RC24556 (W0805-076) May 12, 2008
Other

IBM Research Report

Relationships between Molecular Clock Deviations and High Nonsynonymous to Synonymous Ratios among Some Older Haplogroups

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

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Relationships between molecular clock deviations and high nonsynonymous to synonymous ratios among some older haplogroups.

Daniel E. Platt

Abstract

This study presents an exhaustive computation of the deviations of substitution counts from what would be expected by a maximum likelihood Poisson regression model of 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, primarily in leaf nodes, suggesting most nonsynonymous substitutions are recent, and not yet excluded by selection pressure, in agreement with prior studies. We have verified prior studies also reporting some deviations of nonsynonymous to synonymous substitution ratios between northern and temperate climes, but, at the finer level of analysis, 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. We have also identified deviations in older interior clades that share relationships with each other, suggesting drift effects fixing nonsynonymous substitutions before selection removed them from the population.

Introduction

The molecular clock hypothesis [1] represents an ideal against which deviations both reveal information and inject difficulties. Genetic forces that may promote deviations from the molecular clock are a topic of increasing interest in recent research on the human expansion.[2, 3, 4, 5, 6, 7, 8, 9] The difficulties associated with deviations from the molecular clock on estimation of times of most recent common ancestors,[10,11] and on phylogeny construction (particularly across taxa), and limited available data, has 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.[13,14,15,16,17]

Appendix A revisits the standard development of Markov processes in describing substitution processes. Instantaneous substitution rate matrices that emerge from that description form the basis of substitutions as a Poisson process, both in terms of actual substitution events as well as genetic distance. If substitution rate matrices reflect the differences present in mtDNA samples, they will reflect molecular substitution processes, heteroplasmy, selection, and population effects such as drift. These issues inject variations in the relative timescales of selection and fixation, producing differences in the rates observed in different parts of the phylogenetic tree. Other issues include the impact of correlated substitutions on the ability to sum substitution rates, as well as the distinct question of the role that correlation between rate variations must play in order to observe significant deviations from Poisson clock-like behavior. Deviations from the molecular clock are tied to relationships between selection and drift, allowing a Poisson maximum likelihood regression to act as a probe for processes inducing the differential rates described here.

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.[7,8,9] 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] 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.[18] 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.[19] Ancestral states were inferred using a modified Sarich-Wilson algorithm,[20] consistent with the relative rates test of the molecular clock.

Materials and Methods

Alignments and SNP Counts

A dataset of 3839 complete mtDNA sequences were compiled from public databases Mitomap [21] (http://www.mitomap.org/euk_mitos.html), and NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?term=Homo[Organism]%20AND%20mitochondrion[All%20Fields]%20AND%2015000:17000[SLEN]%20NOT%20pseudogene[All%20Fields]&cmd=Search&db=nuccore&QueryKey=1) around November 21, 2007 . These sequences were pair-wise aligned with the Revised Consensus Reference Sequence rCRS [22] by applying the linear global alignment algorithm "stretcher" [23] implemented in Emboss [24] (http://emboss.sourceforge.net), and by ClustalW [25] and all SNPs identified across all of the samples 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 haplotypes bearing such SNPs. Deviations from the rCRS that represent insertions or deletions represent special problems in comparing two haplotypes [26,27] 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 significant blind spots that could be important to groups outside of the rCRS H haplogroup, as well as 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 haplotypes in which such a SNP may appear would be expected to be low. For each alignment, the number of haplotypes supported by each SNP was determined. The relationship between support and putative error is explored.

A similar analysis, applying ClustalW to data collected by Herrnstadt et al [7,28] was aligned with the rCRS, and SNPs were identified. Following a neighbor-joining analysis,[29] it was pointed out by Bandelt [30] that there are a number of sources of error, some of which specifically identified in the Herrnstadt et al neighbor-joining study. Herrnstadt responded by acknowledging Bandelt's contribution,[28] identified and corrected those errors, and more, 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 haplotypes, all with control region sites. This new data was published with the supplementary material online.[7]

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.[31] At each site, an index of all possible substitutions (A, C, G, T) against the RCRS, using the tRNA tables[31] to determine peptides for each specific possible nucleotide substitution in the coding segments.

Haplogroup Assignment

These haplotypes were assigned to haplogroups as outlined by the Genographic Project public participation markers.[32] The phylogenetic tree shows polytomy, which was reduced to bifurcations [33] with branch orders selected to roughly reflect observed mutation counts. Note that labels marked with "x", representing "complement," provide references for otherwise un-named clades necessary for discussion, but do not reflect any intent to introduce nomenclature. The haplotypes were assigned to each clade and

subclade down to the leaves of the phylogenetic tree following the protocol defined in the Genographic report. 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/) [34] The L2 clades are more consistently marked. The Mitomap phylogeny avoids some of these issues by simply not labeling the branches except by marker. Other phylogenies consulted include those of Bandelt [35] and Macaulay.[36]

