Co-evolution between nuclear and mitochondrial genes in the Drosophila genus; a computational analysis

Stefanos Papadantonakis

July 27, 2019

Contents

1 Introduction

1.1 Co-evolution

The term co-evolution is generally used to describe reciprocal evolutionary influence between two or more lineages. The term can be tracked back to *On the Origin of Species*, where Darwin describes the ecological interactions between numerous species and the effect on their phenotypic adaptations (this is commonly described as the Entangled Bank Hypothesis). The article "When Is It Co-evolution?" [13] is one of the first attempts to distinguish between different types of co-evolutionary thinking, and, strictly speaking, define co-evolution. The term however is used in many different ways depending on the discipline of the publication. The field of Molecular co-evolution is fairly recent and refers to evolutionary influences between proteins and protein families. Pollack [1] describes molecular co-evolution as the influence of an amino acid substitution at one position of a sequence to the substitution rate of other positions in the sequence. Since then, various studies have developed methods in order to detect intra-sequence, co-evolving residues. These methods try to infer coevolving signatures from DNA data and can be divided into two distinct groups: parametric and non-parametric [11, 15, 14, 8]. The parametric methods muster likelihood approximations (Maximum Likelihood or Bayesian approaches), phylogenetic information etc. Non-parametric models mainly employ mutual information (MI), an entropy based measurement taken from information theory. All of these studies, though, are performed on protein multi-sequence alignments (MSA), and try to find co-evolving amino-acid residues within the same protein. In this study, we expand the scope of the analysis on the DNA level and try to detect potentially co-evolving sites between different, interacting proteins.

1.2 Mitochondrial Oxidase Phosphorilation Complex

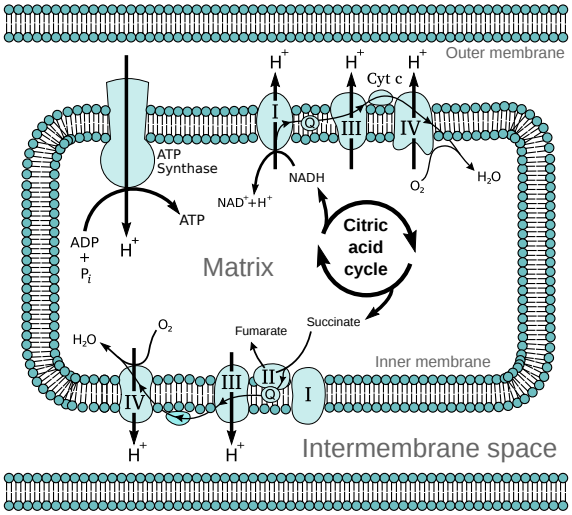

Figure 1: Mitochondrial Electron Transport Chain.

Oxidase Phosphorliation is a metabolic pathway present in almost all aerobic organisms, in which the cell oxidizes nutrients and uses the energy produced to form ATP. In eucaryotes Oxidase Phosphorilation usually takes place in the mitochondria, where electrons are transfered from electron donors to electron acceptors with a series of redox reactions. Specifically, free electrons travel within a series of protein complexes (OXPHOS complex) that form an electron transport chain. The energy that this transfer produces, is used to transport protons across the inner mitochondrial membrane, generating a pH gradient and an electric potential across the membrane. This potential is quenched by protons flowing into the mitochondrion through an ATP synthase. This flow causes a mechanical rotation in a part of the enzyme, that phosphorylates ADP into ATP [10]. Importantly, the mitochondrial DNA contains only a small number of genes needed to build the complexes of the electron transport chain. But these genes are susceptible to the special conditions that affect the evolution of mtDNA. First, mitochondria are more prone in fixing deleterious mutations because of their nearly maternal mechanism of inheritance (a.k.a. Muller 's Ratchet).

Second, it is hypothesized that mitochondria small but frequent bottlenecks even though most cells contain numerous mitochondrial genomes. These may come from the mode of inheritance or simply be a result of mitochondrial degradation due to cell stress [4].

1.3 Motivation

The OXPHOS complex is an ideal candidate for studying coevolution for two main reasons: a)The proteins that comprise the electron transport chain have subunits that originate from genes of mitochondrial and/or nuclear origin and b)The mutation rate in mitochondria is of an order of magnitude greater than the nuclei in animals. Thus, in order for these proteins' 3D structure to remain functional, either the nucleic loci have to compensate for the higher mitochondrial mutation rate by an increase of fixation of compensatory mutations, or the mitochondrial loci are subjects of very strong purifying selection. The idea of compensating mutations is not new. It is thought as an important factor influencing the evolution of mitochondria and even as a driver for speciation events [9]. Most of the studies, however, explore compensation within the same protein [17, 16, 23]. We employ a linkage disequilibrium (LD) product, T_2 [25], that describes correlation between two sites of the same or, in our case, different MSAs, in order to discover functional "linkage" between nuclear and mitochondrial sites.

