Genetics & Applications

Institute for Genetic Engineering and Biotechnology University of Sarajevo

Research article Open access

In silico Prediction of *mt*DNA Gene Expression Based on Codon Usage Bias in Ants (Formicidae Latreille, 1802) that Inhabit Limestone Quarry Ecosystems

Semir Dorić[1*], Dinko Osmanković[2] and Lada Lukić Bilela[3]

[1] *University of Sarajevo, Institute for Genetic Engineering and Biotechnology, Sarajevo, Bosnia and Herzegovina*
[2] *University of Sarajevo, Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina*
[3] *University of Sarajevo, Faculty of Science, Sarajevo, Bosnia and Herzegovina*

DOI: 10.31383/ga.vol2iss1pp32-37

***Correspondence**
E-mail: semir.doric@ingeb.unsa.ba

Received
February, 2018
Accepted
April, 2018
Published
June, 2018

Keywords

codon usage bias, MILC, MELP, ants, gene expression prediction

Abstract

Codon usage is considered as a modulator of gene expression, due to high correlation between codon usage, tRNA abundance and the level of gene expression. Adaptability is primarily manifested at gene level therefore *mt*DNA gene expression analysis may indicate trends toward the development of adaptive traits for specific environmental conditions. Moreover, modified gene expression patterns may result from such adaptations. Due to their sensitivity to environmental disturbances, great functional importance and accessibility ants (Family: Formicidae Latreille, 1802) are excellent model organisms for molecular and bioinformatics genome analysis. This *in silico* simulation is based on the comparison of codon usage bias and the level of gene expression of currently available mitochondrial protein-coding genes of ant species that were sampled at quarry Ribnica (Kakanj, Bosnia and Herzegovina). MILC and MELP algorithms were used forcodon usage bias analysis and mitochondrial gene expression prediction, respectively. The analysis included four *mt*DNA protein-coding genes from eight selected species of ants totaling in 32 protein sequences. The results of codon usage analysis indicated no statistically significant differences in codon usage bias, as well as relative frequencies of the gene expression level. The next step should be directed to molecular ecology studies, even using whole genome measures of gene expression (RNA-seq; transcriptomics) to capture molecular response to environmental challenges.

Introduction

Mitochondrial genome of Formicide encodes for small and large ribosomal subunit, 13 proteins and up to 22 tRNA molecules. Coding sequences use synonymous codons with unequal frequencies. Thus, highly expressed genes use a subset of 'optimal' codons which are recognized by the most abundant tRNAs while in weakly expressed genes codons are

recognized by rare tRNAs. Although most proteins are expressed near their optimal levels, abundantly expressed proteins should be highly optimized owing to the high energetic cost of production (Gout et al., 2010; Vishnoi et al., 2010). According to high correlation between codon usage, tRNA abundance and level of gene expression, it has been suggested that codon usage is a modulator of gene expression (Holm, 1986). Nonrandom codon usage, or codon bias, is a common phenomenon in a wide variety of organisms. It is generally thought to be determined by a balance among mutation, genetic drift, and natural selection based on translation efficiency and/or accuracy (Ingvarson, 2007).

The aim of this study was to explore possible influence of ecosystems degradation level on gene expression pattern changes in ants by in silico prediction of *mt*DNA gene expression based on codon usage bias. However, further studies should include transcriptome analysis in order to reveal particular gene expression patterns and their relation with possible adaptations to new adaptive zones in degraded ecosystems with strong anthropogenic impact. Molecular ecology has moved beyond the use of a relatively small number of markers and application of transcriptomics has become an imperative in an ecological context (Richards et al., 2009). Alvarez et al., 2015 have made a major Ten years of transcriptomics in the wild contribution towards understanding of significant impacts that stress response may have on numerous categories of genes, but have also shown that even small changes in environment may affect transcription (Richards et al., 2012).

