with *each sequence* j in node i were all included in the Poisson regression as specified in Appendix B. There was some variation between the number of mutations within each node. The averages of the number of mutations within each node is labeled $\overline{m}$, and may be fractional. The standard deviation is labeled σ_m. The contribution of a mutation to the average $\overline{m}$ is effectively the fraction of sequences bearing that mutation within the node, which are assumed to have accumulated during the time from the parent node down that branch to the node in question.

The maximum likelihood estimate of the Poisson parameter λ for each node is obtained. This parameter corresponds to the total expected mutation rate r times the time t of the branch leading to the node. This implies that $\lambda = rt$. Since rates are essentially assumed to behave as sums of large numbers of Γ-distributed variables, with times of sibling *clades* constrained to be equal, a maximum likelihood estimate of the λ's provides a measure of a most compatible molecular clock consistent with the data. The curvature of the maximum likelihood function yields both measures of variability σ_λ expected for the estimate of λ, together with correlations between the various nodes' λs.

A Poisson distributed variable m is expected to have a mean and variance λ. The probability of seeing $m \overset{\leq}{\underset{\geq}{-}} M$ is then $\sum_{m \overset{\leq}{\underset{\geq}{-}} M} P(m;\lambda) = \sum_{m \overset{\leq}{\underset{\geq}{-}} M} \frac{\lambda^m}{m!} e^{-m}$, where the upper(lower) is selected according to whether $M \overset{\leq}{\underset{\geq}{-}} \lambda$, respectively.

There is a challenge in that this would represent the probability of finding an individual sequence with that much deviation, not the average for the haplogroup node. A difficulty exists for computing a probability for such an aggregate of sequences in that the sequence substitutions tend to be strongly correlated in the haplogroup (the Hg H haplogroup is distinct in having a wide range of markers distributed among its large number of sequences, yet each sequence shows an average of around two substitutions, indicating very wide diversity but consistently shallow time).

Silent and Nonsynonymous Substitutions

Upon identification of each mutation, the peptide associated with each mutation within each coding region is identified. This is compared to all the peptide assignments at that site in the sibling group sequences. The *fraction* of synonymous and nonsynonymous substitutions are accumulated in the node, yielding measures of k_s and k_a.

The ratio k_a/k_s is computed for each node, and a global ratio K_a/K_s is obtained. These counts are compared to a binomial distribution where the probability $\sum_{x \underset{\leq}{\overset{\geq}{}} n_a} P(x;p,N) = \sum_{x \underset{\leq}{\overset{\geq}{}} n_a} \binom{N}{x} p^x (1-p)^{N-x}$ of observing a count equal to or more/less than that expected for the global count $p = K_a/(K_a + K_s)$ is computed and tabulated. If $k_a/k_s \underset{\leq}{\overset{\geq}{}} K_a/K_s$, then the sum for $x \underset{\geq}{\overset{\leq}{}} n_a$ is appropriate, respectively. Since the process being modeled by the binomial does not draw sequences from a finite collection suggesting sampling without replacement, which would be better described by a Fisher exact test, but is rather sampling a process that produces substitutions randomly among the coding regions and then excludes some according to deleterious effects, a binomial as a null test is deemed more appropriate.

Results

Haplogroup Assignments and Ancestral States

The alignments of the Herrnstadt set and subsequent SNP identification produced 1779 SNPs. These were retained. The identification of substitutions in the L clades corresponded to those reported elsewhere.[7; 8; 27]

Determination of mutations – and simultaneously, determination of ancestral states – shows a number of very clean assignments, where no examples appear in either sibling or in out-group. Some show evident homoplasy, such as 3666A, which appears in L0/L1*, with none in sibling and 3 out of 60 in outgroup, in Hg D, with no examples in sibling, and 2 out of 520 in its outgroup, and in H, with zero examples in sibling or outgroup. While such examples are evident candidates for interaction, any single substitution could have been promoted by an interaction. Further, interactions can inhibit as well as promote transitions. Hg H is unusual in the number of substitutions that appear in only one or two sequences out of the 209 assigned to H, but do not appear in any siblings or outgroups. A fair number of these are homoplasy, appearing in other haplogroups as well. Examples include 2098A in Hg H and K, 2352C in H and L3*, 3316A in H, T and xX, 3338C in L2a, H, and B, etc.

Maximum Likelihood Poisson Regression

The Poisson regression shows the required constraints (i.e. $\lambda_V = \lambda_H + \lambda_{xV}$). Larger deviations, many of which are significant, though non are highly significant, are identified at Hgs L0a, L2b, xL2a*, D, B, U, K, U*, xR9, xW, and X, many of which have been identified in prior studies.[3; 4; 7-9] All are leaf nodes except for xR9 and xW, though leave nodes associated with these are balanced with the local Poisson regression estimates. Hg B appears significant, but the variation σ_m among sequences within Hg B is also large, so this result is not such a clear result. Similarly for Hgs K, U*, and xW. Only a fraction of sequences register any substititons in Hg U.

Haplogroup H is unusual. It has a large number of sequences, and a large number of substitutions in the group. However, the number of substitutions per sequence is lower than expectation. While the probability of seeing only 2 or less given an expected number of substitutions of 3.34 is 0.35, the largely independent character of the sequences and their substitutions sampled appears to indicate some unusual character in or around this clade.

Coding Nonsynonymous vs. Synonymous ratios

Haplogroups showing larger deviations are Hgs L0a, L2, L2b, xL2b, D, N, xR0, xR9, and L3*. All are leaves except for xR0, and xR9. Of those that show deviations, only N, xR0 and L3* do not show significant variation from the molecular clock.

The estimated $K_a/K_s = 0.5885$ obtained here is significantly larger than the ratio $k_a/k_s = 0.198 \pm 0.054$ reported by Hasegawa et al[42] for human and other species.

Prior studies seeking to identified regional groups as groups of clades, such as Hg's C and D as representative of northern groups.[2; 4] Hgs A and B do not appear in their list of deviants. However, haplogroup-by-haplogroup analysis presented here shows a significantly high k_a/k_s for D, but not for C. Hgs A and B also are not significant in deviation of k_a/k_s. Deviations from the mean clock show up also for Hg D, but not A, B, or C. Among those interior nodes showing both deviations from the mean clock and with significant k_a/k_s deviations, namely xR0 and xR9.

Summary and Conclusions

The preliminary results quoted here are presented in Appendix A, together with detailed references for cited effects. Markov models emerge naturally in describing any discrete state transitions in continuous time, such as nucleotide substitutions. As such, transition rates are defined by, and only have meaning in the context of, Markov models. The matrix of transition rates determines the stochastic behavior of such models.

A number of simplified transition matrices have been studied to explore the impact of increasing complexities. Perhaps the most general that guarantees convergence to a stable distribution is the General Time Reversible model that satisfies detailed balance given constant rates. Even in a state of such an equilibrium distribution, transitions would be expected to accumulate, but the overall distribution of nucleotides would generally fluctuate about that stable solution. It is important to note that the human mtDNA genome is far from such equilibrium since the number of substitutions accumulated since the most recent common ancestor is somewhat less than 40, which is far less than the multiple number of transitions at the more than 15,000 bases required to equilibrate achieved through sampling multiple transitions at each site. The overall probability distribution would not be expected to change significantly over such a short time scale and can be treated as effectively constant.

Interactions between sites are expected to be present. First, there is selection, which can produce correlated selection pressure for one site depending on the value of another site. The simplest example might be where a substitution may be silent for one triplet, but nonsynonymous for another transition: $ACU \rightarrow ACA$ would be silent, but $AGU \rightarrow AGA$ would not be silent. More complex interactions, where one substitution affects the selection pressure applied to another substitution, for example, can produce correlated mutations. Measurements of the rate matrices even just for pairs of interacting sites would require multiple events involving the combinations of pairs, which implies a time scale much longer than that afforded by coalescence in the human mtDNA genome. Surprisingly, a simple Markov model, even in the presence of interactions, can describe the mutation rate measured at a specific site. However, substitution rates of two interacting sites are not additive.

Treating rates as additive injects an effective variability in the rates. Other effects, such as deanimation of methylated C's (CpG hotspots), also inject correlated variability into the rates. Such variation, together with the variation across sites, has been modeled successfully with Γ-distributions resulting in an expected negative binomial distribution for the accumulation of transition events, which results in an expected and observed overdispersion within fairly small regions within the mtDNA genome. By the central limit theorem, such variability would be expected to produce only small variations when the rates of 15,000+ sites are combined.

The validity of such an approximating assumption would hold most effectively if sampled over a heterogeneous collection of substitutions, where all the members of the samples were mixed together. However, sampling on a phylogenetic tree essentially isolates samples according to substitutions, which may affect correlated rates of other substations differentially between samples. Further, since the branches are time-ordered, time-dependent sampling effects, such as purifying selection, will tend to show higher rates for those substitutions identified nearer the leaves than those that survived in interior nodes. Purifying selection, by itself, could account for gradients in substitution rates among substitutions marking older haplogroups to substitutions that occurred within younger haplogroups. Population size changes without selection will not affect substitution rates.[43] Therefore, population size interactions with substitution rates must be enabled by selection. This interaction is realized in the relative time it may take for a deleterious substitution to become fixed in the population vs. the time it takes for selection to remove the deleterious substitution. These considerations imply that a Poisson regression would be expected to identify violations due to interactions and purifying selection when applied to a phylogenetic tree.

