Genomes of Animal Mitochondria have not Evolved According to Common Models of Substitution Probabilities

Abstract

Common probabilistic models of substitutions of bases (Jukes-Cantor, Kimura 2-parameter, Tamura-Nei, F84, HKY, and the 6-parameter models used in linear invariants methods) must be rejected, at least for mitochondrial genomes of animals. They are rejected by a new test that is simple and lenient.

1. Introduction

It would be impossible to infer any aspect of an evolutionary tree from nucleic acid sequences without some model of the evolution of sequences. Here I treat the Jukes-Cantor model (Felsenstein 2004, p. 156), the Kimura two parameter model (Felsenstein 2004, p. 196), the Tamura-Nei model, F84, and HKY (Felsenstein 2004, p. 200 *et seq.*). These are all special cases of models used in linear invariants methods. For mitochondrial genomes of animals, I will reject the less general models by rejecting linear invariants models.

People who derive methods of phylogenetic inference from modeling assumptions are showing that those assumptions are sufficient. They sometimes don't investigate whether their assumptions are necessary. Here I show that if any of the models just mentioned is necessary for a method to succeed, then that method will not succeed, at least for mitochondrial genomes of animals.

The method of linear invariants for inferring evolutionary trees from nucleic acid sequences has the advantage that it depends for its success on very weak assumptions about the way the sequences evolve. The mathematical theory is presented in Lake 1987, Cavender 1989, 1991, and Nguyen and Speed 1992. (But Lemma 2 of Cavender 1991 is wrong.)

2. Necessary Conditions of Linear Invariants

In this paper, I treat two-species trees with a root species **R** and leaf species **S** and **T**. For a particular site in a nucleic acid sequence, the Markov matrix,

$$P_{RS} = \begin{pmatrix} p_{AA} & p_{AG} & p_{AC} & p_{AU} \\ p_{GA} & p_{GG} & p_{GC} & p_{GU} \\ p_{CA} & p_{CG} & p_{CC} & p_{CU} \\ p_{UA} & p_{UG} & p_{UC} & p_{UU} \end{pmatrix}$$

says that the conditional probability of finding G in **S** given A in the root **R** is p_{AG} and similarly for the other components of the matrix. (I am using the convention that U means “U or T, depending on your acid”). A matrix P_{RT} is defined analogously. The linear invariants method requires these matrices to have the form

$$\begin{pmatrix} e & f & g & g \\ h & i & j & j \\ k & k & l & m \\ n & n & p & q \end{pmatrix} \quad (1)$$

and satisfy the two constraints,

$$e - f = i - h \quad (2)$$

and

$$l - m = q - p \quad (3)$$

Each site in the sequence may have its own Markov matrices P_{RS} and P_{RT} satisfying (1), (2), and (3). The probabilities of A, G, C, and U in the root may be anything. Different sites are assumed to be statistically independent. Yes, (1), (2), and (3) look strange, arbitrary, and implausible at first. But the popular Kimura two-parameter model satisfies (1), (2) and (3), so if you can accept that model you may not reject this much less restrictive one. And mathematically (1), (2) and (3) are not arbitrary. They specify a maximal semigroup of Markov matrices under which linear invariants will work, that is, necessary conditions. This the *balanced transversion* model.

There is a family of alternatives, the *nonbalanced transversion* models (or “beyond balanced transversion” models). In these, (1), (2) and (3) are replaced by

$$\begin{pmatrix} e & f & g & \alpha g \\ h & i & j & \alpha j \\ k & \beta k & l & m \\ n & \beta n & p & q \end{pmatrix},$$

$$\beta e - f = i / \beta - h,$$

and

$$\alpha l - m = q / \alpha - p$$

where α and β are positive constants. From Felsenstein 2004 (equation 13.11, P. 203) one can deduce that the Tamura-Nei model, F84, and HKY are nonbalanced models.

