Edmond J. Safra
Bioinformatics Program

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Factors Identification for Mitochondrial DNA Substitution Models

Keren Levinstein Hallak and Saharon Rosset
Department of Statistics and Operations Research, School of Mathematical Sciences

Introduction

Key Questions

Which factors are significant in determining:

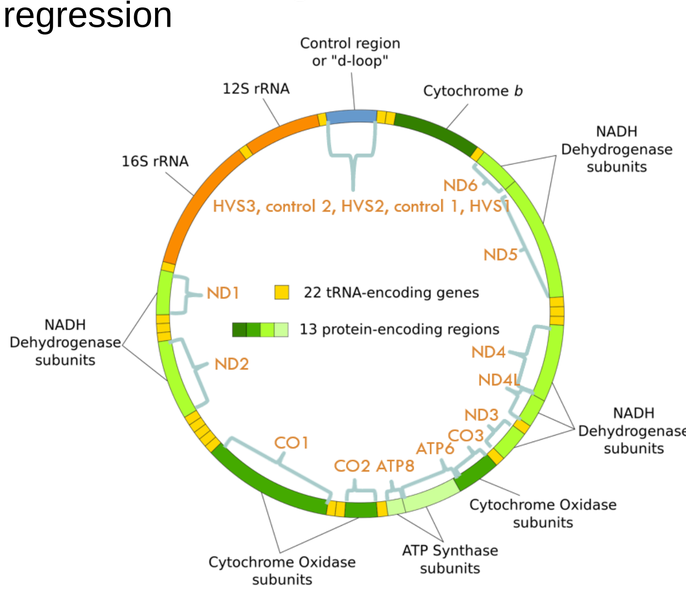
1. The substitution rate?
2. The substitution type?
   - Transition vs. transversion
   - Synonymous vs. non-synonymous

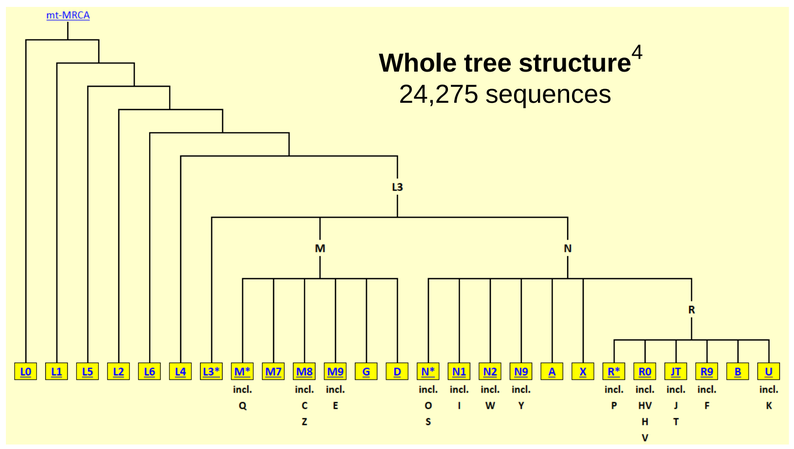
Relevant candidates:

Region/sub region, site, time, nucleotide, neighbor nucleotides, codon, codon position, amino acid, |synonymous amino acid group|, clade, CpG dinucleotides

Assumption:

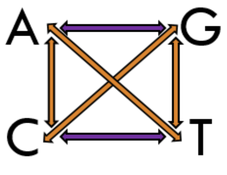
If all relevant factors are known then the substitution rate can be modeled via a Poisson regression, and the substitution type via a logistic regression