Candidates for selection driving deviations from the molecular clock would be identified by deviations from the molecular clock accompanied by unusual k_a/k_s ratios. Those that may mark population size-selection interactions will be those that are identified in older haplogroups interior to the tree.

Prior studies seeking to identify environmental selection effects have tended to lump multiple haplogroups into regional groups.[2; 4] For example, Northern Asians would include the C and D clades, but not A and B. We show a significantly high k_a/k_s for D but not C. Using A and B for comparison, hgs A and B also are not significant in deviation of k_a/k_s. Deviations from the mean clock show up also for Hg D, but not A, B, or C. It is difficult to conclude from this that northern environments can account for the apparent observed selection pressure.

Hgs xR0 and xR9 are interior nodes with both deviations from the mean clock and with significant k_a/k_s deviations. These groups, are branch points of a number of sequences representing populations that moved through a diverse range of environments. Given these considerations, it would appear that environmental selection is not strongly supported by k_a/k_s of multiple haplogroups associated with differences in northern vs. southern climes.

Alternatively, almost all of the deviants from both the molecular clock and the k_a/k_s ratio appear in or near the leaves, consistent with results of other studies,[6; 11] and also consistent with the identification of “private substitutions” by Howell et al.[7; 8] Further, there is significant overlap between lists of groups showing deviations from the molecular clock and k_a/k_s. This would be consistent with a picture of deleterious nonsynonymous substitutions being removed from their populations by selection over time, ultimately achieving some equilibrium in deeper nodes.

It is also notable that while many deviations are associated with leaves, this is not universal for all leaves. Only some leaves show deviations. It would be expected that the mutation rate matrix would produce a binomially distributed spectrum of nonsynonymous substitutions insensitive to the deleterious character of the mutation except for those that are immediately lethal. Selection will act to remove many of these over time. At the 0.1 level that was chosen as a threshold, it would be expected that between 4 and 6 leaf nodes would have been in this list, and a similar number from deeper in the tree. The number of leaves observed exceeding the expected nonsynonymous to synonymous ratio was about expectation, while the number exceeding the molecular clock expectation was rather larger than this number. The number from non-leaf nodes was well under expectation.

Significantly, those haplogroups deeper in the tree that *were* identified as deviant are also closely associated with each other, indicating systematic deviation in those clades. Specifically, xR0 and xR9 were introduced to resolve polotomy ambiguities. SNPs associated with them are placed in the interior due to the K/U* and T/R* splits. Therefore, these k_a/k_s ratios and notable (though not significant) deviations from the mean clock may indicate unusual levels of nonsynonymous substitutions becoming fixed during some period of small population sizes in the clades.

While the L clades that have shown deviations are technically leaf nodes according to this phylogeny, they are very deep clades, showing 20’s of substitutions per sequence. This suggests that they should be dominated by older substitutions. These therefore do not appear to show purifying selection on the time scale suggested by the non-L clades, which suggests some effect other than purifying generating such large numbers of deviations compared to other clades. Further, the L clades show a number of subclades with fairly significant deviations in synonymous vs. nonsynonymous substitutions.

Among all of the cited results, the most striking deviations are among the L0-L2 clades. These deviations are apparent even at the course-grained resolution that this study accomplished. A much more detailed study[44] of the L clades has suggested that the early expansion of H. sapiens through Africa was characterized by a long-term isolation of numbers of very small matrilineal groups. Numbers of groups with small effective population sizes would be consistent with higher rates of fixed nonsynonymous substitition rates than in other populations. Numbers of substitutions from the present to clade MRCAs are consistent with those obtained here assuming the same 5138 yrs/substitution[2] employed in that study.

While nonsynonymous-synonymous ratios may provide insight into this situation, they are by no means exhaustive of deleterious substitutions, considering tRNA's,[5] and rRNA impacts. This is highlighted by xL2a*, where the number of substations is larger than expected, yet the nonsynonymous to synonymous substitution ratio is close to the phylogenetic average. Further, synonymous substitutions are not without selection pressure.[20]

The human mtDNA represents a unique laboratory to explore the relationship between processes that lead to molecular clock-like behavior, as well as violations of that behavior due to the wide range of sampling of the human species, and due to the very short –termed snapshot since the most recent common ancestor. At the same time, the short time scale limits the presence of homoplasy that would provide more insight into the effects of interactions, allowing a more direct measurement of effects that are partly obscured, to the benefit of a molecular clock by the number of sites available in the mtDNA genome. The observation of violations of the molecular clock indicate clearly that a modelling interactions between substitutions and of selection pressure by a simple independent and identically Γ-distributed variable rates fails to capture the observed behavior. This implies that, while the time scale to most recent common ancestor is too short to accurately measure interactions and selection effects, their effect is being detected through application of a simple Poisson-regression. This suggests that there may be utility in forging stronger connections to underlying processes of interaction and selection to characterize what role each of the specific substitutions involved in the deviation of the L, C and D clades.

Acknowledgements – The author has enjoyed the guidance of a number of individuals. Foremost was that of Chris Tyler-Smith, who gave many helpful comments throughout the course of the development of this study, beginning with the problems involving the quality of the data, identifying sources of better quality data, pushing for a deeper appreciation for the impacts of various correlating processes affecting mutation, and giving the insight of an editor's eyes. Thanks also go to Neil Howell for a careful reading and questions that lead to a realization and deeper understanding of the processes that can be identified throughout our phylogenetic tree. Thanks also go to Pierre Zalloua for suggesting deeper consideration of biological effects. Our deepest thanks. Next, thanks goes to Saharon Rosset, questioning the validity of a Poisson distribution for substitution events, which pushed the author to question what should be expected from such substitution events, and what conditions bound the emergence of Poisson-like behavior. Thanks also goes to Doron Behar for insight into the difficulties of noise in the published full sequence mtDNA data, specifically involving apparent homoplasy in the HV lineages. Ultimately, this project began as a part of IBM's contribution to the

Genographic Project of the National Geographic Society, as a way to compare phylogenetic trees produced by several methods, including those of Gabriela Alexe and Gyan Bhanot. We thank them and to all the participants in those early discussions at IBM.

References

1. Zukerkandl E, Pauling L (1962) In: Kasha M, Pullman B (eds) Horizons in Biochemistry. Academic Press, New York

2. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100:171-176

3. Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373-387

4. Ingman M, Gyllensten U (2007) Rate variation between mitochondrial domains and adaptive evolution in humans. Hum Mol Genet 16:2281-2287

5. Xing Y, Lee C (2006) Can RNA selection pressure distort the measurement of Ka/Ks? Gene 370:1-5

6. Elson JL, Turnbull DM, Howell N (2004) Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. Am J Hum Genet 74:229-238

7. Howell N, Elson JL, Turnbull DM, Herrnstadt C (2004) African Haplogroup L mtDNA sequences show violations of clock-like evolution. Mol Biol Evol 21:1843-1854

8. Howell N, Elson JL, Howell C, Turnbull DM (2007) Relative rates of evolution in the coding and control regions of African mtDNAs. Mol Biol Evol 24:2213-2221

9. Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69:1348-1356

10. Rosset S (2007) Efficient inference on known phylogenetic trees using Poisson regression. Bioinformatics 23:e142-147

11. Endicott P, Ho SY (2008) A Bayesian evaluation of human mitochondrial substitution rates. Am J Hum Genet 82:895-902

12. Felsenstein J (2003) Inferring Phylogenies. Sinauer Associates, Sunderland, MA

13. Bromham L, Penny D (2003) The modern molecular clock. Nat Rev Genet 4:216-224

14. Ho SY, Larson G (2006) Molecular clocks: when times are a-changin'. Trends Genet 22:79-83

15. Hedges SB, Kumar S (2003) Genomic Clocks and Evolutionary Timescales. Trends in Genetics 19:200-206

16. Welch JJ, Bromham L (2005) Molecular dating when rates vary. Trends Ecol Evol 20:320-327

17. Rutschmann F (2006) Molecular Dating of Phylogenetic Trees: A Brief Review of Current Methods that Estimate Divergence Times. Diversity Distrib 12:35-48

18. Ota T, Nei M (1994) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 38:642-643

19. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-314

20. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98-108

21. Wu CI, Li WH (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A 82:1741-1745

22. Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823-833

23. Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144-1147

24. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376

25. Herrnstadt C, Preston G, Howell N (2003) Errors, phantoms and otherwise, in human mtDNA sequences. Am J Hum Genet 72:1585-1586

26. Bandelt HJ, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743-753

27. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70:1152-1171

28. Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71:1150-1160

29. Ingman M, Gyllensten U (2006) mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34:D749-751

30. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

31. Meyers EW, Miller W (1988) Optimal Alignments in Linear Space. Computer Appliations in the Biosciences 4:11-17

32. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277

33. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

34. Wilson MR, Allard MW, Monson K, Miller KW, Budowle B (2002) Recommendations for consistent treatment of length variants in the human mitochondrial DNA control region. Forensic Sci Int 129:35-42

35. Wilson MR, Allard MW, Monson K, Miller KW, Budowle B (2002) Further discussion of the consistent treatment of length variants in the human mitochondrial DNA control region. For Sci Comm 4(4)

36. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457-465

37. Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, Quintana-Murci L, Tyler-Smith C, Wells RS (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3:e104

38. Bandelt HJ, Kong QP, Richards M, Macaulay V (2006) Estimation of mutation rates and coalescence times: some caveats. In: Bandelt HJ, Macaulay V, Richards M (eds) Human Mitochondrial DNA and the Evolution of Homo Sapiens. Springer-Verlag, Berlin, pp 47-90

39. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034-1036

40. Ingman M, Gyllensten U (2001) Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J Hered 92:454-461

41. Fitch WM (1971) Toward defining the course of evolution: defining the minimum change for a specific tree topology. Systematic Zoology 20:406-416

42. Hasegawa M, Cao Y, Yang Z (1998) Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. Mol Biol Evol 15:1499-1505

43. Kimura M (1983) The Neutral Theory of Evolution. Cambridge University Press, New York

44. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S (2008) The dawn of human matrilineal diversity. Am J Hum Genet 82:1130-1140

Figures and Tables

Figure 1. Phylogenetic tree and haplogroup markers used for classification of sequences.