2 Materials and Methods

2.1 Dataset

86 nuclear loci and 13 mitochondrial loci were obtained from UCSC database for 13 Drosophila species as well as their phylogenetic relationships.

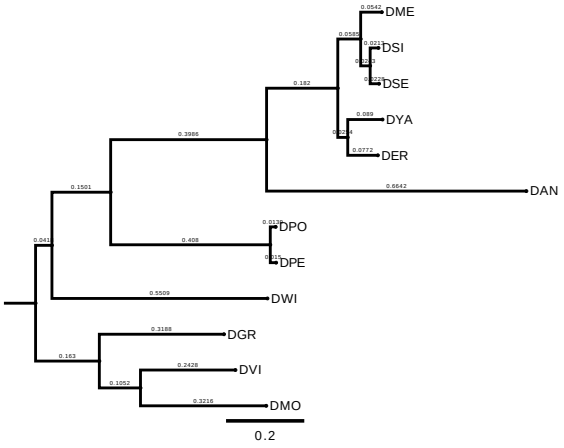

Figure 2: Phylogenetic tree of the 12 Drosophila species that were used in the analysis. The tree was generated by analysis of 4-fold degenerate sites.

2.2 Linkage Disequilibrium

We employed the T_2 formula as described in Zaykin et al., in order to calculate the correlation between nuclear and mitochondrial sites. With an unambiguous gametic phase, observations can be arranged into a kXm table. The cells of the table represent N haplotype observations: the (i, j)th cell is the number (n_{ij}) of haplotypes that carry allele i at the nuclear site and j at the mitochondrial one. The observed haplotype frequencies are $p_{ij} = \frac{n_{ij}}{N}$ and the observed correlation for sites i and j is:

$$r_{ij} = \frac{pij - p_i q_j}{\sqrt{p_i(1-p_i)q_j(1-q_j)}} \tag{1}$$

p_i and q_j are simply the haplotype frequencies at nuclear and mitochondrial sites respectively. The formula for the T_2 statistic approximates a X^2 distribution and is described as follows:

$$T_2 = \frac{(k-1)(m-1)N}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} r_{ij}^2 \sim X^2_{(k-1)(m-1)} \tag{2}$$

Only biallelic sites were used in the calculations, as mathematical inference that violates the infinite site model becomes tedious. All the biallelic sites of nuclear loci were compared with all the biallelic sites of mitochondrial loci. These comparisons comprise three possible categories: a) a site that has a synonymous polymorphism versus a site that also has a synonymous polymorphism, b) a site that has a synonymous polymorphism versus a site that has a non-synonymous polymorphism and c) two sites that both contain non-synonymous polymorphisms. For each pair of nuclear and mitochondrial loci we calculated the likelihood of observing this many or greater pair of sites that belong to category (c) among the sites that have the highest T_2 value. This likelihood was calculated under the hyper-geometric distribution.

2.3 Scripting and Code

Implementation of the T_2 formula was performed with C programming language. R packages "Phangorn" [20] and "PopGenome" [18] were used to describe whether biallelic polymorphisms were synonymous or not. The "Invertebrate Mitochondrial Code" [3] of NCBI was used to describe mitochondrial polymorphisms.

2.4 Hypothesis Testing

In order to find out whether coevolution actually affects the value of the statistic we are employing, we need to isolate all other parameters that may have an effect on these values (e.g. phylogeny, branch lengths, etc.). For this reason, we simulated codon sequences that do not interact with each other using the INDELible package [7]. Mitochondrial and nuclear codon rates and equilibrium frequencies, were inferred using the codeML program from PAML package [24], by concatenating the mitochondrial and nuclear loci respectively. While this method often produces biases under a phylogenetic perspective, we have already assumed that the phylogeny is known. Instead, we are trying to find the codon rates and equilibrium frequencies that have the maximum likelihood given the phylogenetic tree and the sequence data. The phylogenetic tree provided by UCSC was generated exclusively by 4-fold degenerate sites, that evolve faster than the other coding sites. This affects the branch lengths of the phylogeny of the species and consequently the number of polymorphisms in our supposedly neutral simulations. In order to evaluate the effect that the branch lengths have in the analysis, we simulated datasets, where we scaled the phylogenetic tree's branch lengths according to analysis performed by the RAxML [22] program.