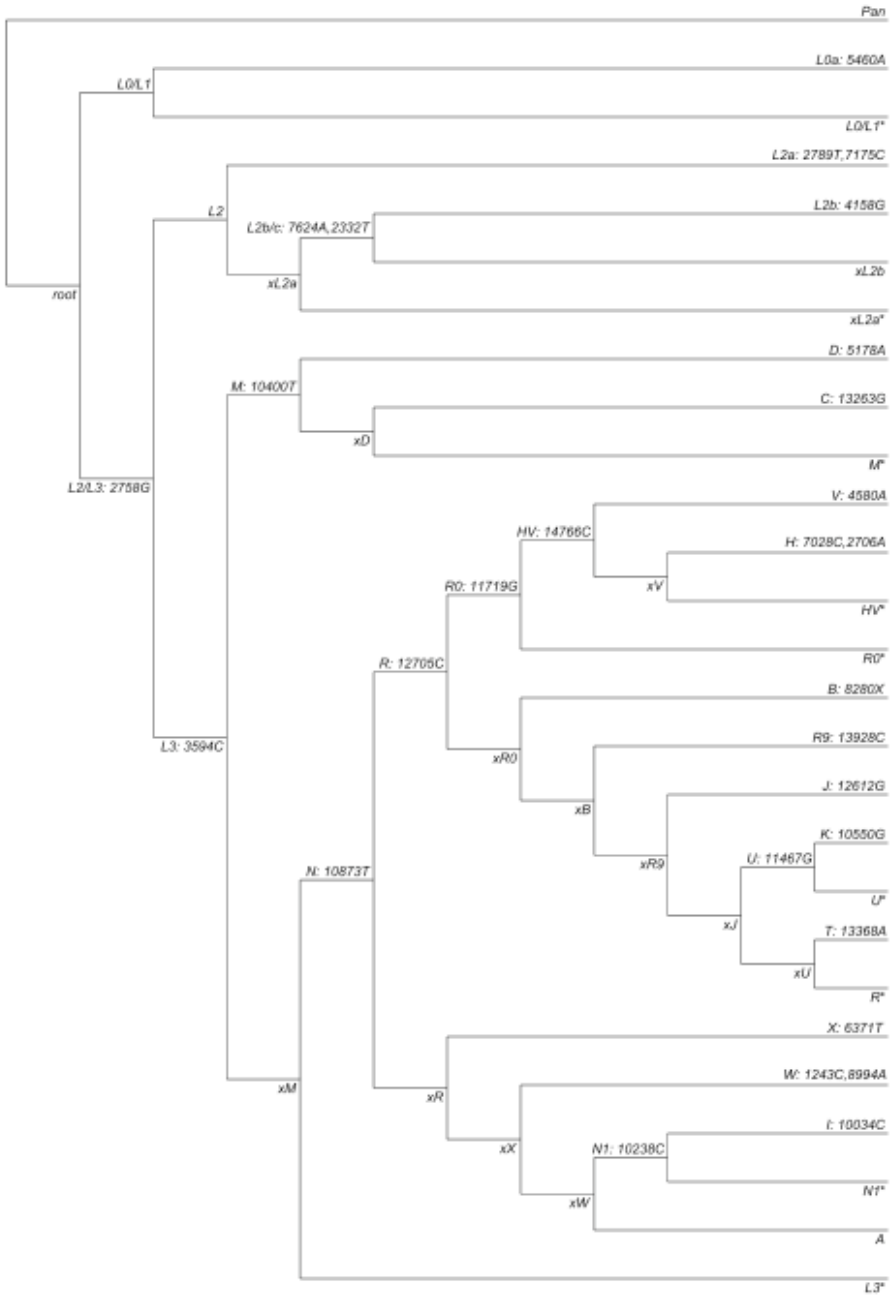
Pan
L0a: 5460A
L0/L1
L0/L1*
L2a: 2789T,7175C
L2b: 4158G
L2
L2b/c: 7624A,2332T
xL2b
root
xL2a
xL2a*
D: 5178A
M: 10400T
C: 13263G
xD
M*
L2/L3: 2758G
V: 4580A
HV: 14766C
H: 7028C,2706A
xV
R0: 11719G
HV*
R0*
R: 12705C
B: 8280X
R9: 13928C
L3: 3594C
xR0
J: 12612G
xB
K: 10550G
xR9
U: 11467G
N: 10873T
U*
xJ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

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

Inferring Ancestral States

The algorithm employed in this study is an adaptation of Sarich and Wilson's approach.[20] The adaptation described here essentially considers each SNP and mutation candidate individually to assign contribution to the appropriate node. Thresholds are applied to sibling and outgroup clade comparisons to avoid counting low-support errors.

Identification of mutations, or which haplotype nodes should own which mutations given homoplasies, should take cognizance the following. First a mutation that occurred in a clade is likely not going to be present in its sibling. If a SNP is present in both siblings, the mutation marked by the SNP occurred in a parent mutation. Given that a SNP is so marked, as absent from its sibling, looking in the parent clade's sibling can further check the mutation state of that SNP value. If a mutation occurred in a clade, generally, both subclades will contain some haplotypes carrying the marker. An algorithm that places a candidate mutation in a node that 1) is present in some haplotypes above a proportionate threshold in the node, 2) is absent (below some stringent threshold) in neighbor and nearest outgroup (parent's sibling – in the case of the root node, the nearest outgroup is comprised of two chimpanzees and a bonobo obtained from http://www.genpat.uu.se/mtDB/ [34]), 3) and is present (above some threshold) in haplotypes assigned to both children (if it is not a leaf), will guarantee placement as recently in the tree as possible, and can place markers in multiple locations within the phylogeny (homoplasy) where this mutation could occur, except for those so close to each other that it would not be possible to resolve the mutation from sibling and parent's sibling. These conditions are guaranteed to preserve mutations when subclade bifurcations are trimmed yielding lower resolution in the tree, except for the case of homoplasy in the nearest outgroup (parent's sibling). While the number of mutations associated with a trimmed node may increase (approximately the union of both children), the number of such mutations associated with each of the haplotypes, also comprised of

the union that would have been assigned to subclades, does not increase, and may actually decrease if some mutations are lost due to homoplasy in the outgroup. The estimate of the number of mutations that occurred along the lineages within a node is computed from the average and variance of the number of mutations observed across the haplotypes in that node.