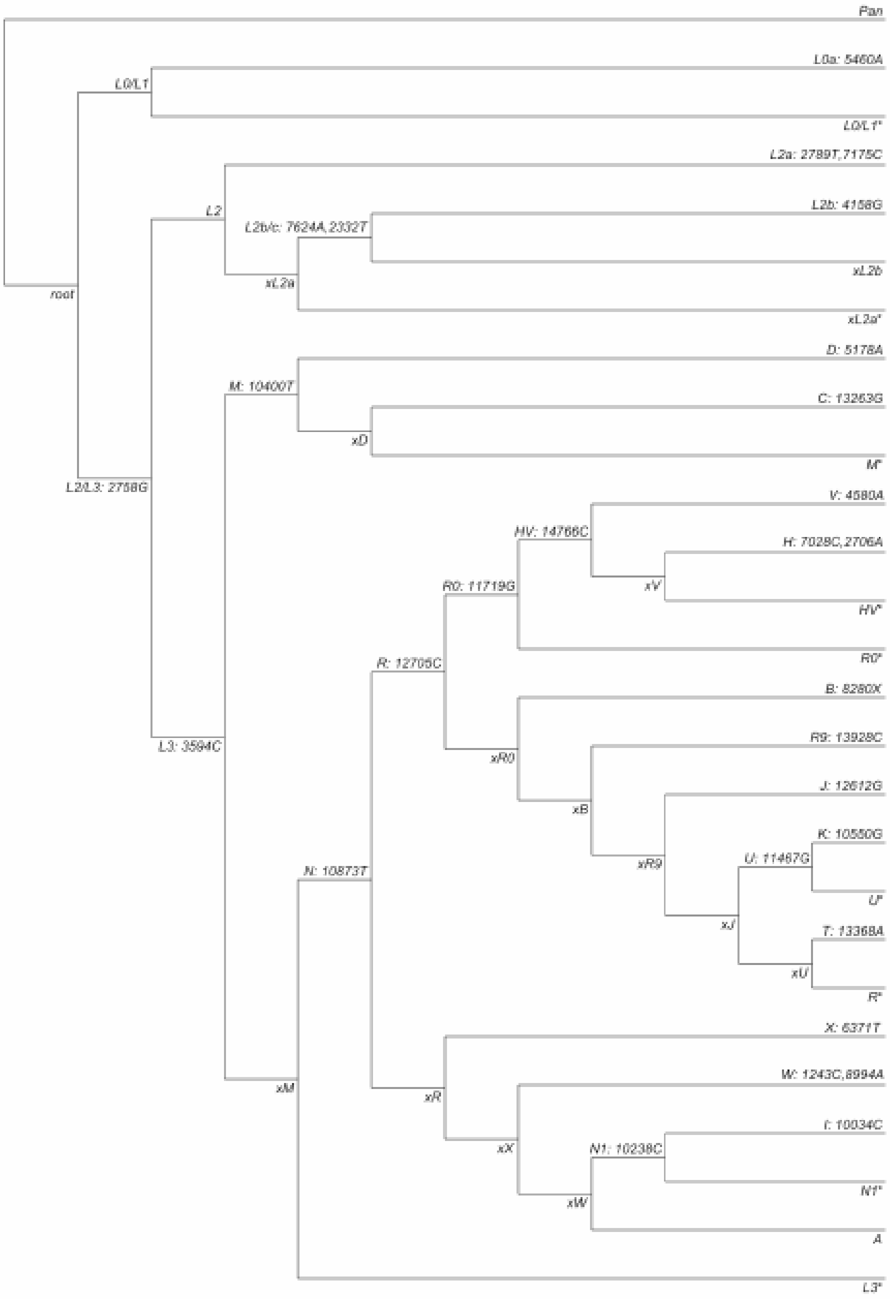
Pan
L0a: 5460A
L0/L1
L0/L1*
L2a: 2789T,7175C
L2
L2b: 4158G
L2b/c: 7624A,2332T
xL2a
xL2b
root
xL2a*
D: 5178A
M: 10400T
C: 13263G
xD
M*
L2/L3: 2758G
V: 4580A
HV: 14766C
H: 7028C,2706A
xV
R0: 11719G
HV*
R0*
R: 12705C
B: 8280X
R9: 13928C
xR0
xB
J: 12612G
K: 10550G
U: 11467G
xR9
L3: 3594C
N: 10873T
U*
xJ
T: 13368A
xU
R*
X: 6371T
W: 1243C,8994A
xM
xR
I: 10034C
xX
N1: 10238C
N1*
xW
A
L3*

Table 1. Estimated times (T), Poisson parameters (λ),number of mutations (m) at each node, and number of sequences N for the full phylogenetic tree from the Herrnstadt set.

Haplogroup	λ	σ_λ	$\overline{m}$	σ_m	L-R tail	$\sum_{m {\le \lceil m \rceil \atop \ge \lfloor m \rfloor}} P(m;\lambda)$
root	0.0000	0.0000	0.0000	0.0000	-	------
L0/L1	2.4812	0.2178	3.0000	0.0000	>	0.2384
L0a	25.8938	0.2891	33.7143	2.8140	>	0.0720
L0/L1*	25.8938	0.2891	30.7812	3.9901	>	0.1808
L2/L3	7.4645	0.1105	7.3645	0.6745	<	0.6668
L2	6.1340	0.2704	6.0159	0.2813	<	0.7253
L2a	14.7765	0.2946	12.6042	2.4895	<	0.3849
xL2a	0.0000	0.0000	0.0000	0.0000	-	------

Haplogroup	λ	σ_λ	$\overline{m}$	σ_m	L-R tail	$\sum_{m \substack{\le \lceil m \rceil \\ \ge \lfloor m \rfloor}} P(m;\lambda)$
L2b/c	3.0225	0.3664	4.1539	1.7908	>	0.1885
L2b	11.7541	0.4307	19.2500	1.0897	>	0.0176
xL2b	11.7541	0.4307	11.2000	4.7497	<	0.6041
xL2a*	14.7765	0.2946	22.0000	0.0000	>	0.0284
L3	6.6431	0.1084	6.5586	0.8104	<	0.6517
M	3.0913	0.3929	2.0000	0.0000	<	0.4031
D	11.1762	0.4097	5.8889	0.7370	<	0.0717
xD	5.3811	0.5300	3.8235	2.0069	<	0.3763
C	5.7950	0.5358	4.1667	0.9860	<	0.4791
M*	5.7950	0.5358	4.0000	0.8944	<	0.3134
xM	1.4266	0.0519	1.4327	0.8962	>	0.4173
N	5.8766	0.1049	6.0865	1.1432	>	0.3739
R	1.3210	0.0543	1.3606	1.1392	>	0.3806
R0	2.1845	0.1281	1.0000	0.0000	<	0.3584
HV	0.0000	0.0000	0.0000	0.0000	-	------

Haplogroup	λ	σ_λ	$\overline{m}$	σ_m	L-R tail	$\sum_{m \substack{\leq \lceil m \rceil \\ \geq \lfloor m \rfloor}} P(m;\lambda)$
V	3.4588	0.1374	2.3750	0.4841	<	0.5455
xV	0.1116	0.0359	0.0502	0.2184	<	0.9942
H	3.3472	0.1376	1.5455	1.4307	<	0.3500
HV*	3.3472	0.1376	0.7000	0.9000	<	0.1529
R0*	3.4588	0.1374	1.0000	0.0000	<	0.1403
xR0	3.4187	0.0894	5.8936	3.6040	>	0.1318
B	2.2246	0.0753	5.1579	3.0134	>	0.0261
xB	0.3142	0.0322	0.5207	1.2922	>	0.2697
R9	1.9104	0.0715	1.5000	0.5000	<	0.7009
xR9	0.0862	0.0171	0.1437	0.3834	>	0.0826
J	1.8242	0.0703	1.4688	1.1452	<	0.7241
xJ	0.1662	0.0247	0.3111	0.6718	>	0.1531
U	0.0438	0.0141	0.1047	0.3061	>	0.0428
K	1.6143	0.0678	3.2444	1.8638	>	0.0808
U*	1.6143	0.0678	4.5366	3.4152	>	0.0245

Haplogroup	λ	σ_λ	$\overline{m}$	σ_m	L-R tail	$\sum_{m \substack{\le \lceil m \rceil \\ \ge \lfloor m \rfloor}} P(m;\lambda)$
xU	0.0000	0.0000	0.0000	0.0000	-	------
T	1.6580	0.0682	1.7381	1.7870	>	0.4936
R*	1.6580	0.0682	0.7143	0.8806	<	0.5064
xR	0.0000	0.0000	0.0000	0.0000	-	------
X	6.9643	0.1134	2.7273	0.8624	<	0.0836
xX	0.3955	0.0795	0.4894	0.5408	>	0.3267
W	6.5688	0.1331	8.6250	1.4087	>	0.2167
xW	0.1049	0.0461	0.1282	0.3343	>	0.0996
N1	2.9580	0.2995	3.8571	0.5151	>	0.3434
I	3.5058	0.3022	5.0000	2.2361	>	0.1432
N1*	3.5058	0.3022	2.0000	1.0000	<	0.3198
A	6.4639	0.1388	7.6000	1.0583	>	0.3220
L3*	12.8409	0.1473	8.7391	2.8775	<	0.1766

Table 2. $K_a / K_s = 0.5885$ for the entire population, so $p = 0.62953$. This table shows nonsynonymous to synonymous ratios for all nodes across all protein coding regions. Probabilities are listed as computed against a null hypothesis of binomial sampling from the entire population, where the probability represents the chances of seeing that many substitutions or more/less (depending on whether the ratio is larger or small than expected) by chance. N is the number of coding region substitutions.