L4 Subtree

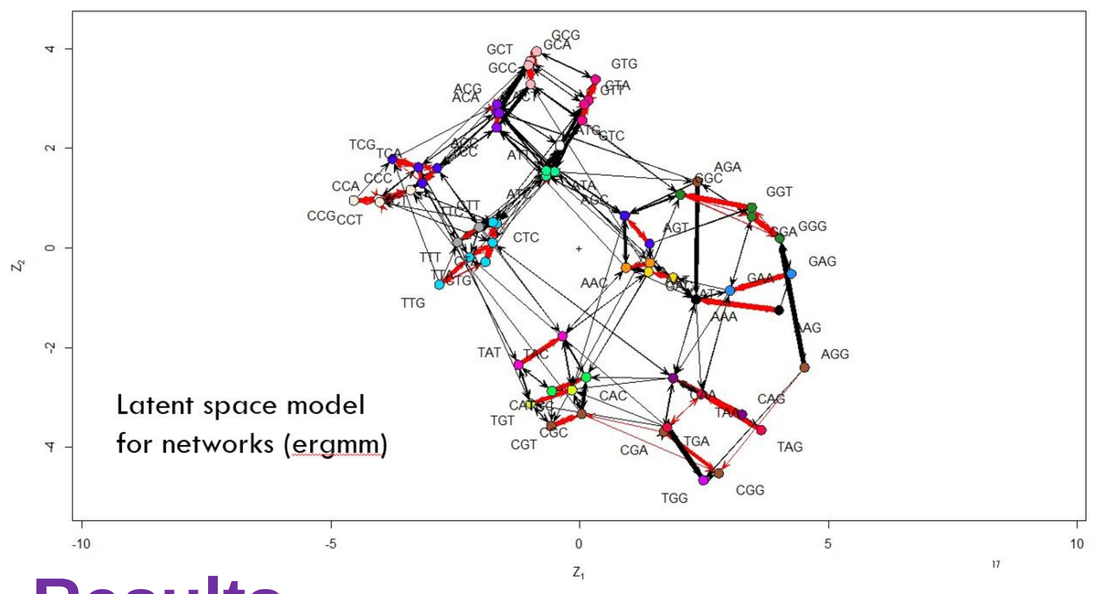
General Statistics



	A	G	C	T
A		5270	639	150
G	4830		79	51
C	299	62		5986
T	97	58	6760	

		# of events	mean	std	median	0.75 quantile	max (100% quantile)
coding region	transitions	14633	0.9473037	3.102829	0	1	103
	transversions	619	0.04007251	0.3348809	0	0	19
	total	15252	0.9873762	3.153734	0	1	103
control region	transitions	8213	7.319964	36.54745	0	3	1035
	transversions	816	0.7272727	10.30587	0	0	315
	total	9029	8.047237	38.10185	0	4	1036

Synonymous vs. non-synonymous mutations



Results

Sub-regions tests

a. Test the distribution of the bases using a two-sample χ^2-test
b. Test the number of substitutions at each site using Kruskal-Wallis
c. Negative binomial fit for the number of substitutions in both sub-regions compared together vs. apart (GLR)
d. Negative binomial regression for the number of substitutions with the sub-region as an explaining variable
e. Compare the distribution of codons using a two-sample χ^2-test (only relevant for codon locations)

	rRNA12S	rRNA16S	Different
rRNA12S	X		b,c,d
rRNA16S		X	b,c,d
Different	b,c,d	b,c,d	X

	Control 1	Control 2	HVS1	HVS2	HVS3
Control 1	X	d	c	b,d	a,b
Control 2	d	X	b,c,d	b,c,d	a,b
HVS1	c	b,c,d	X	b,d	a,b,c,d
HVS2	b,d	b,c,d	b,d	X	a,c,d
HVS3	a,b	a,b	a,b,c,d	a,c,d	X

	ND1	ND2	ND3	ND4	ND5	ND4L	ND6	CO1	CO2	CO3	ATP8	ATP6	CYB
ND1	X						a,e				b,d	b,d	
ND2		X					e	a			b,d	b,d	
ND3			X				a				b,d	b,d	
ND4				X			e	a			b,d	b,c,d	b,c,d
ND5					X		a,e	a			b,d	b,d	
ND4L						X					b,d	b,c,d	b,c
ND6	a,e	e	a	e	a,e		X	a,e	a	a,e	b,d	b,d,e	a,e
CO1		a		a	a		a,e	X			a,b,d	b,c,d	b,c,d
CO2							a		X		a,b,d	b,d	b,d
CO3							a,e			X	a,b,d	b,d	
ATP8	b,d	b,d	b,d	b,d	b,d	b,d	b,d	a,b,d	a,b,d	a,b,d	X		
ATP6	b,d	b,d	b,d	b,c,d	b,d	b,c,d	b,d,e	b,c,d	b,d	b,d		X	b,d
CYB				b,c,d		b,c	a,e	b,c,d	b,d			b,d	X

Conclusions

- We have yet to find a direct Poisson regression model for the number of substitutions per site, suggesting that there are missing relevant factors
- Adding relevant factors results in a higher theta parameter for the Negative Binomial regression (as well as higher likelihood)
- It is beneficiary (GLR wise) to separate the substitution model according to the codon position

Bibliography

1. Behar, D. M., van Oven, M., Rosset, S., Metspalu, M., Loogväli, E.-L., Silva, N. M., … Villems, R. (2012). A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *American Journal of Human Genetics*

2. Rosset, S. (2007). Efficient inference on known phylogenetic trees using Poisson regression. *Bioinformatics*, *23*(2), 142–147.

3. Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., … Richards, M. B. (2009). Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *American Journal of Human Genetics*, *84*(6), 740–759.

4. Van Oven, M. (2015). PhyloTree Build 17: Growing the human mitochondrial DNA tree.

5. Van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, *30*(2), E386–E394.