3. Invariants

Two aligned sequences determine a *spectrum*,

$$S = \begin{pmatrix} s_{AA} & s_{AG} & s_{AC} & s_{AU} \\ s_{GA} & s_{GG} & s_{GC} & s_{GU} \\ s_{CA} & s_{CG} & s_{CC} & s_{CU} \\ s_{UA} & s_{UG} & s_{UC} & s_{UU} \end{pmatrix}$$

in which s_{AG} is the number of sites in the alignment where an A in the first sequence aligns with a G in the second, and similarly for the other components. With (1), (2), and (3) in force, the forms

$$s_{AC} - s_{AU} - s_{GC} + s_{GU} \quad (4)$$

and

$$s_{CA} - s_{UA} - s_{CG} + s_{UG} \quad (5)$$

are random variables with mean zero. In more elaborate language, let

$$V_1 = \begin{pmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad V_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \end{pmatrix}$$

and define an inner product by

$$\langle V, S \rangle = \sum_X \sum_Y (V_{XY} S_{XY}), \quad X, Y \in \{A, G, C, U\}.$$

Then (4) is $\langle V_1, S \rangle$, (5) is $\langle V_2, S \rangle$, and V_1 and V_2 are two *linear invariants* that form a basis of the two-dimensional vector space of all linear invariants of the two-species tree. The nonbalanced transversion models also have linear invariants.

4. Results

I studied mitochondrial genomes because these appear to be long enough to provide statistical significance.

Table 1 shows the spectrum for the complete mitochondrial genomes of two species of Cnidaria. Species S and T are the brown hydra *Hydra oligactis* (GenBank Accession NC_010214) and the brown sea anemone *Metridium senile* (NC_000933.1). I copied

the sequences from GenBank and used the Smith-Waterman algorithm of EMBOSS

(https://www.ebi.ac.uk/Tools/psa/emboss_water/nucleotide.html)

to align them. The entry 275, for example, means that at 275 sites, an A in the *Hydra* sequence aligned with a C in *Metridium*. For this spectrum, (4) is -54 , not zero. If (4) and (5) are significantly far from zero, I must reject (1), (2), and (3) as a hypothesis.

In more correct terms, I want to test whether the *Hydra-Metridium* spectrum lies significantly far from the subspace that is orthogonal to V_1 and V_2 . To do this, set $Y = s_{AC} + s_{GU}$ and $n_Y = s_{AC} + s_{AU} + s_{GC} + s_{GU}$ so that under the null hypothesis Y is binomially distributed with parameters n_Y and $1/2$. Set

$$\Phi_Y = \frac{(Y - n_Y/2)}{\sqrt{n_Y/4}}$$

so that Φ_Y is approximately standard normal. Compute Z and Φ_Z from $s_{CA} + s_{UG}$ and $s_{CA} + s_{UA} + s_{CG} + s_{UG}$ similarly. Set $\chi^2 = \Phi_Y^2 + \Phi_Z^2$ to form a chi-square test statistic with two degrees of freedom. In this case, $\chi^2 = 4.505$ which is not large enough for rejecting the null hypothesis. (The rejection levels for a chi-square test with 2 degrees of freedom are 5.99 and 9.21 at the 5% and 1% levels, respectively.) Strictly, the modeling assumption that each base in the sequence has its own (1), (2), and (3) is not compatible with this statistical analysis. The distributions of Y and Z are mixtures of binomials rather than binomial. As a practical matter, this means the power of the test is reduced but the size is not impaired. That is, acceptances should not be trusted but rejections can be. (But this is ordinarily the case with hypothesis tests.)

Repeating this analysis with the freshwater jellyfish *Craspedacusta sowerbyi* (JN593332) and the jellyfish *Rhopilema esculentum* (NC_035740) gives $\chi^2 = 23.812$. This is decisive grounds for rejection. But this spectrum exactly satisfies the invariants of the nonbalanced transversion model with $\alpha = 3.83$ and $\beta = 0.188$ and also for the model with $\alpha = 0.6642$ and $\beta = 1.1156$. Two nonbalanced models like these can be found for most spectra, as follows.

From Cavender 1991, page 73, or Cavender 1989, page 313, these matrices are a basis of the invariant space under the nonbalanced model:

$$V_1 = \begin{pmatrix} 0 & 0 & \alpha\beta & -\beta \\ 0 & 0 & -\alpha & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad V_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \alpha\beta & -\alpha & 0 & 0 \\ -\beta & 1 & 0 & 0 \end{pmatrix}.$$