Haplogroup	#	k_a / k_s	n_a	N	L-R	$\sum_{x \substack{\geq \\ \leq} n_a} P(x;p,N)$
	Sequences				Tail	
root	648	-------		0	-	-------
L0/L1	39	1.0000	1	2	>	0.6037
L0a	7	0.3333	30	40	<	0.0760
L0/L1*	32	0.6338	71	116	>	0.3819
L2/L3	609	0.6667	6	10	>	0.5413
L2	63	0.0000	6	6	<	0.0622
L2a	48	0.6216	37	60	>	0.4666
xL2a	15	-------		0	-	-------

Haplogroup	#	k_a/k_s	n_a	N	L-R	$\sum_{x \substack{\geq \\ \leq} n_a} P(x;p,N)$
	Sequences				Tail	
L2b/c	13	0.5000	2	3	<	0.6899
L2b	8	0.4706	17	25	<	0.3828
xL2b	5	1.3750	8	19	>	0.0523
xL2a*	2	0.3077	13	17	<	0.1847
L3	546	0.3846	13	18	<	0.2900
M	26	1.0000	1	2	>	0.6037
D	9	2.3333	3	10	>	0.0359
xD	17	0.6667	3	5	>	0.6102
C	12	1.0000	5	10	>	0.2951
M*	5	0.6000	10	16	>	0.5785
xM	520	0.2727	11	14	<	0.1765
N	474	1.5714	7	18	>	0.0328
R	416	1.0000	8	16	>	0.2060
R0	228	0.0000	1	1	<	0.6295
HV	227	-------		0	-	-------

Haplogroup	#	k_a/k_s	n_a	N	L-R	$\sum_{x \gtreqless n_a} P(x;p,N)$
	Sequences				Tail	
V	8	0.4966	2	3	<	0.2495
xV	219	0.0000	1	1	<	0.6295
H	209	0.6667	81	135	>	0.2658
HV*	10	1.0000	2	4	>	0.4732
R0*	1	-------		0	-	-------
xR0	188	0.2914	24	31	<	0.0279
B	19	0.9000	10	19	>	0.2409
xB	169	2.0000	1	3	>	0.3101
R9	2	1.0000	1	2	>	0.6037
xR9	167	4.0000	1	5	>	0.0663
J	32	0.8333	12	22	>	0.2720
xJ	135	0.5000	8	12	<	0.5235
U	86	-------	0	1	>	0.3705
K	45	0.5789	19	30	<	0.5642
U*	41	0.5366	40	63	<	0.4176

Haplogroup	#	k_a/k_s	n_a	N	L-R	$\sum_{x \substack{\geq \\ \leq} n_a} P(x;p,N)$
	Sequences				Tail	
xU	49	-------		0	-	-------
T	42	0.6250	16	26	>	0.5143
R*	7	0.5000	2	3	<	0.6899
xR	58	-------		0	-	-------
X	11	0.5216	6	9	<	0.5572
xX	47	1.0000	1	2	>	0.6037
W	8	0.3077	13	17	<	0.1847
xW	39	0.0000	1	1	<	0.6295
N1	14	1.0000	2	4	>	0.4732
I	12	0.6000	10	16	>	0.5785
N1*	2	0.5000	2	3	<	0.6899
A	25	0.4706	17	25	<	0.3828
L3*	46	0.4208	57	81	<	0.0759

Appendix A

Introduction

The purpose of this section is to explore the impact of interactions between transitions at different sites, selection and the Poisson molecular clock.

The molecular clock hypothesis[1] depends on the notion that molecular and population processes producing nucleotide substitutions in a population or across any number of populations will operate with the same substitution rates throughout the phylogenetic tree. Observed violations of a molecular clock[2] should therefore involve mechanisms relating differences in rates to differences in environment, genetics, or other elements that would cause deviations.[3-6]

Among these are population effects[7-9] including those due to enhanced chance for a mutation, weakly deleterious or not, to survive in smaller effective populations,[10] variations in generation time, geographical environmental effects that differentially impact selection in the phylogenetic tree depending on where different mutations emerge,[11-13] and geometric impacts, such as protein DNA and RNA folding, that impose selection-based correlations between mutations at different sites in the genetic sequence. Yet, founding events and bottlenecks are characterized by a loss of diversity[14] in the population; effect on evolutionary rate (rate that substitutions are accumulated in each lineage) is not so clear. This section primarily explores the effects of assumptions downstream from the action of selection on the formulation of substitution rates, and effective sources of variation in rates on an overall molecular clock.

Conditions leading to Markov processes as descriptions of substitutions in mtDNA.

Eukaryotic cells, including the human germ line, contain multiple mitochondria. Mitochondria are small organelles carrying their own unique genetic material arranged circularly as in prokaryotes, with roughly 16kb, containing tRNA, rRNA, and gene coding segments. Each mitochondrion may possess two to ten copies of mtDNA. Their tRNAs are distinct from other known organisms, but shared among all mitochondria-bearing eukaryotic cells, suggesting a pre-eukaryotic divergence. The closest known modern prokaryote that resembles the mitochondrion is Rickettsia prowazekii.[15] The enzymes coded by the mtDNA are far from sufficient to support an independent organism, only supporting ATPase plus the electron transport chain. ATPase catalyses the production of ATP from ADP in the citric acid cycle. This requires the passage of a bare proton (H^+ ion) through the core of the transmembrane ATPase, which provides the energy required to release the ATP from the ATPase after the ATP is assembled but still bound to the ATPase. The energy comes from the gradient of the transmembrane chemical potential generated by pumping of protons across the membrane powered by the electron transport chain reactions.[16; 17] While such transmembrane potentials clearly drive the reaction, and it is also clear that thermodynamics limits the availability of free energy, definition of a continuum chemical potential is problematical.[18] The membranes in question are the inner membranes of mitochondria, called cristae, which show a characteristic convolved folded structure. The number of folds depends on the intensity of the demand for ATP in the tissues being served. As ATP demand increases, the number of mitochondria will replicate to increase capacity as well.

Most of the diseases known among mitochondria are related by the failure of any of the enzymes in the production of ATP, such as cytochrome c which participates in the electron transport cycle. Such diseases emerge due to deleterious mutations in the mtDNA, producing reductions in the efficiencies of any of those processes necessary to the maintenance of oxidative phosphorylation that drives the metabolic citric acid cycle. These failures can result in myopathies, neuropathies such as Leber’s disease, as well as a failure of any specific organ functions relying upon ATP, resulting in other conditions such as diabetes. In any cell, and distributed throughout the organism, various mitochondria may bear different mtDNA sequences, a condition called heteroplasmy.[19] In any organism, the distribution of alleles may vary from tissue to tissue. More detailed reviews of human mitochondria are available.[20]

Each ovum therefore represents a population bottleneck in the transmission of mutated versions of heteroplasmic mtDNA to the next generation. From generation to generation in a population, drift and selection impacts the measured variation in the population.

Largely, most of the previously covered issues appear not to provide much insight into how mutations would prefer some climates to others that could lead different impacts in different parts of a phylogenetic tree. Two processes that may affect such issues include heat production via “proton leak” facilitated by thermogenin, found in brown adipose tissue.[21] This allows protons to release energy as heat when crossing the cristae membranes. However, this functions independently of the genetic structure of mtDNA, since nuclear DNA provides the code for thermogenin. Another enzyme that takes advantage of mitochondria is glutamine synthase, which detoxifies ammonia in liver tissue mitochondria.[22] This could be a candidate for differential diet, since need for detoxification of ammonia arises from the catabolism of proteins, and northern diets are protein rich. However, glutamine synthase is also coded in nuclear DNA. In this context, is also significant that Leber’s neuropathy shows gender bias, suggesting a disease process that also involves expression of nuclear proteins or other epigenetic factors.[23] The high rate of expression in the J haplogroup [20; 24; 25] suggests confounding factors in identifying selection as a cause of differential rates in northern climates.