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.[37] 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. An advantage to the modified relative-rates approach is that it easily allows the use of thresholds to screen and/or allow for some of the noise identified in the data set.

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

Variations in the number of mutations identified in each node across the haplotypes are also recorded, to estimate overdispersion [38, 39] of the molecular clock relative to the expected Poisson distribution. Such estimate must be done within each node to control for the significant correlation imposed by the phylogenetic tree structure. While overdispersion is a sure demonstration of violation of the molecular clock, correlations due to phylogeny may result in apparent underdispersion in a sample. Unfortunately, measures of overdispersion can be sensitive to sampling weights of haplotypes within any particular clade, confounding results. Therefore, that analysis is not reviewed here.

An estimate of the probability of finding that many or more/less (depending on whether the average substitution count is above or below that expected by ML Poisson regression) is computed. There is a challenge in that this would represent the probability of finding an individual haplotype with that much deviation, not the average for the haplogroup node. A difficulty exists in that the haplotype 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 haplotypes, yet each haplotype shows an average of around two substitutions, indicating low correlation).

Silent and Nonsynonymous Substitutions

The ratio k_a/k_s is computed for each node. These counts are compared to a binomial distribution where the probability 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.

Evidence of interactions

The presence of correlated substitutions, as described in the appendix, would challenge the additivity of substitution rates across sites. Other dependencies would be difficult to detect given the very low substitution rate (there are only around 1700 substitutions out of 15,000+ sites, with each haplotype carrying some number less than 40 substitutions). This implies that dependence of rates on neighboring states is not sampled across a wide range of substitutions. This leaves the question of whether correlated mutations occur in numbers large enough to impact rate calculations. This is distinct from the question of whether such interactions exist, but rather whether such interactions would be sufficient in number to affect rates. Parenthetically, if they were present in such numbers, phylogeny reconstruction efforts would likely reflect these effects in greater number.

Results

Haplogroup Assignments

The alignments of the Herrnstadt set and subsequent SNP identification produced 1779 SNPs. These were retained. The alignments with the NCBI set showed more than 11,000 SNPs, reflecting significant noise. A histogram across sites showed no distinction between coding region and hypervariable regions in SNP counts. Exclusion of SNPs below support thresholds showed reductions in counts that changed qualitatively at around 8 haplotypes, with a much more level fall-off of SNPs at higher supports. The number of recorded SNPs at 10 haplotypes support showed around 2,000 SNPs, with a histogram that clearly identified hypervariable regions as distinct from coding regions. There are some haplogroups with smaller numbers of haplotypes than the support level of 10, rendering any estimate of substitution counts in those groups suspect. Some HV markers showed apparent back-mutations, misplacing significant numbers of H haplotypes in R clades. For these reasons, the NCBI set was not used.

The identification of markers in the L clades corresponded to those reported elsewhere. [7,8,29]

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.[3,4,7,8,9] 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 haplotypes 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 haplotypes register any substititons in Hg U.

Haplogroup H is unusual. It has a large number of haplotypes, and a large number of substitutions in the group. However, the number of substitutions per haplotype 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 haplotypes and their substitutions sampled appears to indicate some unusual character in or around this clade.

Haplogroup	λ	σ_λ	m	σ_m	L-R tail	$\sum_m P(m;\lambda)$
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	-	------
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	-	------
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

Haplogroup	λ	σ_λ	m	σ_m	L-R tail	$\sum_m P(m;\lambda)$
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
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 1. Estimated times (T), Poisson parameters (λ),number of mutations (m) at each node, and number of haplotypes N for the full phylogenetic tree from the Herrnstadt set.

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 [40] for human and other species.

Haplogroup	# Haplotypes	k_a/k_s	N	L-R Tail	$\sum_{n_a} P(n_a;p,N)$
root	648.0000	-------	0.0000	-	-------
L0/L1	39.0000	1.0000	2.0000	>	0.6037
L0a	7.0000	0.3333	40.0000	<	0.0760