Thus we seek α and β that, for the given s , satisfy

$$s_{AC}\alpha\beta - s_{AU}\beta - s_{GC}\alpha + s_{GU} = 0 \quad (6)$$

$$s_{CA}\alpha\beta - s_{UA}\beta - s_{CG}\alpha + s_{UG} = 0 \quad (7)$$

this system is easy to solve if s_{AC} or s_{CA} is 0. Otherwise, eliminate $\alpha\beta$ by adding equations. That is, add

$$-s_{CA}(s_{AC}\alpha\beta - s_{AU}\beta - s_{GC}\alpha + s_{GU}) = 0$$

to

$$s_{AC}(s_{CA}\alpha\beta - s_{UA}\beta - s_{CG}\alpha + s_{UG}) = 0$$

obtaining

$$(s_{CA}s_{AU} - s_{AC}s_{UA})\beta + (s_{CA}s_{GC} - s_{AC}s_{CG})\alpha - (s_{CA}s_{GU} - s_{AC}s_{UG}) = 0$$

or $h\beta + k\alpha = j$ with h, k , and j defined in the obvious way. Then

$$\beta = \frac{j - k\alpha}{h} \quad (8)$$

Substituting (8) into (6) gives

$$s_{AC}\alpha \frac{j - k\alpha}{h} - s_{AU} \frac{j - k\alpha}{h} - s_{GC}\alpha + s_{GU} = 0,$$

a quadratic equation to solve for α . Finally, recover β using (8). But the quadratic equation that gives two solutions might not always give real, positive α and β , if only because spectra are random objects.

The chi-square computation just given for the balanced transversion model can be modified for use with nonbalanced transversion models. Is there a nonbalanced transversion model under which neither of these Cnidarian spectra is rejected by the chi-square test? No. The best is surprisingly close to balanced, with $\alpha=0.842$ and $\beta=0.963$ but has $\chi^2=37.94$ for both spectra, requiring rejection at the 1% level. Fudging parameters to minimize a χ^2 statistic is another dubious procedure. Again, it reduces power but does not impair size.

This analysis of Cnidaria can be repeated for any taxon containing four species whose mitochondrial genomes have been sequenced. I have done it for 11 phyla, proposed phyla and former phyla. The results are shown in Table 2. The balanced model is rejected for almost every phylum. I hoped to find an α and β that would work throughout Metazoa, but they don't exist. No consistent α and β to account for evolution in the various phyla showed up. No linear invariants model is correct. (For Ctenophora no credible nonbalanced model fit both spectra. The minimum-seeking algorithm halted while pursuing α and β to ridiculous extremes. This is doubtless because the mtDNA of *Mnemiopsis leidyi* is anomalous, as Pett, *et al.* 2011 discovered.)

5. Discussion

The failure of mtDNA to comply with linear invariants models is not because different parts of the DNA evolve at different rates and it is not because different rates apply on different edges of the tree. The linear invariants models allow that. (The tests in Goldman 1993 do not.) The rejections happened because (1), (2), and (3) are false hypotheses, or possibly because of dependence among substitutions at different sites.

One might object that it is not plausible that the balanced model is *exactly* right, and therefore any sufficiently powerful test is bound to reject it. But my test is hardly powerful! Moreover, Table 2 shows that perfect balance is more than a little false, according to my best estimates.

The models that I have rejected, all proposed before 1992, look like very sound applied probability. It is surprising to find that molecular evolution is so non-uniform, especially in mitochondria, so uniform in their environments and functions. These 22 spectra do not tell us whether their divergence from reasonable models indicates grand, stately flows, or rapid, wide fluctuations. With a careful choice of mitochondrial genomes, it should be possible to tell. Many of the investigators who produced the 44 sequences that I have studied here did so in order to clarify phylogenetic relations. This might be a futile program.

Mitochondrial genomes of animals are CG-poor, as seen in Table 1, for example. To maintain this imbalance over time requires small α and large β , if in each edge of the tree all sites in the sequences are governed by the same Markov matrix. Since this is not seen, I conclude that a large majority of the Metazoan mitochondrial sequence is CG-poor and very stable. The minority is most of what I have analyzed here. This is a new line of evidence for old knowledge (Aquadro, Kaplan and Risko 1984).