Materials and methods

Sample Collection and Taxonomic Identification

Research methodology used in this study is shown in Figure 1. Myrmecofauna collecting within three areas [quarry (tertiary ecosystem), mesophyll's meadow (secondary ecosystem) and pine forest and hornbeam (primary ecosystem)] was performed during field research (Jun-Sep, 2014). The collected specimens were stored in tubes containing 96% ethyl alcohol. The ant species were identified in the lab according to two taxonomic keys (Agosti & Collingwood, 1987; Seifert, 2007). The identified ant species were photographed under a binocular microscope (10 x 40 magnification, Figures 2A-H). The identified ant species are as follows: Primary ecosystem (1) *Lasisus niger* Linnaeus, 1758, (2) *Camponotus vagus* (Scopoli, 1763); Secondary ecosystem (3) *Formica cunicularia* Linnaeus, 1758, (4) *Camponotus ligniperdus* (Latreille, 1802), (5) *Formica fusca* Linnaeus, 1758; Tertiary ecosystem (6) *Tetramorium caespitum* (Linnaeus, 1758), (7) *Aphaenogaster subterranea* (Latreille, 1798), (8) *Tapinoma erraticum* (Latreille, 1798).

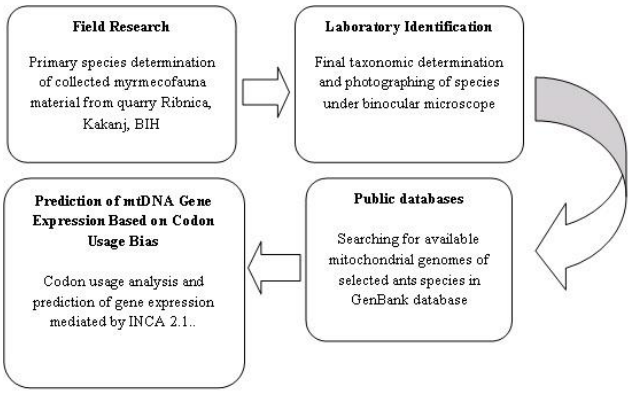

Figure 1. Schematics of research methodology used

In silico analysis

Genomic data was obtained from the public biological database NCBI/GenBank and consists of complete sequences of available protein-coding genes from mitochondrial genomes of eight selected formicid species retrieved in FASTA format. Mitochondrial genes used for codon usage bias analysis and prediction of gene expression are *cox1*, *cox2*, *cytb* and *nad6* (Tab 1). INCA 2.1 (Interactive Codon Usage Analyzer) software was used for synonymous codon usage analysis of nucleotide sequences (Supek & Vlahoviček, 2006). Two main algorithms for codon usage calculations, MILC and MELP were used. Measure Independent of Length and Composition (MILC) is, mathematically, based on a log-likelihood ratio score used in the statistical G-test for goodness-of-fit. This methodology yields numerically similar results to the more commonly used $\chi 2$ test, but may hold theoretical advantages over it in statistical analyses. MILC- based Expression Level Predictor (MELP) is a method of quantitatively predicting gene expressivity and was computed simply as the ratio of respective distances

of a gene's codon usage from the genomic average, and a predefined reference set.

Fortunately, optimal codon usage in genes seems to coincide with transcription enhancing factors. Therefore it is sensible to explore a correlation between codon usage (acting at translation level) and the abundance of transcript. Reference set was defined simply by including protein genes that appear to be valid starting point for expression level predictions in the sampled formicid species.