Given the flexibility of mitochondria to respond to differential demand for ATP, and the relative insensitivity of those parts of mitochondrial metabolism to respond to environmental changes, it is difficult to target which features may be susceptible to specific environmental selection pressure, or more generally, how many of them contributing to observable rates might contribute. A test of differences in overall substitution rates in various parts of the phylogenetic tree have offered opportunities to search for such effects, as does this study. However, it is important to realize that the only instruments of sensitivity available to this study are aggregated substitution rates and ratios of nonsynonymous to synonymous substitutions. Further, the establishment of rates provides the context of, and limitations to, how effects of selection and neutral drift may interact with each other. Therefore, it is important to understand the formulation of rates, how they may be affected by selection, and limits to interpretation of variations in substitution rates.

A standard approach is to develop a phenomenological probability model, and explore the constraints that such a model imposes.

First, it is possible to at least conceptually consider probabilities such as $p(x_s,t)=P(X_s(t)=x_s)$, where $X_s(t)$ is a random variable representing a nucleotide type $x \in \{C,G,A,T\}$ at site s at time t. Now consider the relationship of probabilities of observing nucleotide values at some site at one time to those observed at a prior time. Specifically, the sample space of observable events at some site must satisfy

$$\{X_s(t_1)=x_s\}=\left\{X_s(t_1)=x_s \cap \left[\bigcup_{x_s'} X_s(t_2)=x_s'\right]\right\}=\bigcup_{x_s'}\left\{X_s(t_1)=x_s \cap X_s(t_2)=x_s'\right\}.$$

It follows that $p(x_s,t)=\sum_{x_s'} p\left(x_s,t \middle| x_s',t_0\right)p\left(x_s',t_0\right)$, a Markov process. This could be applied yet again iteratively, yielding the Chapman-Kolmogorov equation $p\left(x_s,t \middle| x_s',t_0\right)=\sum_{x_s''} p\left(x_s,t \middle| x_s'',t'\right)p\left(x_s'',t' \middle| x_s',t_0\right)$. Therefore, under very general considerations, a Markov model describing the behavior at a single site emerges naturally from an assumption of a substitution process regardless of dependence on other sites. Essentially, the above applies to the marginal summed over all other sites that specific substitution rates (not yet defined here) might depend upon.

The following caveat must be noted: while it is possible to describe transitions at an individual site with this model, and this description is valid even in the presence of interactions between sites (developed below), such interactions will invalidate a rigorous additivity of transition rates (also developed below).

Substitution Rates: the Differential Chapman-Kolmogorov Equation

A stochastic substitution rate may be constructed by defining

$$A_{ij}(t) = \lim_{\delta t \to 0} \frac{p(i,t+\delta t|j,t) - p(i,t|j,t)}{\delta t} = \lim_{\delta t \to 0} \frac{p(i,t+\delta t|j,t) - \delta_{ij}}{\delta t}.$$

The assumption that such a limit exists imposes constraints on the analytical form of $p\left(x_s,t\middle|x_s',t_0\right)$. The postulation of such a limit represents a model of substitution processes. As a description of substitution processes across a phylogenetic tree, which would be required of a Poisson regression "assuming" a molecular clock as applied to the whole phylogenetic tree, it must encompass molecular substitution processes, drift in and interactions between numerous populations that have shifted, formed, and evaporated over time, and both weak and strong selection effects. Specifically, it is assumed the effective δt is short compared to time measured, but long compared to transition events (generations) and selection and drift events (multiple generations). The assumption of a limiting rates imposes a diffusive character on the analytical form of the probabilities.

The existence of such a limiting form implies $p(i,t+\delta t|j,t) = \delta_{ij} + A_{ij}(t)\delta t + o(\delta t)$. The Chapman-Kolmogorov equation becomes $\frac{\partial p(i,t|j,t_0)}{\partial t} = \sum_k A_{ik}(t)p(k,t|j,t_0)$. Noting $1 = \sum_i p(i,t+\delta t|j,t) = 1 + \sum_i A_{ij}(t)\delta t + o(\delta t)$, it follows that $A_{jj}(t) = -\sum_{i,i\neq j} A_{ij}(t)$. The Chapman-Kolmogorov equation may then be rewritten $\frac{\partial p(i,t|j,t_0)}{\partial t} = \sum_{k,k\neq i}\left[A_{ik}(t)p(k,t|j,t_0) - A_{ki}(t)p(i,t|j,t_0)\right]$, a form called the "master equation."

Stationary and Equilibrium Solutions

There are a number of special cases for the A_{ij}'s that have been studied and used extensively. These include the Jukes-Cantor model,[26] Kimura 2-Parameter model,[27] Felsenstein '81 model,[28] Felsenstein '84 model,[28] Hasegawa, Kishino, and Yano '85 model,[29] Tamura-Nei '92 model,[30] Tamura-Nei '93 model,[31] and the General Time Reversible model (REV/GTR).[32; 33] One of the primary reasons so many variants exist is the diagonalization characteristics of these particular matrices allowing for closed-form analytical expressions of the genetic distance by means of eigentheory.

The expectation is that the $p\left(x_s, t \middle| x_s{}', t_0 \right) \to \pi_s$ independent of the initial state after some long time. Certainly the existence of such a stationary state can be guaranteed given a term-by-term cancellation in the master equation, yielding $A_{ik}\pi_k = A_{ki}\pi_i$. In this case, this states that the total rate of transition $i \to k$ is equal to the total rate of transition $k \to i$, a condition called detailed balance. If this is true in a tree, then it does not matter which direction one transverses an edge, and the phylogeny may be constructed in an unrooted manner, which is what is meant by a time-reversible model. This matrix relationship is satisfied by $A_{ik} = \pi_i S_{ik}$ for some symmetric S_{ik}.

It is possible to prove that a limiting stationary solution exists under GTR. Consider $u_i(t) = \left[p(i,t|k,t_0) - \pi_i \right] / \pi_i$. Then $\pi_i \frac{\partial u_i}{\partial t} = \sum_{k \neq i} \pi_i S_{ik} \pi_k (u_k - u_i)$. Next construct $\frac{\partial}{\partial t}\left(\sum_i \pi_i u_i^2 \right) = 2 \sum_{i,k,i \neq k} \pi_i S_{ik} \pi_k u_i (u_k - u_i) = -\sum_{i,k,i \neq k} \pi_i S_{ik} \pi_k (u_k - u_i)^2$. As long as the $A_{ij} \geq 0$ where $i \neq j$, then $\frac{\partial}{\partial t} \sum_i \pi_i u_i^2 < 0$. This implies that this positive number must decrease, approaching some greatest lower bound (0 is a lower bound, so some greatest lower bound possibly larger than 0 exists), at which time the rate of change of the $p\left(x_s, t \middle| x_s{}', t_0 \right)$ approaches zero. It is clear that the right-hand side is zero when all the u_i are equal. The only value $u_i = U$ that satisfies $\sum_i \pi_i u_i = 0$ is when all the $u_i = U = 0$, which leads to the equilibration condition $p\left(x_s, t \middle| x_s{}', t_0 \right) = \pi_{x_s}$ being *the* unique solution, and limit towards which this process must converge regardless of initial state.

This equilibration assumes sampling over long enough times so that each substitution would be thoroughly sampled. In practice, the effect of coalescence asserts that the entire human phylogenetic tree is less than 40 substitutions into the past acting on $S = 16kb$ along any one lineage. This implies that the number of expected mutations would be roughly $40 = \langle n \rangle = S\bar{r}t$, so that $\bar{r}t \approx 2.5 \times 10^{-3}$. This is not nearly long enough for the random substitution processes to sample the entire space necessary for equilibration of $p\left(x_s, t \middle| x_s', t_0 \right)$. In other words, the entire phylogenetic tree represents a nearly instantaneous time slice on the time scale that equilibration would operate.

Interactions between sites

Consider the case where two sites s_1 and s_2 are *not* independent. The simplest example might be where a substitution may be silent for one triplet, but nonsynonymous for another transition, as in where $ACU \rightarrow ACA$ would be silent, but $AGU \rightarrow AGA$ would not be silent. Other examples occur where one substitution may mitigate selection of a transition at another site, Other evidence indicates a prevalence of some pathogenic substitutions in certain haplogroups.[20; 24; 25] In the case of interactions, it follows that $P\left(X_{s_1}(t) = x_{s_1} \middle| X_{s_2}(t) = x_{s_2}\right) \neq P\left(X_{s_1}(t) = x_{s_1}\right)$, or $p\left(x_{s_1}, x_{s_2}, t\right) \equiv P\left(X_{s_1}(t) = x_{s_1} \cap X_{s_2}(t) = x_{s_2}\right) \neq p\left(x_{s_1}, t\right) p\left(x_{s_2}, t\right)$. In this case, the dependence of the joint distribution on time must satisfy $p\left(x_{s_1}, x_{s_2}, t\right) = \sum_{x_{s_1}', x_{s_2}'} p\left(x_{s_1}, x_{s_2}, t \middle| x_{s_1}', x_{s_2}', t_0 \right) p\left(x_{s_1}', x_{s_2}', t_0 \right)$. The equivalent of the Chapman-Kolmogorov equation satisfies $p\left(x_{s_1}, x_{s_2}, t \middle| x_{s_1}', x_{s_2}', t_0 \right) = \sum_{x_{s_1}'', x_{s_2}''} p\left(x_{s_1}, x_{s_2}, t \middle| x_{s_1}'', x_{s_2}'', t' \right) p\left(x_{s_1}'', x_{s_2}'', t' \middle| x_{s_1}', x_{s_2}', t_0 \right)$. The continuous time-like rate may be defined $A_{x_{s_1} x_{s_2} x_{s_1}' x_{s_2}'}(t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \left[p\left(x_{s_1}, x_{s_2}, t + \delta t \middle| x_{s_1}', x_{s_2}', t \right) - p\left(x_{s_1}, x_{s_2}, t \middle| x_{s_1}', x_{s_2}', t \right) \right]$.