Haplogroup	# Haplotypes	k_a/k_s	N	L-R Tail	$\sum_{n_a} P(n_a;p,N)$
L0/L1*	32.0000	0.6338	116.0000	>	0.3819
L2/L3	609.0000	0.6667	10.0000	>	0.5413
L2	63.0000	0.0000	6.0000	<	0.0622
L2a	48.0000	0.6216	60.0000	>	0.4666
xL2a	15.0000	-------	0.0000	-	-------
L2b/c	13.0000	0.5000	3.0000	<	0.6899
L2b	8.0000	0.4706	25.0000	<	0.3828
xL2b	5.0000	1.3750	19.0000	>	0.0523
xL2a*	2.0000	0.3077	17.0000	<	0.1847
L3	546.0000	0.3846	18.0000	<	0.2900
M	26.0000	1.0000	2.0000	>	0.6037
D	9.0000	2.3333	10.0000	>	0.0359
xD	17.0000	0.6667	5.0000	>	0.6102
C	12.0000	1.0000	10.0000	>	0.2951
M*	5.0000	0.6000	16.0000	>	0.5785
xM	520.0000	0.2727	14.0000	<	0.1765
N	474.0000	1.5714	18.0000	>	0.0328
R	416.0000	1.0000	16.0000	>	0.2060
R0	228.0000	0.0000	1.0000	<	0.6295
HV	227.0000	-------	0.0000	-	-------
V	8.0000	0.4966	3.0000	<	0.2495
xV	219.0000	0.0000	1.0000	<	0.6295
H	209.0000	0.6667	135.0000	>	0.2658
HV*	10.0000	1.0000	4.0000	>	0.4732
R0*	1.0000	-------	0.0000	-	-------
xR0	188.0000	0.2914	31.0000	<	0.0279
B	19.0000	0.9000	19.0000	>	0.2409
xB	169.0000	2.0000	3.0000	>	0.3101
R9	2.0000	1.0000	2.0000	>	0.6037
xR9	167.0000	4.0000	5.0000	>	0.0663
J	32.0000	0.8333	22.0000	>	0.2720
xJ	135.0000	0.5000	12.0000	<	0.5235
U	86.0000	-------	1.0000	>	0.3705
K	45.0000	0.5789	30.0000	<	0.5642
U*	41.0000	0.5366	63.0000	<	0.4176
xU	49.0000	-------	0.0000	-	-------
T	42.0000	0.6250	26.0000	>	0.5143
R*	7.0000	0.5000	3.0000	<	0.6899
xR	58.0000	-------	0.0000	-	-------
X	11.0000	0.5216	9.0000	<	0.5572
xX	47.0000	1.0000	2.0000	>	0.6037
W	8.0000	0.3077	17.0000	<	0.1847
xW	39.0000	0.0000	1.0000	<	0.6295
N1	14.0000	1.0000	4.0000	>	0.4732
I	12.0000	0.6000	16.0000	>	0.5785
N1*	2.0000	0.5000	3.0000	<	0.6899
A	25.0000	0.4706	25.0000	<	0.3828

Haplogroup	# Haplotypes	k_a/k_s	N	L-R Tail	$\sum_{n_a} P(n_a;p,N)$
L3*	46.0000	0.4208	81.0000	<	0.0759

Figure 2. $K_a/K_s = 0.5885$ for the entire population. 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.

Interaction Candidates

Substitutions that appeared in L2b, one of the candidates for unusual numbers of substitutions, and in other non-L haplogroups in the Herrnstadt dataset include 3 from H: 4185T, 4767G, and 5237A, and single entries K: 5237A, 6: 6026A, U*: 6629G (two haplotypes), xJ: 12406A (three haplotypes out of 135), M*: 15236G, and R: 15326A (eight haplotypes out of 416). These do not appear to represent significant contributors to rates.

Conclusions

Analysis of the Markov character of stochastic substitution models indicates that independent rate variations should show molecular clock-like behavior in a fairly robust manner. Deviations must imply either interactions between substitutions, or correlations between substitution rates. Substitution interactions show two types of signatures: 1) rates of multiple substitution events of $O(\delta t)$, or 2) rates at one site that depend on nucleotide values at other sites. While some rare homoplasies were identified that may represent candidates, significant numbers would need to be apparent in order to significantly impact substitution rates. Given less than 40 substitutions back to the most recent common ancestor for the human mtDNA phylogeny, and around 1000 SNPs out of the roughly 15,000 coding region sites, there is not significant opportunity for these interactions to affect rates sufficiently to account for deviations from the molecular clock. What remains are correlations between variations in substitution rates. Without such correlations, the cumulative rate would converge to a fairly stable mean due to the action of the central limit theorem. Therefore the effective number of independently varying rates must be reduced by correlation. This leaves the question of which processes can lead to correlations between large numbers of sites.

Selection, by itself, could account for gradients in substitution rates among substitutions marking older haplogroups to substitutions that occurred within younger haplogroups. Population changes without selection will not affect substitution rates.[41] 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.

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. 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. Hags 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, which are branch points of a number of haplotypes showing movements 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 most deviations occur in 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, they are very deep clades, showing 20's of substitutions per haplotype. This suggests that they should be dominated by older substitutions, with many of the deleterious nonsynonymous substitutions having been removed by selection. Yet, they show significant deviations from expectation not typical of clades closer in age to their common branch points. In this respect, they share some similarities to the xR9 clade.

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[42] 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.

Acknowledgements – The authors have 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 the use of measurements such as nonsynonymous to synonymous substitution ratios for understanding the quality of the data, to consideration of more factors that may be

contributing to the organization of deviations from a mean molecular clock, to pushing for the strongest conclusions that might be drawn from the analysis, and giving the insight of an editor's eyes. Our deepest thanks. Next, thanks goes to Saharon Rosset, questioning the validity of a Poisson distribution for substitution events pushed the question of 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. Zuckerkandl, E. and Pauling, L.B. (1962). "Molecular disease, evolution, and genetic heterogeneity", in Kasha, M. and Pullman, B (editors): Horizons in Biochemistry. Academic Press, New York, 189–225.

2. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, Easley K, Chen E, Brown MD, Sukernik RI, Olkers A, Wallace DC, "Natural Selection Shaped Regional mtDNA Variation in Humans," PNAS 100, 171-176 (2003).

3. Kisiveld T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Unerhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knijff P, Feldman M, Cavalli-Sforza LL, Oefner PJ, "The Role of Selection in the Evolution of Human Mitochondrial Genomes," *Genetics* **172**, 373-387 (2006).

4. Ingman M, Gyllensten U, "Rate Variation Between Mitochondrial Domains and Adaptive Evolution in Humans," Human Molecular Genetics 16(19), 2281-2287 (2007).

5. Xing Y, Lee C, "Can RNA Selection Pressure Distort the Measurement of Ka/Ks?" Gene 370: 1-5 (2006).

6. Elson JL, Turnbull DM, Howell N, "Comparative Genomics and the Evolution of Human Mitochondrial DNA: Assessing the Effects of Selection," Am. J. Hum. Genet. 74: 229-238 (2004).

7. Howell N, Elson JL, Turnbull DM, Herrnstadt C, "African Haplogroup L mtDNA Sequences Show Violations of Clock-like Evolution," Mol. Biol. Evol. 21(10): 1843-1854 (2004).

8. Howell N, Elson JL, Howell C, Turnbull DM, "Relative Rates of Evolution in the Coding and Control Regions of African mtDNAs," Mol. Biol. Evol. 24(10): 2213-2221 (2007).

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, "Do the Four Clades of the mtDNA Haplogroup L2 Evolve at Different Rates?" Am. J. Hum. Genet. 69, 1348-1356 (2001).

10. S. Rosset (2006), "Efficient Inference on Known Phylogenetic Trees Using Poisson Regression" Proc. of the 5th European Conference on Computational Biology (ECCB-2007), Bioinformatics 23, e142-e147.

11. Endicot P, Ho SYW, "A Bayesian Evaluation of Human Mitochondrial Substitution Rates," *Am. J. Hum. Gen.* **82**, 895-902 (2008).

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

13. Bromham L, Penny D, "The Modern Molecular Clock," Nature Reviews Genetics, 4: 216-224 (2003).

14. Ho SYW, "Molecular Clocks: When Times are A-Changin'", Trends in Genetics, 22: 79-83 (2006).

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

16. Welch JJ, Bromham L, "Molecular Dating when Rates Vary," Trends in Genetics, 20: 320-327 (2005).

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

18. Wu C-I, Li W-H, "Evidence for higher rates of Nucleotide Substitution in Rodents than in Man," PNAS 82: 1741-1745 (1985).

19. Takezaki N, Rzhetsky A, Nei M, "Phylogenetic Test of the Molecular Clock and Linearized Trees," Mol. Biol. Evol. 12(5):823-833 (1995).

20. Sarich VM, Wilson AC, "Generation Time and Genomic Evolution in Primates," Science 179, 1144-1147 (1973).

21. Kogelnik, A.M., Lott, M.T., Brown, M.D., Navathe, S.B. & Wallace D.C. (1996) MITOMAP: a human mitochondrial genome database. Nucleic Acids Res. 24, 177-179.

22. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N, "Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA", Nature Genetics, 23(2), 147-147 (1999).

23. Myers E W, and Miller W. (1988) Optimal Alignments in Linear Space.Computer Applications in the Biosciences 4, 11-17.

24. Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16(6), 276-277.

25. Higgins D., Thompson J., Gibson T., Thompson J.D., Higgins D.G., Gibson T.J. "CLUSTAL W: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice." Nucleic Acids Research 22: 4673-4680 (1994).

26. 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.

27. 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).

28. Hernnstadt C, Preston G, Howell N, "Errors, Phantom and Otherwise, in Human mtDNA Sequences," Am. J. Hum. Genet. 72: 1585-6 (2003).

29. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N "Reduced-median-network analysis of complete mitochondrialDNAcoding-region sequences for the major African, Asian, and European haplogroups." Am J Hum Genet 70:1152–1170 (2002).

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

31. Anderson S, Bankier A T, Barrell B G, de Bruijn M H L, Coulson A R, Drouin J, Eperon I C, Nierlich D P, Roe B A, Sanger F, Schreier P H, Smith A J H, Staden R, and Young I G, (1981) "Sequence and organization of the human mitochondrial genome." Nature 290, 457–465.

32. Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, Quintana-Murci L, Tyler-Smith C, Wells RS, the Genographic Consortium, "The

Genographic Project Public Participation Mitochondrial DNA Database," PLoS Genetics 3, 1083-1095 (2007).

33. Ingman M, Gyllensten U, "Analysis of the Complete Human mtDNA Genome: Methodology and Inferences for Human Evolution," The Journal of Heredity, 92(6), 454-461 (2001).

34. Ingman M, Gyllensten U, "mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences," Nucleic Acids Research (Database Issue) 34: D749-D751 (2006).

35. Bandelt H-J, Kong Q-P, Richards M, Macaulay V "Estimation of mutation rates and coalescence times: some caveats," In: Bandelt H-J, Macaulay V, Richards M (eds). Human Mitochondrial DNA and the Evolution of Homo Sapiens. Springer-Verlag: Berlin Heidelberg. pp 47–90 (2006).