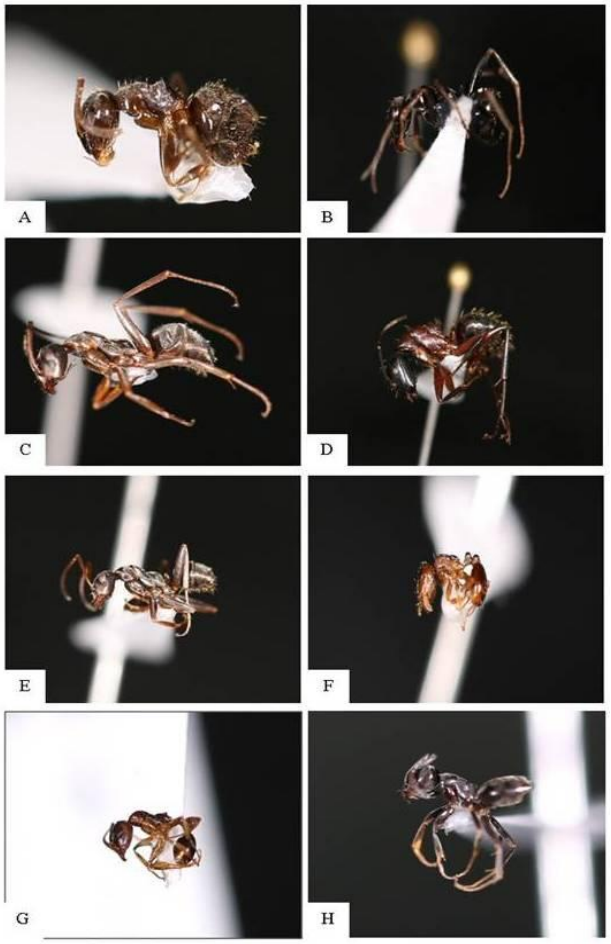

Figure 2. Myrmecofauna species found at Ribnica area: Primary ecosystem - A) *Lasisus niger*; (B) *Camponotus vagus*; Secondary ecosystem - (C) *Formica cunicularia*; (D) *Camponotus ligniperdus*; (E) *Formica fusca*; Tertiary ecosystem - (F) *Tetramorium caespitum*; (G) *Aphaenogaster subterranean*; (H) *Tapinoma erraticum* (Foto: Adi Vesnić)

Results and Discussion

Relative values of codon usage frequency (MILC) were used for the calculation of relative protein-coding gene expression level (MELP) values. We found MELP values ranging from 0.9672 for *cytb* gene of *Aphaenogaster subterranea* to 1.4302 for *nad6* gene of *Camponotus ligniperdus*. Average MELP value for 32 genes of eight ants species was 1.741.

Table 1. Selected ant species and their *mt*DNA genes used for *in silico* analysis, with accession numbers in GenBank database.

No	Species	Genes	Accession numbers
1	*Lasius niger*	*cox1, cox2, cytb, nad6*	AY452155, AY456155, JQ681115, JQ681010
2	*Camponotus vagus*		AB019415, AB019442, AY185274, AY185224
3	*Formica cunicularia*		AB010926, HQ853322, HQ651084, DE170885
4	*Camponotus ligniperdus*		JN134886, JN134992, JN134777, JX179785
5	*Formica fusca*		AB010925, AY334398, HQ651077, JA170888
6	*Tetramorium caespitum*		KC510134, DQ335195, EU616819, JX199785
7	*Aphaenogaster subterranea*		FJ824414, AL348726, EF518580, GA178331
8	*Tapinoma erraticum*		GU388394, AE314322, JE683715, AS183414

Average MELP value for *cox1* gene is 1.0834, *cox2* gene 1.3232, 1.0854 for *cytb* gene and for *nad6* gene 1.3652. MELP values for *cox1* gene ranged between 1.0428 to 1.1080, for *cox2* gene from 1.0957 to 1.4047, for *cytb* gene from 0.9672 to 1.1508 and for *nad6* gene from 1.0972 to 1.4302 (Fig 3-6). Variability of MELP values for 32 gene sequences in eight selected species of the family Formicidae are shown in Figure 7.

Codon usage analysis results of selected formicid mitochondrial genomes point to small differences in relative frequencies of codon usage as well as in the expression levels of genes of species that inhabit three distinct ecosystems. There was a statistically insignificant difference between MELP values for 12 genes for species that inhabit tertiary ecosystems, relative to the mean MELP value of the same genes for species that inhabit primary and secondary

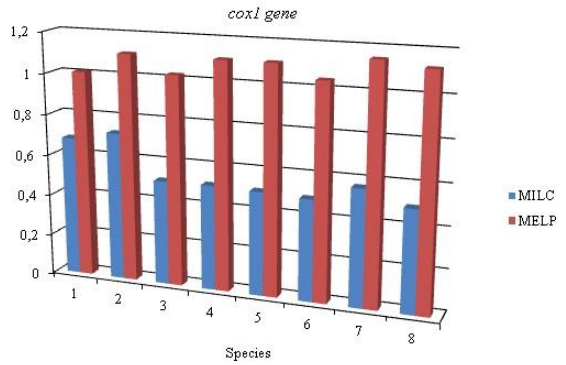

Figure 3. Relative frequencies of codon usage (MILC) and the level of gene expression (MELP) for *cox1* gene of eight selected species from the family Formicidae Latreille, 1802 mediated by INCA 2.1. software