3 Results and Discussion

3.1 Distribution of P-value statistic

The distribution of the Log10 p-values, as described in the 2.4 section, is shown in **Figure 3**. A striking result is the distance between the threshold generated from simulations under the UCSC phylogenetic tree and the one generated from simulations after the scaling with the RAxML best tree length. In particular, we found out that the total length of the UCSC tree length was 2.993864 times larger that the RAxML one. This practically means, that all the sites in the simulated dataset evolve just as fast as the 4-fold degenerate sites that were used to generate the tree. The result is inflated number of non-synonymous polymorphisms in the unscaled dataset, that cause the threshold to shift upwards. In contrast, the scaled dataset 's threshold appears to be too lenient compared to the real data. This is the reason we chose to operate with the stricter approach.

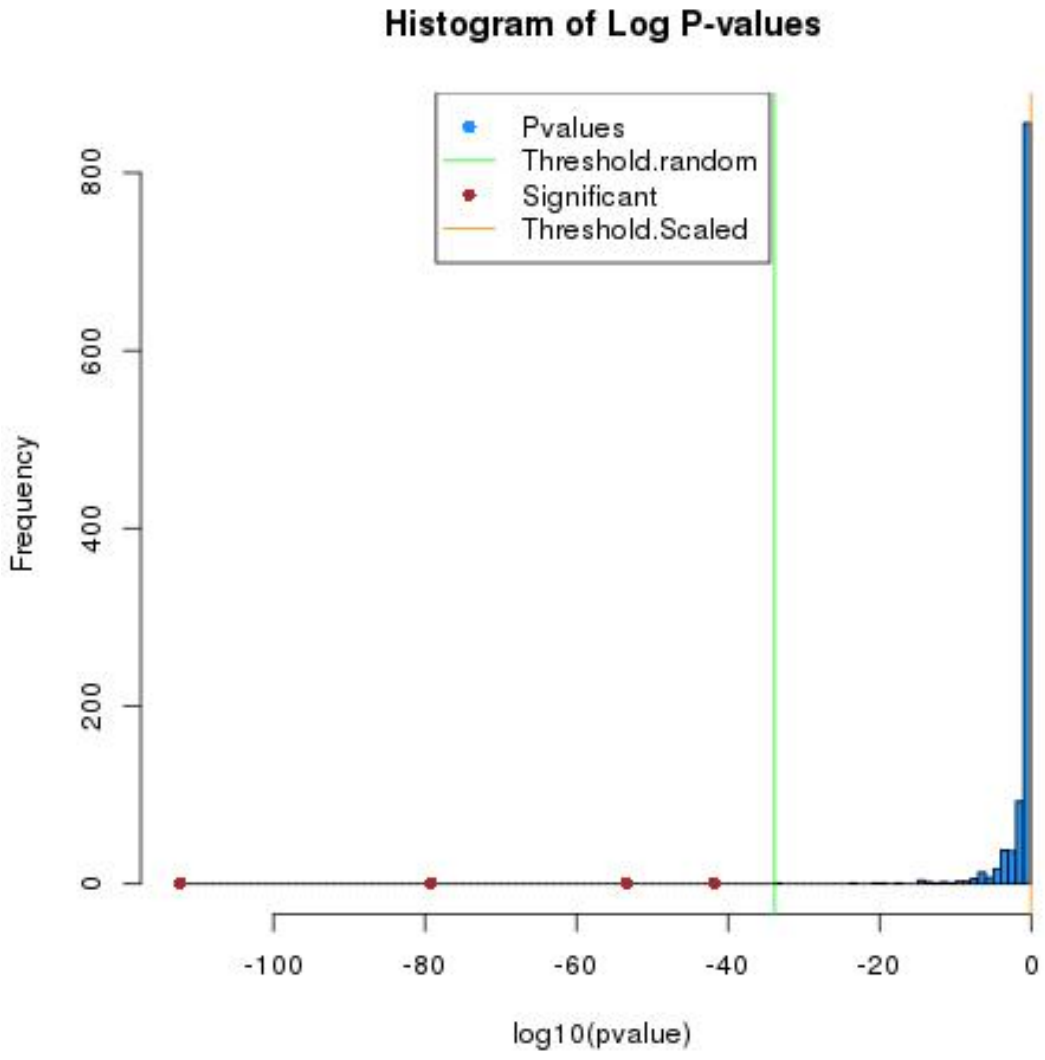

Figure 3: Histogram of Log-Pvalues. Green line represents the threshold generated from simulations under the tree generated from 4-fold degenerate sites. Orange line was generated from simulations under after the tree was scaled according to RAxML tree length. P-values represent the probability that the sites with the highest T_2 contain that many or more comparisons between sites that contain non-synonymous substitutions in both nucleic and mitochondrial loci, under the hypergeometric distribution.