36. 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 H-J, Oppenheimer S, Torroni A, Richards M, "Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes," Science 308: 1034-1036 (2005).

37. W. M. Fitch (1971), "Toward defining the course of evolution: defining the minimum change for a specific tree topology," Systematic Zoology 20, 406-416.\

38 Cutler DJ, "Understanding the Overdispersed Molecular Clock," Genetics 154: 1403-1417 (2000).

39. Cutler DJ, "Estimating Divergence Times in the Presence of an Overdispersed Molecular Clock," Mol. Biol. Evol. 17(11): 1647-1660 (2000).

40. Hasegawa M, Cao Y, Yang Z, "Preponderance of Slightly Deleterious Polymorphism in Mitochondrial DNA: Nonsynonymous/Synonymous Rate Ratio is Much Higher Within Species than Between Species," Mol. Biol. Evol. 15(11): 1499-1505 (1998).

41. Kimura M, "*The Neutral Theory of Molecular Evolution*" Cambridge University Press, Cambridge, UK (1983).

42. Behar et al., "The Dawn of Human Matrilineal Diversity," *AJHG* **82**, 1-11(2008), doi:10.1016/j.ajhg.2008.04.002

Appendix A

Introduction

The purpose of this section is to consider premises necessary for the description of the molecular clock, and to determine whether and in what ways a constrained Poisson regression constitutes an effective test of the molecular clock hypothesis for the human mtDNA phylogeny.

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.[3,4,5,6]

Among these are population effects [7,8,9] 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 [11,12,13], 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.

Conditions leading to Markov Processes

Cells, including the germ line, contain multiple mitochondria, among which various mitochondria may bear different mtDNA sequences, a condition called heteroplasmy.[15] In any organism, the distribution of alleles may vary from tissue to

tissue. Each ovum represents a population bottleneck. From generation to generation in a population, drift and selection impacts the measured variation in the population.

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 $p(x_s,t) = P\left(X_s(t) = x_s\right)$, 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

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

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 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.

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 $p\left(x_s,t\middle|x_s',t_0\right)$. Specifically, no cyclic behavior, and the distribution evolves with a diffusive character. Phenomenologically, the assumption being made is that a time scale

δt describes a time scale very small compared to the time over which $p\left(x_s,t\middle|x_s',t_0\right)$ varies, but long compared to the time that defines a stable presence of a substitution in the population. However, there are a number of processes acting, two of which being selection and drift. Selection operates on a characteristic time scale $\tau \approx 1/s$, while drift operates on a time scale $\tau \approx N_e$. If the population is small, drift may dominate over selection: more deleterious substitutions may survive in that smaller population more probably than in larger ones. Further, those lineages with more deleterious substitutions are more likely to die out over time within any haplogroup, leaving more depleted nonsynonymous substitutions for older haplogroups than younger subhaplogroups. Even if bottlenecks and varying population sizes in the history of a haplogroup did not play a role, the formation of subhaplogroups may probe time scales at some $\tau < \delta t$ assumed for selection to have uniformly weeded out nonsynonymous deleterious substitutions to some equilibrium threshold. Some of those haplotypes in leaf nodes bearing deleterious substitutions may be selected out later, when the node moves to the interior of the tree. The Poisson processes (described later) will measure these rates as they *appear* now. The result is that rates measured for the same events in the haplogroup (node) will appear to change over time as that haplogroup ages and moves deeper in the tree. The tree can probe on a time scale $\tau < \delta t$, smaller than the time scale assumed for the limit to be a valid description.

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

$$\frac{\partial p\left(i,t\middle|j,t_0\right)}{\partial t} = \sum_{k\neq i}\left[A_{ik}(t)\,p\left(k,t\middle|j,t_0\right) - A_{ki}(t)\,p\left(i,t\middle|j,t_0\right)\right],$$

a form called the "master equation."

Stationary 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,[16] Kimura 2-Parameter model,[17] Felsenstein '81 model,[18] Felsenstein '84 model,[18] Hasegawa, Kishino, and Yano '85 model,[19] Tamura-Nei '92 model,[20] Tamura-Nei '93 model,[21] and the General Time Reversible model (REV/GTR).[22,23] 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 such a condition 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 $u_i(t)=\left[p(i,t|k,t_0)-\pi_i\right]\big/\pi_i$. Then

$$\pi_i\frac{\partial u_i}{\partial t}=\sum_{k\neq i}\pi_i S_{ik}\pi_k\left(u_k-u_i\right)$$

Then construct

$$\frac{\partial}{\partial t}\left(\sum_i \pi_i u_i^2\right)=2\sum_{i,k,i\neq k}\pi_i S_{ik}\pi_k u_i\left(u_k-u_i\right)=-\sum_{i,k,i\neq k}\pi_i S_{ik}\pi_k\left(u_k-u_i\right)^2.$$

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 solution $p\left(x_s,t\middle|x_s{}',t_0\right) = \pi_{x_s}$ being unique.