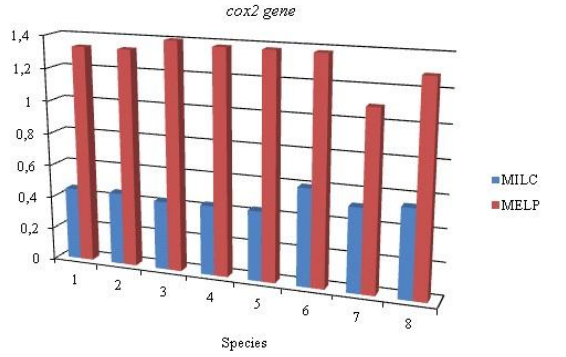

Figure 4. Relative frequencies of codon usage (MILC) and the level of gene expression (MELP) for *cox2* gene of eight selected species from the family Formicidae Latreille, 1802 mediated by INCA 2.1. software

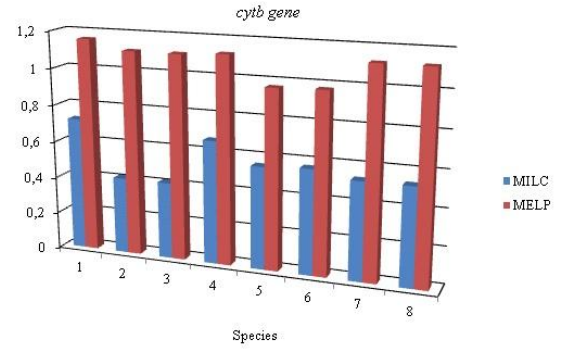

Figure 5. Relative frequencies of codon usage (MILC) and the level of gene expression (MELP) for *cytb* gene of eight selected species from the family Formicidae Latreille, 1802 mediated by INCA 2.1. software

ecosystems. Variability of MELP values for 12 genes from species found at tertiary ecosystems relative to the mean MELP value for genes from species of primary ecosystems (MVP) and the mean MELP value for genes from species of secondary ecosystems (MVS) are shown in Figure 8, while the variability of MELP values for 12 genes from species of tertiary ecosystem in relation to the MELP values for genes of species which inhabit primary and secondary ecosystem is shown in Figure 9.

The expression of the analyzed protein-coding genes in the species that inhabit three different ecosystems were >95% identical.

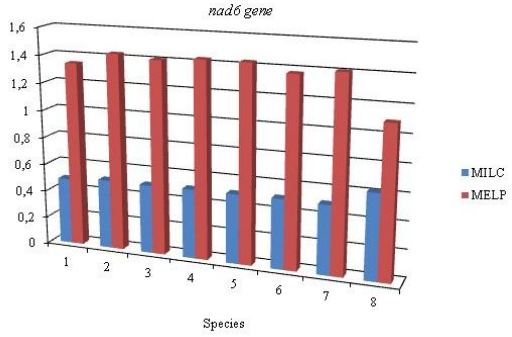

Figure 6. Relative frequencies of codon usage (MILC) and the level of gene expression (MELP) for *nad6* gene of eight selected species from the family Formicidae Latreille, 1802 mediated by INCA 2.1. software

The most commonly used codon in the pool of 32 analyzed genes was UUA (Leu) with the number of 122 and relative synonymous codon using value of 3.3. Codon CGC (Arg) is the least frequently used codon with the number of 2.3, and with relative synonymous codon usage value of 0.37.

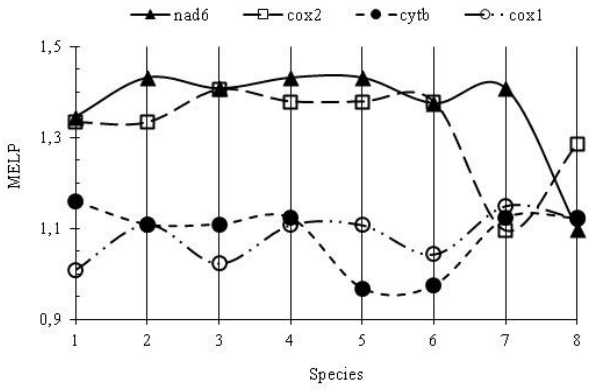

Figure 7. Variability of MELP values for 32 genes (*cox1*, *cox2*, *cytb* and *nad6*) of eight selected species of the family Formicidae Latreille, 1802