The differential form of the Chapman-Kolmogorov equation may then be written

$\frac{\partial p\left(x_{s_1}, x_{s_2}, t \middle| x_{s_1}', x_{s_2}', t_0\right)}{\partial t} = \sum_{x_{s_1}'', x_{s_2}''} A_{x_{s_1} x_{s_2} x_{s_1}'' x_{s_2}''}(t) \cdot p\left(x_{s_1}'', x_{s_2}'', t \middle| x_{s_1}', x_{s_2}', t_0\right)$, and this relationship holds for specific states: $\frac{\partial p\left(x_{s_1}, x_{s_2}, t\right)}{\partial t} = \sum_{x_{s_1}', x_{s_2}'} A_{x_{s_1} x_{s_2} x_{s_1}' x_{s_2}'}(t) \cdot p\left(x_{s_1}', x_{s_2}', t\right)$.

These rates are constrained in their relationship to the one-site marginal rates. $\frac{\partial p\left(x_{s_1}, t\right)}{\partial t} = \sum_{x_{s_2}} \frac{\partial p\left(x_{s_1}, x_{s_2}, t\right)}{\partial t} = \sum_{x_{s_2}, x_{s_1}', x_{s_2}'} A_{x_{s_1} x_{s_2} x_{s_1}' x_{s_2}'}(t) \cdot p\left(x_{s_1}', x_{s_2}', t\right) = \sum_{x_{s_1}'} A_{x_{s_1} x_{s_1}'}(t) \cdot p\left(x_{s_1}', t\right)$, so that $A_{x_{s_1} x_{s_1}'}(t) = \sum_{x_{s_2}, x_{s_2}'} A_{x_{s_1} x_{s_2} x_{s_1}' x_{s_2}'}(t) \cdot p\left(x_{s_2}', t \middle| x_{s_1}', t\right)$. Therefore, even in the presence of interactions, the Markov substitution process for a single site can be described by a single rate; however, the possibility of adding the rates of two sites together is excluded.

If the sites are independent, then $P\left(X_{s_1}(t) = x_{s_1} \cap X_{s_2}(t) = x_{s_2}\right) = P\left(X_{s_1}(t) = x_{s_1}\right) P\left(X_{s_2}(t) = x_{s_2}\right)$. Inserting this into the rate equation, this yields the following constraint on rates:
$A_{x_{s_1} x_{s_2} x_{s_1}' x_{s_2}'}(t) = A_{x_{s_1} x_{s_1}'}(t) \delta_{x_{s_2} x_{s_2}'} + A_{x_{s_2} x_{s_2}'}(t) \delta_{x_{s_1} x_{s_1}'}$.
In other words, rates are then additive. The probability of observing multiple transitions in time δt is $O\left(\delta t^2\right)$, which is the effect of the Kronecker-δ's. The contribution to the rate of the transition of s_1 is not influenced by the value of x_{s_2}.

Interactions between substitutions at multiple sites could produce deviations as catalogued above: specifically, that 1) substitutions at any individual site is a Markov process without considering other sites even if interactions between sites exist; 2) simultaneous considerations of rates at two sites cannot be considered separately if interactions exist, and specifically, they are not additive, 3) the contributions of individual rates depends on the value of other sites' nucleotides, and double substitutions could be promoted with probabilities $O(\delta t)$. All of these effects result in the effective total rates not being additive.

Equilibration of interacting sites may be expected for extended general time-reversal systems satisfying detailed balance, where such balance involves a more complicated alphabet of combinations of sites. Here again, equilibration would assume processes that have proceeded long enough to effectively sample all of the combinations with as little phylogeny-induced linkage surviving as possible.

Independent Markov Substitution Events Must be Poisson Processes

Poisson processes are Markov processes, where the distribution of the number of transitions that occurred are counted as a function of time rather than the end-states resulting from specific transitions. If the rate that transition events happen is r, construct a distribution of transition events counting events $N(t)$. Then the probability of seeing n transitions at time $t+\delta t$ is equal to the total probability of seeing n at time t and none in time δt plus that of seeing $n-1$ at time t and one in time δt, or $P\big(N(t+\delta t)=n\big)=r\delta t\cdot P\big(N(t)=n-1\big)+(1-r\delta t)P\big(N(t)=n\big)$. The resulting differential equation satisfied by this is $\frac{\partial P\big(N(t)=n\big)}{\partial t}=r\cdot P\big(N(t)=n-1\big)-r\cdot P\big(N(t)=n\big)$. The similarity to the master equation is very visible. If the conditions are stationary, or the rates are relatively constant over the period being sampled, the distribution of events is Poisson distributed $P\big(N(t)=n\big)=\frac{(rt)^n}{n!}e^{-rt}$.

If a number of events occur independently, rates are additive. Given two groups of events, with cumulative substitution counts n_1 and n_2 over time t with rates r_1 and r_2, the distribution of the total number of counts $n=n_1+n_2$ will be distributed according to $P_n(t)=\sum_{\substack{n_1,n_2\\ n_1+n_2=n}}\frac{(r_1t)^{n_1}}{n_1!}e^{-r_1t}\frac{(r_2t)^{n_2}}{n_2!}e^{-r_2t}=\frac{[(r_1+r_2)t]^n}{n!}e^{-(r_1+r_2)t}$. In other words, the rates are cumulative.

It is clear from the above that Poisson processes are Markov processes where the state changes counted are number of events that have occurred. There is therefore a direct relationship between substitutions and the number of substitution events. At a particular site, the rate that a nucleotide species j transitions to species i is $A_{ij}\pi_j$. The total rate that all species j contribute to the probability of a specific species i is $\sum_{j \neq i} A_{ij}\pi_j$. The total rate of all transition *events* over all species i is then $\sum_{i, j \neq i} A_{ij}\pi_j$. If all sites s are independent, then the total rate is $r = \sum_{s,i, j \neq i} A_{sij}\pi_{sj}$.

Modeling Interactions and Variation

Any of the above interaction and selection effects can lead to an invalidation of a rigorously Poisson molecular clock, and the list is not necessarily exhaustive. However, the question remains as to how much of an effect these otherwise uncharacterized effects may be expected to have on a molecular clock. A standard way to account for variations introduced by weakly characterized, or “uncontrolled,” variables is to describe the effects of their variation with a random distribution. In this case, for any given specific site, the variations at other sites are assumed to be uncontrolled, resulting in an effective variation in the rate at the selected site. Since the relationship of any specific site to interactions induced by a neighbor may be different, the assumption may be made that the site’s transition is independent of the others, and to replace the fixed transition rates at each site with a randomly distributed variable. This model of variation makes it possible to treat the transitions as if they were independent, and therefore more specifically, to treat them as if they were additive. Likewise, an assumption of additivity of rates induces an effective variation in the rates of each of the sites.

Distinct from the assumption of the independence of transitions is an assumption that the effective variations between the rates imposed by the statistical model described here are independent. The assumption of independence in this variation in rates is a distinct and explicit assumption that becomes testable. This random variation is assumed to cover both variations between sites, and variations at each site due to interactions between sites due to not understood molecular interactions, differential correlation in selection, differential selection effects, and interactions between selection and population effects, as well as any other biological or population processes that are not understood, or which cannot adequately be characterized due to paucity of data. At the same time, other sources of variations in rates exist. Deamidation of methylated C's results in high rates of spontaneous mutation (CpG hotspots). It is notable that CpG mutations appears to involve complex interactions,[34,35] and that the effect of such enhancements appears to suppress their presence in the mitochondrial genome.[36; 37] Lastly, variations between sites can also be rolled into the mix. The validity of such an approach is not developed here. Its justification has stemmed primarily through its empirical connection to overdispersion. However, connecting these sources of variation to a model of uncontrolled variation is deferred to another study. The following section explores the impact of the assumption of variable but independent rates on the molecular clock.

Overdispersion and Violation of the Molecular Clock Hypothesis

Even in the case of variation in rates, the constancy of the total rate *could* be expected to be fairly robust. One way to view effects of variations in substitution rates sampled by the phylogenetic process is to allow each site to vary independently. This yields roughly 15000-16000 independent rates, which are distributed from site to site, and over time. Such variation has been modeled with fair success previously with a Γ-distribution.[38; 39] The Poisson distribution may be summed over the Γ-distributed rates. This can be considered to be a simple phenomenological model for the variations in the molecular clock.