Table 1
A spectrum

					First sequence
	AA	AG	AC	AU	A
	2899	439	275	355	3968
	25.33%	3.84	2.40	3.10	34.67
	GA	GG	GC	GU	G
	84	1066	65	91	1306
	0.73	9.31	0.57	0.80	11.41
	CA	CG	CC	CU	C
	72	83	859	103	1117
	0.63	0.73	7.51	0.90	9.76
	UA	UG	UC	UU	U
	317	340	343	4054	5054
	2.77	2.97	3.00	35.42	44.16
Second sequence:	3372	1928	1542	4603	11445
	29.46	16.85	13.47	40.22	

Table 2: Fit of models to spectra

	Phylum	Species	GenBank	Balanced chi-square	Unbalanced alpha	Unbalanced beta	Unbalanced chi-square
	Acanthocephala	Centrorhynchus aluconis	NC_029765	26.21	6.107	0.825	0.0778
		Plagiorhynchus transversus	NC_029767				
		Oncicola luehei	NC_016754.1	55.75			
	Bryoazoa	Leptorhynchoides thecatus	AY562383		0.941	0.613	2.188
		Tubulipora flabellaris	EU563937	0.481			
		Bugula neritina	AY690838				
	Cnidaria	Celleporella hyalina	NC_018344	19.21	0.841	0.962	0.125
		Flustrellidra hispida	DQ157889				
		Craspedacusta sowerbyi	JN593332	23.812			
	Ctenophora	Nemopilema nomurai	NC_035740		*	*	*
		Hydra oligactis	NC_010214	4.505			
		Metridium senile	NC_000933				
	Platyhelminthes	Coeloplana loyai	LN898113	3.454	3.467	1.004	0.504
		Pleurobrachia bachei	NC_016697				
		Mnemiopsis leidyi	NC_016117	178.371			
	Nemertea	Vallicula multiformis	LN898115		10.41	0.693	0.41
		Echinococcus multilocularis	NC_000928	0.755			
		Trichobilharzia szidati	NC_036411				
	Mollusca	Opisthorchis felinus	EU921260	27.824	1.382	0.784	0.326
		Explanatum explanatum	NC_027958				
		Emplectonema gracile	JF727825	35.217			
	Annelida	Prosadenoporus spectaculum	NC_023931		2.205	0.488	1.761
		Lineus alborostratus	JN234382	6.577			
		Iwatanemertes piperata	NC_023523				
	Arthropoda	Pyramidella dolabrata	AY345054	9.85	1.693	0.585	4.587
		Ascobulla fragilis	AY345022				
		Siphonaria pectinata	AY345049	5.149			
	Chaetognatha	Myosotella myosotis	AY345053		0.626	0.349	0.111
		Erpobdella japonica	NC_036150	174.063			
		Eusyllis blomstrandii	KX752423				
	Chordata	Whitmania laevis	KM655839	106.457	1.215	0.846	36.924
		Hirudo nipponia	KC667144				
		Pelecinus polyturator	KM104167	26.057			
	Chordata	Tigriopus californicus	NC_008831		0.626	0.349	0.111
		Scutigera coleoptrata	AJ507061	102.683			
		Nephila clavata	AY452691				
	Chordata	Zonosagitta nagae	AP011545	8.22	1.215	0.846	36.924
		Spadella cephaloptera	NC_006386				
		Sagitta enflata	AP011547	39.293			
	Chordata	Paraspadella gotoi	NC_006083		1.215	0.846	36.924
		Diomedea chrysostoma	AP009193	17.767			
		Rattus rattus	NC_012374				
	Chordata	Epigonichthys cultellus	KC896827	63.859	1.215	0.846	36.924
		Clavelina lepadiformis	FJ839918				

References

Aquadro CF, Kaplan N, Risko KJ 1984, An analysis of the dynamics of mammalian mitochondrial DNA sequence evolution, *Mol Biol Evol* 1(5):423-434.

Cavender JA 1989, Mechanized derivation of linear invariants. *Mol Biol Evol* 6(3):301-316.

Cavender JA 1991, Necessary conditions for the method of inferring phylogeny by linear invariants. *Math Biosc* 103:69-75.

Felsenstein J. Inferring phylogenies. 2004. Sunderland: Sinauer Associates.

Goldman N 1993, Statistical tests of models of DNA substitution. *J Mol Evol* 36:182-198.

Lake JA. 1987, A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Mol Biol Evol* 4:167-191.

Nguyen T, Speed T 1992, A derivation of all linear invariants for a nonbalanced transversion model. *J Mol Evol* 35(1):60-76.

Pett W, Ryan JF, Pang K, Mullikin JC, Martindale MQ, Baxevanis AD, Lavrov DV 2011, Extreme mitochondrial evolution in the ctenophore *Mnemiopsis leidyi*: Insight from mtDNA and the nuclear genome, *Mitochondrial DNA* 22(4): 130-142.