3.2 Interactions

COX2 a gene that encodes the second subunit of the cytochrome oxidase apparatus. It is located in the mitochondrial DNA and is part of a membrane protein with 6 subunits, that comprise of products of the COX1 and COX2 genes. Cytochrome Oxidase2 is the main catalytic molecule of the enzyme and accepts 4 electrons from two cytochrome c units and transfers them to O_2 [21]. The **ATP6** gene encodes the 6th subunit of the F_1F_o ATP synthase. As a result, it is in close proximity with the other subunits of complex V. Mutations in the ATP6 gene have been associated with Leigh syndrome [6]. **CYTB** is the only component of respiratory complex III, that is encoded in the mitochondia. Cytochrome *b*, along with Cytochrome *c* and iron-sulfur protein (ISP) form the catalytic center of the enzyme. This complex is a membrane bound enzyme that transfers electrons from Ubiquinol to Cytochrome *c*. Mutations in the CYTB gene have been associated with different encephalopathies and myopathies [5]. **ISP** encodes a Rieske protein (iron-sulfur protein) that operates inside the complex III of the electron transport chain. Mutations of ISP1 gene are shown to result in delayed development and increased lifespan of *C.elegens*. It has been shown that this molecule passes electrons to the *cyt-b* module [12]. **ATPaseBeta** encodes part of the main catalytic subunit of the F_1F_o ATPase motor complex (complex V). Together with the a subunit, they form the ADP binding site of the enzyme [2]. **NDUFV1** is the nuclear gene that expresses the NADH Dehydrogenase Ubiquinone Flavoprotein 1. This protein is the main subunit of complex I and the NDUFV1 gene encodes the binding site of NADH that transfers electrons from NADH towards the electron transport chain [19].

Nuclear Locus	Mitochondrial Locus	Log10 P-value
ATPaseBETA	COX2	4.662501e-113
NDUFV1	ATP6	7.059310e-80
NDUFV1	CYTB	3.645180e-54
ISP	ATP6	1.526968e-42

Table 1: Table of comparisons that pass the significance threshold.

3.3 Conclusions

In this study, we have provided a novel framework for studying coevolutionary relationships between proteins on the DNA level. While our results may suggest significance over loci that do not interact physically, a lot of the structures of the gene products for the OXPHOS complexes remain hypothetical. Moreover, it has been shown that physical interaction is not necessary for sequences to develop coevolutionary signatures. Merely being in the same network, interacting with the same protein molecules, generates enough pressure and results in similarities [11]. A naive example is that of the binding sites of a specific transcription factor. They may have a high degree of similarity, even though they may be located in different chromosomes. Even though our methodology needs refining, it can be extended to detect pairs of interacting sited in MSAs, and will be extremely useful in discovering patterns of coevolution among other cyto-nuclear interacting systems, e.g. chloroplast-nucleus interactions of the photosynthetic mechanism.

4 References

References

[1] P. J. Angeline and J. B. Pollack. Competitive Environments Evolve Better Solutions for Complex Tasks. In S. Forrest, editor, *Genetic Algorithms: Proceedings of the Fifth International Conference (GA93)*. The Ohio State University, 1993.

[2] I. Antes, D. Chandler, H. Wang, and G. Oster. The Unbinding of ATP from F 1 -ATPase. *Biophysical Journal*, 85(August):695–706, 2003.

[3] J. L. Boore and W. M. Brown. Complete DNA Sequence of the Mitochondrial Genome of the Black Chiton, Katharina Tunicata. *Genetics*, 138(2):423–443, Oct 1994.

[4] J.-y. Chou and J.-y. Leu. The Red Queen in mitochondria: cyto-nuclear co-evolution, hybrid breakdown and human disease. *Frontiers in Genetics*, 6(May):1–8, 2015.

[5] S. Dasgupta, M. O. Hoque, S. Upadhyay, and D. Sidransky. Forced Cytochrome B gene mutation expression induces mitochondrial proliferation and prevents apoptosis in human uroepithelial SV-HUC-1 cells. *International Journal of Cancer*, 125(12):2829–2835, 2009.