If each individual site were identically and independently distributed according to a gamma distribution[38; 39] $\Gamma(r;\alpha,\beta)dr = \frac{\beta^{\alpha}}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r} dr$, then cumulative rate $r = r_1 + r_2$

where r_1 and r_2 are distributed as $\Gamma(r_1;\alpha_1,\beta)dr_1$ and $\Gamma(r_2;\alpha_2,\beta)dr_2$ will be distributed as $\Gamma(r;\alpha_1+\alpha_2,\beta)dr$. So this distribution also allows a cumulative description of molecular mutation rates for aggregates of sites. Now, the distribution of the number of mutations expected to occur at one site given it is drawn from a random selection of sites whose rates are distributed according to the gamma distribution is distributed according to the negative binomial distribution $P(n,t;\alpha,\beta)=\int_0^\infty dr\frac{(rt)^n}{n!}e^{-rt}\frac{\beta^\alpha}{\Gamma(\alpha)}r^{\alpha-1}e^{-\beta r}=\frac{\Gamma(\alpha+n)}{n!\Gamma(\alpha)}\left(\frac{t}{\beta+t}\right)^n\left(\frac{\beta}{\beta+t}\right)^\alpha$. The rules allowing aggregation of sampling sites immediately implies that the number n expected in time t out of L sites will be distributed as $P(n,t;L\alpha,\beta)$. Identifying $\bar{r}=\alpha/\beta$ and $\sigma_{\bar{r}}^2=\alpha/\beta^2$, this becomes $P(n,t;\alpha,\beta)=\frac{\Gamma(\alpha+n)}{n!\Gamma(\alpha)}\left(\frac{\bar{r}t}{\alpha+\bar{r}t}\right)^n\frac{1}{\left(1+\frac{\bar{r}t}{\alpha}\right)^\alpha}\xrightarrow[\alpha\to\infty]{}\frac{(\bar{r}t)^n}{n!}e^{-\bar{r}t}$. The mean and variance of n are $E(n)=\bar{r}t$ and $\mathrm{var}(n)=\bar{r}t\left(1+\frac{\bar{r}t}{\alpha}\right)$.

Any regime of the negative binomial distribution with large enough α to approximate the Poisson distribution has an α large enough to cause overdispersion to be negligible, which, even if significantly overdispersed for individual sites, since $\alpha\propto L$, it follows that the total is likely not significantly overdispersed. However, for the hypervariable regions, where L is significantly smaller and $\bar{r}$ is larger, it has been reported that the Γ distributed rates fits the observed data quite satisfactorily.[40] Even accounting for rather broad overdispersion in individual sites, the effect of summing over independently distributed contributions will tend to converge to a Poisson distribution, which is expected from more general considerations from the law of large numbers.

More to the point, if $\alpha=\alpha_o L$, then $\Gamma(r;\alpha,\beta)dr\approx\frac{\beta^\alpha}{\Gamma(\alpha)}\bar{r}^{\alpha-1}e^{-\beta\bar{r}}e^{-\frac{\alpha-1}{2\bar{r}^2}(r-\bar{r})^2}$, so the distribution becomes tightly peaked around $\bar{r}=\frac{L\alpha_0}{\beta}$, with a standard deviation proportional to $\sqrt{L}$. This implies that all the information about variation involving interactions, variations between sites, and epigenetically induced effects such as methylation of CpG sites, which were combined into just two parameters tends to become diluted when considering the combined rates of a large number of sites.

Even if individual sites show broad fluctuation, the sum of independent variants will tend towards a Poisson distribution. Given Γ-distributed variability, the only way to cause a significant change in a sum of a large number of samples would be if the variations among the rates were not independent (this is distinct from saying that the sites' substitutions interact).

Incomplete Phylogenies: collapsing nodes

Also an important consideration in this development, which has been freely exercised here, is that if a phylogeny is incomplete, and some bifurcations have been collapsed to single nodes, then counts m from two layers of nodes will have been combined into one node. Given a rate r, the distribution of total counts over time $t = t_1 + t_2$ becomes $P_m(t) = \sum_{\substack{m_1, m_2 \\ m_1 + m_2 = m}} \frac{(rt_1)^{m_1}}{m_1!} e^{-rt_1} \frac{(rt_2)^{m_2}}{m_2!} e^{-rt_2} = \frac{[r(t_1 + t_2)]^m}{m!} e^{-r(t_1+t_2)}$. In other words, collapsed clades in a phylogenetic tree will obey the same kind of statistics that a more detailed tree will show. This is just the non-differential form of the Chapman-Kolmogorov equation, that all continuous time Markov processes must satisfy.

Sources of violations of the molecular clock

Two sources of variations in rates have been identified. First, there are variations in the processes captured in the "limit" in which continuum rates are defined, which includes mutation, selection, and drift. Second, there is the effect of site-site interactions.

While it is possible to consistently model transitions and measure rates for individual sites as a Markov process, the effect of interactions is to render additivity invalid. Insisting on additivity effectively injects a random variation in modeling the interacting sites as uncontrolled in determining site rates. However, for large numbers of sites, if the variations in sites are uncorrelated, the distribution of the sum will approach a Poisson process. Correlated mutations tend to be expressed through selection where one mutation renders another mutation either more deleterious, or one mutation removes the deleterious effect of another mutation. Other examples include CpG type interactions where methylation at one site promotes errors in error correction at a neighboring site resulting in highly increased mutations at that site.

Alternatively, there are mechanisms that could promote correlations between the variations in the rates of substitutions that do not involve direct interactions. Processes of coalescence and of selection *both* remove lineages from a population. Removal of such lineages will result in a reduction in variation within the population. However, drift with no selection leads to a uniform rate of accumulation of substitutions in the population equivalent to that at the individual sites. The effect of selection becomes important when its rate exceeds the effect of drift.[41] The effect of selection is to remove substitutions from the population, not just variation. Selection, therefore, can result in a gradient in the effective substitution rate observed, if the probe provided by the phylogenetic tree acts on a scale δt that is smaller than the time scale that selection operates. This would produce a signature of systematically high substitution rates in the leaves compared to deeper nodes.

Conclusions

A molecular clock described by a Poisson distributed stochastic event is expected to follow from a Markov substitution model given no interactions between sites and constant rates at each site. However, numbers of effects, from correlated substitutions, CpG and other interactions invalidating the additivity of site-by-site substitution rates, to effects of selection and of drift acting on time scales longer than those probed by phylogentic bifurcations are all capable of invalidating the Poisson description of a molecular clock. If such violations are modeled by treating the effects as uncontrolled variation in the rates via inserting a random distribution, it is expected that, over the short time to the human mtDNA MRCA, the molecular clock would be expected to be robust unless there are correlations between the modeled variations *of rates* for a significant number of sites, or unless certain interactions may boost rates by nearly singular multiples.

Conditions involving interactions that would most closely satisfy the assumptions where such an approximation could be made would be if the transitions were sampled from a mixed set of phylogenies where the whole group had equilibrated. Clearly this is not the case here. Further, sampling along branches of a phylogeny specifically isolate groups in ways that will differentiate correlated mutations dependent on the markers that identify each phylogenetic branch. Further, the partition by branches also isolates time-dependent effects, such as purifying selection, which will tend to identify different rates near the leaves compared to older branches. Therefore, deviations from a Poisson clock applied to a phylogenetic tree will act as evidence for purifying selection or interactions producing correlated mutations.

Bibliography

1. Zukerkandl E, Pauling L (1962) Molecular Disease, Evolution, and Genic Heterogeneity. In: Kasha M aPB (ed) Horizons in Biochemistry. Academic Press, New York, pp 189-225
2. Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69:1348-1356
3. Bromham L, Penny D (2003) The modern molecular clock. Nat Rev Genet 4:216-224
4. Ho SY, Larson G (2006) Molecular clocks: when times are a-changin'. Trends Genet 22:79-83
5. Hedges SB, Kumar S (2003) Genomic Clocks and Evolutionary Timescales. Trends in Genetics 19:200-206
6. Welch JJ, Bromham L (2005) Molecular dating when rates vary. Trends Ecol Evol 20:320-327
7. Ohta T, Kimura M (1971) On the constancy of the Evolutionary Rate of Cistrons. J Mol Evol 1:18-25

8. Ohta T (1987) Very slightly deleterious mutations and the molecular clock. J Mol Evol 26:1-6

9. Ohta T (2002) Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci U S A 99:16134-16137

10. Ohta T (2002) The Nearly-Neutral Theory of Molecular Evolution. Annu Rev Ecol Syst 23:263-286

11. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olckers A, Wallace DC (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100:171-176

12. Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373-387

13. Ingman M, Gyllensten U (2007) Rate variation between mitochondrial domains and adaptive evolution in humans. Hum Mol Genet 16:2281-2287

14. Nei M, Maruyama T, Chakraborty R (1975) The Bottleneck Effect and Genetic Variability in Populations. Evolution 1:1-10

15. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature 396:133-140

16. Mitchell P, Moyle J (1967) Chemiosmotic hypothesis of oxidative phosphorylation. Nature 213:137-139

17. Mitchell P (1967) Proton current flow in mitochondrial systems. Nature 214:1327-1328

18. Platt DE (2001) Are mitochondria mesoscopic? Biophys Chem 91:245-252

19. Fan W, Waymire KG, Narula N, Li P, Rocher C, Coskun PE, Vannan MA, Narula J, Macgregor GR, Wallace DC (2008) A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. Science 319:958-962

20. Chinnery PF (2006) Mitochondrial DNA in Homo Sapiens. In: Bandelt HJ, Macaulay V, Richards M (eds) Human Mitochondrial DNA and the Evolution of Homo Sapiens. Springer-Verlag, Berlin, pp 3-15

21. Mozo J, Emre Y, Bouillaud F, Ricquier D, Criscuolo F (2005) Thermoregulation: what role for UCPs in mammals and birds? Biosci Rep 25:227-249

22. Matthews GD, Gould RM, Vardimon L (2005) A single glutamine synthetase gene produces tissue-specific subcellular localization by alternative splicing. FEBS Lett 579:5527-5534