Interactions between sites

Consider the case where two sites s_1 and s_2 are *not* independent. In this case, 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 $p\left(x_{s_1},x_{s_2},t\right) = P\left(X_{s_1}(t) = x_{s_1} \cap X_{s_2}(t) = x_{s_2}\right) \neq p\left(x_{s_1},t\right)p\left(x_{s_2},t\right)$. In this case, the dependence of the joint distribution on time must satisfy

$$p\left(x_{s_1},x_{s_2},t\right) = \sum_{x_{s1}{}',x_{s2}{}'} p\left(x_{s_1},x_{s_2},t\middle|x_{s_1}{}',x_{s_2}{}',t_0\right)p\left(x_{s_1}{}',x_{s_2}{}',t_0\right)$$

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:

$$\frac{\partial p\left(x_{s_1},x_{s_2},t\right)}{\partial t} = \sum_{x_{s_1}{}',x_{s_2}{}'} A_{x_{s_1}x_{s_2}x_{s_1}{}'x_{s_2}{}'}(t)\cdot p\left(x_{s_1}{}',x_{s_2}{}',t\right)\cdot$$

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 $A_{x_{s_1}x_{s_1}'}(t) = \sum_{x_{s_2},x_{s_2}'} A_{x_{s_1}x_{s_2}x_{s_1}'x_{s_2}'}(t) \cdot p\left(x_{s_2}', t \middle| x_{s_1}', t\right)$.

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 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) double substitutions could be promoted with probabilities $O(\delta t)$ resulting in the effective total rates not being additive, 2) the contributions of individual rates depends on the value of other sites' nucleotides. Of these, the issue related to δt most directly is the promotion of simultaneous substitutions. In this case, a situation where such might be observed would be multiple mutations with selection promoting specific pairs in a time short compared to fixation.

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 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

$$P\left(N(t+\delta t) = n\right) = r\delta t \cdot P\left(N(t) = n-1\right) + (1 - r\delta t)P\left(N(t) = n\right).$$

The resulting differential equation satisfied by this is

$$\frac{\partial P\left(N(t)=n\right)}{\partial t} = r \cdot P\left(N(t)=n-1\right) - r \cdot P\left(N(t)=n\right).$$

The similarity to the master equation is very visible. If the conditions are stationary, the distribution of events is Poisson distributed

$$P\left(N(t)=n\right) = \frac{(rt)^n}{n!} e^{-rt}$$

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

$$P_n(t) = \sum_{\substack{n_1, n_2 \\ n_1+n_2=n}} \frac{(r_1 t)^{n_1}}{n_1!} e^{-r_1 t} \frac{(r_2 t)^{n_2}}{n_2!} e^{-r_2 t} = \frac{\left[(r_1 + r_2)t\right]^n}{n!} e^{-(r_1+r_2)t}.$$

In other words, the rates are cumulative.

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 transitions 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 $r = \sum_{s,i,j \neq i} A_{sij}\pi_{sj}$.

From the above argument, it is clear the result depends on the notion that r does not change significantly over the course of sampling. This could be true if the system had equilibrated. However, it could also be approximately true for constant A_{sij} for a time-slice short compared to the rate of change expressed in the Chapman-Kolmogorov equation. The observable distribution is not anywhere near this equilibrium given the short time to the human MRCA (less than 40 substitutions). That requires that sampling implied in the limit definition has occurred in a way so that the sample time is long compared to selection and population drift effects, that these effects were uniform over the phylogenetic tree, and that the time probed by phylogenetic branching is long compared to the drift and selection effects. Independence of substitution events between

sites is also required, but this can be examined explicitly. Therefore the presence of such selection/population interactions can be tested with a maximum likelihood Poisson estimation as a null hypothesis, assuming correlated substitutions can be eliminated.

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.[24,25] 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[24,25] $\Gamma(r;\alpha,\beta)dr = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r} dr$, 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

$$E(n) = \bar{r}t$$

$$\mathrm{var}(n) = \bar{r}t\left(1+\frac{\bar{r}t}{\alpha}\right)$$

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 $\alpha \propto S$, it follows that the total is likely not overdispersed. However, for the hypervariable regions, where S is significantly smaller and $\bar{r}$ is larger, it has been reported that the Γ distributed rates fits the observed data quite satisfactorily.[26] 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.

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, if a phylogeny is incomplete, and some bifurcations have been collapsed, 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

$$P_m(t) = \sum_{\substack{m_1, m_2 \\ m_1 + m_2 = m}} \frac{(rt_1)^{m_1}}{m_1!} e^{-rt_1} \frac{(rt_2)^{m_2}}{m_2!} e^{-rt_2} = \frac{\left[r(t_1 + t_2)\right]^m}{m!} e^{-r(t_1 + t_2)}$$

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.

Conclusions

The preceding shows that a Poisson distributed molecular clock can follow from non-correlated substitutions in $O(\delta t)$ resulting in additive rates, and rates that may vary, but whose variations are not correlated with each other. An assumption of time-independent rates might be suggested by the idea that substitutions along specific lineages are governed primarily by molecular rates, whereas selection will act to remove haplotypes bearing deleterious substitutions from the population. If such elimination is differential according to population size and fixation time effects relative to selection, or by environmental effects, substantial correlations between rates could be introduced that would cause deviations from Poisson processes. Therefore, a maximum likelihood Poisson regression applied to the phylogenetic tree may act as a test of a null hypothesis against which deviations can mark candidates for selection population, drift, and inter-site correlated substitutions.