[6] L. De Meirleir, S. Seneca, W. Lissens, E. Schoentjes, and B. Desprechins. Bilateral Striatal Necrosis With a Novel Point Mutation in the Mitochondrial ATPase 6 Gene. *Pediatric Neurology*, 13(3):242–246, 1995.

[7] W. Fletcher and Z. Yang. INDELible: A Flexible Simulator of Biological Sequence Evolution. *Molecular biology and evolution*, 26(8):1879–1888, 2009.

[8] H. Gao, Y. Dou, J. Yang, and J. Wang. New methods to measure residues coevolution in proteins. *BMC Bioinformatics*, 12(1):206, 2011.

[9] M. Gershoni, A. Fuchs, N. Shani, Y. Fridman, M. Corral-Debrinski, A. Aharoni, D. Frishman, and D. Mishmar. Coevolution predicts direct interactions between mtDNA-encoded and nDNA-encoded subunits of oxidative phosphorylation complex I. *Journal of Molecular Biology*, 404(1):158–171, 2010.

[10] M. Gershoni, L. Levin, O. Ovadia, Y. Toiw, N. Shani, S. Dadon, N. Barzilai, A. Bergman, G. Atzmon, J. Wainstein, A. Tsur, L. Nijtmans, B. Glaser, and D. Mishmar. Disrupting Mitochondrial–Nuclear Coevolution Affects OXPHOS Complex I Integrity and Impacts Human Health. *Genome Biology and Evolution*, 6(10):2665–2680, 2014.

[11] C.-s. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-evolution of Proteins with their Interaction Partners. *Journal of Molecular Biology*, pages 283–293, 2000.

[12] G. Jafari, B. M. Wasko, A. Tonge, N. Schurman, C. Dong, Z. Li, R. Peters, E.-b. Kayser, J. N. Pitt, P. G. Morgan, M. M. Sedensky, A. R. Crofts, and M. Kaeberlein. Tether mutations that restore function and suppress pleiotropic phenotypes of the C. elegans isp-1(qm150) Rieske iron – sulfur protein. *PNAS*, 1(13):6148–6157, 2015.

[13] D. H. Janzen. When is it Coevolution? *Evolution*, 34(3):611–612, 1980.

[14] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues. *Bioinformatics*, 21(22):4116–4124, 2005.

[15] L. Oliveira, A. C. M. Paiva, and G. Vriend. Correlated Mutation Analyses on Very Large Sequence Families. *ChemBioChem*, 3:1010–1017, 2002.

[16] N. Osada and H. Akashi. Mitochondrial-nuclear interactions and accelerated compensatory evolution: evidence from the primate cytochrome C oxidase complex. *Molecular biology and evolution*, 29(1):337–46, jan 2012.

[17] F. Pazos and A. Valencia. Protein co-evolution, co-adaptation and interactions. *The EMBO Journal*, 27(20):2648–2655, 2008.

[18] B. Pfeifer, U. Wittelsbuerger, S. E. Ramos-Onsins, and M. J. Lercher. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*, 31:1929–1936, 2014.

[19] M. Scheulke, J. Smeitink, E. Mariman, J. Leoffen, B. Plecko, F. Trijbels, S. Stöckler-Ipsiroglu, and L. van den Heuvel. Mutant NDUFV1subunit of mitochondrial complex I causes leukodystrophy and myoclonic epilepsy. *Nature Genetics*, 21(march):260–261, 1999.

[20] K. Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[21] T. Silkjaer, C. Nyvold Guldborg, C. Juhl-christensen, P. Hokland, and J. Maxwell. Mitochondrial cytochrome c oxidase subunit II variations predict adverse prognosis in cytogenetically normal acute myeloid leukaemia. *European Journal of Haematology*, 91:295–303, 2013.

[22] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[23] B. Szamecz, G. Boross, D. Kalapis, K. Kovacs, G. Fekete, Z. Farkas, V. Lazar, M. Hrtyan, P. Kemmeren, M. J. A. Groot Koerkamp, E. Rutkai, F. C. P. Holstege, B. Papp, and C. Pal. The Genomic Landscape of Compensatory Evolution. *PLoS Biology*, 12(8), 2014.

[24] Z. Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8):1586–1591, 2007.

[25] D. V. Zaykin, A. Pudovkin, and B. S. Weir. Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics*, 180(1):533–45, sep 2008.