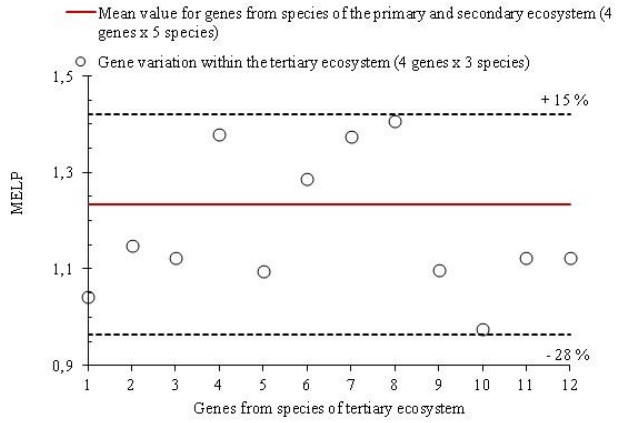

Figure 8. Variability of MELP values for 12 genes from species of tertiary ecosystems relative to the mean MELP value for genes from species of primary ecosystems (MVP) and the mean MELP value for genes from species of secondary ecosystems (MVS)

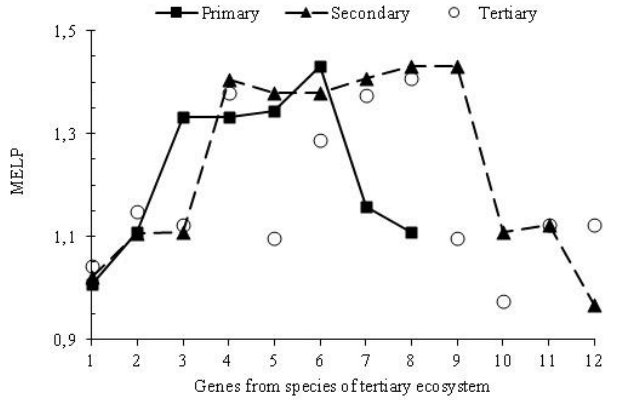

Figure 9. Variability of MELP values for 12 genes from species of tertiary ecosystem in relation to the MELP values for genes of species which inhabit primary and secondary ecosystem

Conclusions

This *in silico* prediction of *mt*DNA gene expression prediction based on codon usage bias showed to be not informative. Further studies should include a wider set of genes/proteins involved in crucial metabolic processes. Furthermore, transcriptome analysis of selected formicid species could help in elucidation of possible adaptations to ecosystems with strong antropogenic impact, considering variation in gene expression affected by environmental stimuli.

Acknowledgement

This research was part of the Quarry Life Award 2014 contest, organized and financed by HeidelbergCement Group. We extend our gratitude to Dr. Adi Vesnić (University of Sarajevo, Faculty of Science) for his help with final taxonomic identification and revision of previously identified ant species.

References

Alvarez M, Schrey AW, Richards CL (2015) Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Mol Ecol, 24(4):710-725.

Agosti D, Collingwood CA (1987) A provisional list of the Balkan ants (Hym. Formicidae) with a key to the worker caste. II. Key to the worker caste, including the European species without the Iberian. Mitteilungen der Schweizerischen Entomologischen Gesellschaft, 60:261-293.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2004) GenBank. Nucleic Acids Research, 23:3-6.

Gout J-F, Kahn D, Duret L (2010) Paramecium Post-Genomics Consortium, The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet. 6:e1000944.

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York.

Seifert B (2007) Ameisen Mittel- und Nordeuropas. Tauer: lutra Verlags – und Vertreibsgesellschaft, 368.

Supek F, Vlahoviček K (2006) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics, 6:182.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution, 30:2725-2729.

Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, Yoder RJ, Robbertse B, Spatafora JW, Rappé MS, Giovannoni SJ (2011) Phylogenomic evidence for

a common ancestor of mitochondria and the SAR11 clade. Scientific Reports, 1:13.

Homl L (1986) Codon usage and gene expression. Nucleic Acid Research, 14:7.

Ingvarson PK (2007) Gene Expression and Protein Length Influence Codon Usage and Rates of Sequence Evolution in Populus tremula. Oxford University Press.

Richards CL, Hanzawa Y, Katari MS, Ehrenreich IM, Engelmann KE, Purugganan MD (2009) Perspectives on Ecological and Evolutionary Systems Biology. Annual Plant Reviews Volume 35: Plant Systems Biology, 331-349.

Richards CL, Rosas U, Banta J, Bhambhra N, Purugganan MD (2012) Genome-wide patterns of Arabidopsis gene expression in nature. PLoS genetics, 8(4):e1002662.

Vishnoi AS, Kryazhimskiy G, Hannenhalli BS, Plotkin J (2010) Young proteins experience for variable selection pressures than old proteins. Genome research, 20:1574-1581.