23. Yen MY, Wang AG, Wei YH (2006) Leber's hereditary optic neuropathy: a multifactorial disease. Prog Retin Eye Res 25:381-396

24. Torroni A, Petrozzi M, D'Urbano L, Sellitto D, Zeviani M, Carrara F, Carducci C, Leuzzi V, Carelli V, Barboni P, De Negri A, Scozzari R (1997) Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing

the penetrance of the primary mutations 11778 and 14484. Am J Hum Genet 60:1107-1121

25. Howell N, Kubacka I, Halvorson S, Howell B, McCullough DA, Mackey D (1995) Phylogenetic analysis of the mitochondrial genomes from Leber hereditary optic neuropathy pedigrees. Genetics 140:285-302
26. Jukes T, Cantor C (1969) Evolution of Protein Molecules. Mammalian Protein Metabolism. Academic Press, New York
27. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111-120
28. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376
29. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174
30. Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 9:678-687
31. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512-526
32. Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105-111
33. Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39:315-329
34. Zhao Z, Jiang C (2007) Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. Mol Biol Evol 24:23-25
35. Fryxell KJ, Moon WJ (2005) CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 22:650-658
36. Karlin S, Doerfler W, Cardon LR (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J Virol 68:2889-2897
37. Cardon LR, Burge C, Clayton DA, Karlin S (1994) Pervasive CpG suppression in animal mitochondrial genomes. Proc Natl Acad Sci U S A 91:3799-3803
38. Ota T, Nei M (1994) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 38:642-643
39. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306-314
40. Bandelt HJ, Kong QP, Richards M, Macaulay V (2006) Estimation of mutation rates and coalescence times: some caveats. In: Bandelt HJ, Macaulay V, Richards M (eds) Human Mitochondrial DNA and the Evolution of Homo Sapiens. Springer-Verlag, Berlin, pp 47-90
41. Kimura M (1983) The Neutral Theory of Evolution. Cambridge University Press, New York

Appendix B: Maximum Likelihood Poisson Regression

The method presented here echoes Sarich and Wilson's approach,[1] which compared differences between Poisson-distributed variables, determined by simulation to be normally distributed for Poisson counts $N \geq 20$, yielding χ^2 distributed variables.[2] A later study compared observed counts with those of counts placed by simulation on a phylogenetic tree, essentially implementing a Poisson distribution to measure probabilities branch-by-branch.[3] Ancestral states were inferred using a modified Sarich-Wilson algorithm,[1] consistent with the relative rates test of the molecular clock. The approach seeks in large part to assess the level of noise in the system, and, as much as possible, to exclude it, rather than to accommodate it.

An earlier Poisson regression study[4] by Rosset introduced the Poisson regression method with molecular clock constraints, and explored the question of deviations from the molecular clock. That study employed the Fitch algorithm,[5] which, in the case of evidence of multiple mutations along a single site (homoplasy), identifies multiple numbers of candidate ancestral states that yield the same number of mutations (maximum parsimony), or, if scored with a substitution rate matrix, the same maximum score. Such ambiguities are identified in a two-step process, the second step recognizing the possibilities of multiple mutations along the lineage. Without such mutations, the first step yields results equivalent to the results of Sarich and Wilson. Rosset did also show that the probability of seeing such mutations given current estimates of mutation rates is very small in the human mtDNA phylogeny. The Sarich and Wilson algorithm lends itself to comparisons across a number of haplotypes per node at once, allowing for threshold-limited selection of mutations. It is more difficult to adapt the Fitch algorithm to this purpose. While this was adapted to attempt to deal with some of the datasets, the dataset ultimately identified showed very little need to manage for errors.

For each node, given a parameter

$$\lambda = rt \geq 0$$

for rate of mutation r over time t, the probability of observing n mutations is

$$P_m(t) = \frac{\lambda^m}{m!} e^{-\lambda}.$$

The log-likelihood function for n observed m_j, one for each haplotype, is

$$L(\lambda) = \sum_j \left[m_j \ln \lambda - \lambda + c(m_j) \right]$$

which extremizes at

$$\lambda = \frac{\sum_j m_j}{n}.$$

The variance in λ may be computed from

$$\sigma_\lambda^2 = -\left(\frac{d^2 L}{d\lambda^2} \right) = \frac{\lambda}{n} = \frac{\sigma_m^2}{n}.$$

This does not account for the molecular clock constraint that the time from this node to each of its leaves must be the same as the time for its sibling to each of the sibling's leaves. For each node i, the time for the two children will be labeled t_{i1} and t_{i2}. There is therefore a constraint imposed at each node that

$$t_{i1} = t_{i2}.$$

Rather than managing multiple constraints, a cost may be added to the maximum likelihood function that maximizes with value 0 where the $t_{i1} = t_{i2}$. Such a function could be represented by a simple quadratic of the form $-K(t_{i1} - t_{i2})^2$. This imposes a "ridge" on the maximum likelihood function, which satisfies the constraint more closely as $K \to +\infty$. The modified log-likelihood function becomes

$$L_K = \sum_i \left(M_i \ln \lambda_i - N_i \lambda_i - K(t_{i1} - t_{i2})^2 \right)$$

where the $M_i = \sum_j m_{ij}$ for the n_i haplotypes in node i, and where the extremization of L_K converges to the values for L subject to the equal-time constraints as $K \to +\infty$. This is equivalent to introducing a representation of a Dirac-δ representation $\delta(x) = \lim_{\sigma \to 0^+} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ in order to require convergence. The effective sequence of penalty functions has benign computational characteristics that allow easy convergence.

In seeking to apply simple maximization techniques to this problem, another issue is the constraint that the $\lambda_i \geq 0$. This condition may be transformed into a form that is amenable to numerical methods by mapping $\lambda_i = e^{x_i}$. The auxiliary log-likelihood function may then be smoothly differentiated using simple numerical techniques, such as Richardson's extrapolation, which was employed here to compute first and second partial derivatives of the log-likelihood function with respect to the x_i's.

Until now, there has been no explicit assumption that the mutation rate is the same for all branches. At this point, it is explicitly inserted. For ease of notation and computation, time units are chosen so that the rate of mutation is unity (i.e., the unit of time is the amount of time for the expected number of mutations to be one). Then, at each node, the estimated time passed up to the parent node is defined to be

$$t_i = e^{x_i} + \frac{t_{i1} + t_{i2}}{2}.$$

(Note, the average could be arbitrarily weighted, so long as the result is between the two times; as $|t_{i1} - t_{i2}|$ is reduced as K increases, both terms in the average approach each other). The auxiliary log-likelihood function may be expressed as

$$\begin{aligned} L_K\left(x + \overset{\circledR}{\delta x}\right) &= L_K(x) + b \cdot \delta x + \frac{1}{2}\delta x \cdot C \cdot \delta x + O\left(dx^3\right) \\ &= L_K(x) - \frac{1}{2} b \cdot C \;\; \cdot b + \frac{1}{2}\left(\delta x + C \;\; \cdot b\right) \cdot C \cdot \left(\delta x + C \;\; \cdot b\right) + O\left(dx^3\right) \end{aligned}$$

where

$$\begin{aligned} b &= \nabla_x L_K \\ C &= \nabla_x \nabla_x L_K. \end{aligned}$$

and

$$C \;\; = \lim_{\varepsilon \to 0}(C + I\varepsilon)^{-1} C (C + I\varepsilon)^{-1}$$

represents the inverse with contributions from the zero-valued eigenvalues of C removed since there are no contributions to L_K from such components of $\overset{\omega}{\delta x}$. These vectors and matrices are computed via Richardson's extrapolation.[6] This maximizes where

$$\delta x = -C \;\; \cdot b.$$

As in the simple Newton's method, this tends to show quadratic convergence for the sequence $\vec{x}_{n+1} = \vec{x}_n + \delta\vec{x}_n$, so long as the curvature is small over the length scale of the step size. This can be a problem for large K, so a schedule of increasing K's are applied to a succession of roots, checking that the converged value for large K is independent of the schedule. The convergence misbehaves if empty nodes are included. Then the MLE λ's are

$$\lambda_j = e^{x_j}$$
$$\sigma_{\lambda_i \lambda_j} = -e^{x_i + x_j} \left(C \ \right)_{ij}.$$

For a linear thread from a node up to some ancestor node, passing through a set of nodes D, the time estimates are

$$t_D = \sum_{j \in D} \lambda_j,$$
$$\sigma_{t_D}^2 = \sum_{j, j' \in D} \sigma_{\lambda_j \lambda_{j'}}^2.$$

Note that the covariances between variables must be included to correctly propagate errors since the constraints impose correlations in the variations of the parameters.

References

1. Sarich VM, Wilson AC (1973) Generation time and genomic evolution in primates. Science 179:1144-1147
2. Wu CI, Li WH (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci U S A 82:1741-1745
3. Takezaki N, Rzhetsky A, Nei M (1995) Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol 12:823-833
4. Rosset S (2006) Efficient Inference on Known Phylogenetic Trees Using Poisson Regression. Bioinformatics (Proc of the 5th European Conference on Computational Biology ECCB-2007) 23:e142-e147
5. Fitch WM (1971) Toward defining the course of evolution: defining the minimum change for a specific tree topology. Systematic Zoology 20:406-416
6. Richardson LF (1927) The deferred approach to the limit. Philosophical Transactions of the Royal Society of London, Series A 226:299-349