Bibliography

---

1. Zukerkandl E, Pauling L, in "Horizons in Biochemistry," (Eds Kasha M and Pullman B), pp189-225 (Academic Press, NY, 1962).

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, "Do the Four Clades of the

mtDNA Haplogroup L2 Evolve at Different Rates?" Am. J. Hum. Genet. 69: 1348-1356 (2001).

3. Bromham L, Penny D, "The Modern Molecular Clock," Nature Reviews Genetics, 4: 216-224 (2003).

4. Ho SYW, "Molecular Clocks: When Times are A-Changin'", Trends in Genetics, 22: 79-83 (2006).
5. Hedges SB, Kumar S, "Genomic Clocks and Evolutionary Timescales," Trends in Genetics, 19: 200-206 (2003).

6. Welch JJ, Bromham L, "Molecular Dating when Rates Vary," Trends in Genetics, 20: 320-327 (2005).

7. Ohta T, Kimura M, "On the constancy of the Evolutionary Rate of Cistrons, " J. Mol. Evol 1: 18-25 (1971).

8. Ohta T, "Very Slightly Deleterious Mutations and the Molecular Clock," J. Mol. Evol. 26, 1-6 (1987).

9. Ohta T, "Near-neutrality in Evolution of Genes and Gene Regulation," PNAS 99, 16134-16137 (2002).

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

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, "Natural Selection Shaped Regional mtDNA Variation in Humans," PNAS, 100: 171-176 (2003).

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, Cavelli-Sforza L, Oefner PJ, "The Role of Selection in the Evolution of Human Mitochondrial Genomes," Genetics 172: 373-387 (2006).

13. Ingman M, Gyllensten U, "Rate Variation Between Mitochondrial Domains and Adaptive Evolution in Humans," Human Molecular Genetics, 16: 2281-2287 (2007).

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

15. Fan W, Waymire KG, Narula N, Li P, Rocher C, Coskun PE, Vannan MA, Narula J, MacGregor GR, Wallace DC, "A Mouse Model of Mitochondrial Disease Reveals Germline Selection Against Sever mtDNA Mutations," Science, 319: 958-962 (2008).

16. Jukes T and Cantor C: Evolution of protein molecules. In: Mammalian Protein Metabolism. Academic Press, New York, 1969.

17. Kimura M: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16(2):111–120.

18. Felsenstein J: Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol 1981;17(6):368–376.

19. Hasegawa M, Kishino H, and Yano T: Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22(2):160–174.

20. Tamura K: Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G_C-content biases. Mol Biol Evol 1992;9(4):678–687.

21. Tamura K and Nei M: Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 1993;10(3):512–526.

22. Yang Z: Estimating the pattern of nucleotide substitution. J Mol Evol 1994;39(1):105–111.

23. Zharkikh A: Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39(3): 315–329 (1994).

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

25. Yang Z: Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol 1994;39(3):306–314.

26. Bandelt H-J, Kong Q-P, Richards M, Macaulay V "Estimation of mutation rates and coalescence times: some caveats," In: Bandelt H-J, Macaulay V, Richards M (eds). Human Mitochondrial DNA and the Evolution of Homo Sapiens. Springer-Verlag: Berlin Heidelberg. pp 47–90 (2006).

Appendix B: Maximum Likelihood Poisson Regression

The method 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.[1] 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.[2] Ancestral states were inferred using a modified Sarich-Wilson algorithm,[3] 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

$$L_K(\vec{x} + \delta\vec{x}) = L_K(\vec{x}) + \vec{b} \cdot \delta\vec{x} + \frac{1}{2}\delta\vec{x} \cdot C \cdot \delta\vec{x} + O(dx^3)$$
$$= L_K(\vec{x}) - \frac{1}{2}\vec{b} \cdot C^\circ \cdot \vec{b} + \frac{1}{2}(\delta\vec{x} + C^\circ \cdot \vec{b}) \cdot C \cdot (\delta\vec{x} + C^\circ \cdot \vec{b}) + O(dx^3)$$

where

$$\vec{b} = \nabla_{\vec{x}} L_K$$
$$C = \nabla_{\vec{x}} \nabla_{\vec{x}} L_K.$$

and

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

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 $\delta\vec{x}$. These vectors and matrices are computed via Richardson's extrapolation.[6] This maximizes where

$$\delta\vec{x} = -C^\circ \cdot \vec{b}.$$

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

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

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

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

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. Wu C-I, Li W-H, "Evidence for higher rates of Nucleotide Substitution in Rodents than in Man," PNAS 82: 1741-1745 (1985).

2. Takezaki N, Rzhetsky A, Nei M, "Phylogenetic Test of the Molecular Clock and Linearized Trees," Mol. Biol. Evol. 12(5):823-833 (1995).

3. Sarich VM, Wilson AC, "Generation Time and Genomic Evolution in Primates," Science 179, 1144-1147 (1973).

4. S. Rosset (2006), "Efficient Inference on Known Phylogenetic Trees Using Poisson Regression" Proc. of the 5th European Conference on Computational Biology (ECCB-2007), Bioinformatics 23, e142-e147.

5. W. M. Fitch (1971), "Toward defining the course of evolution: defining the minimum change for a specific tree topology," Systematic Zoology 20, 406-416.

6. Richardson, L. F. (1927). "The deferred approach to the limit". Philosophical Transactions of the Royal Society of London, Series A 226: 299–349.