Nucleotide Frequency Distribution in Mitochondrial Genomes: Analysis and Utilisation

Preliminary Results and Outlook

Abdullah Sahyoun [1,2]

Matthias Bernt [2]

Peter F. Stadler [2]

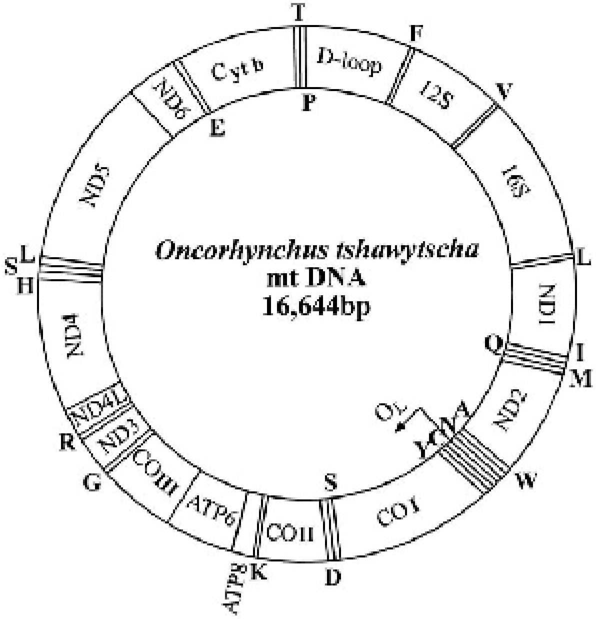
Kifah Tout [1]

[1]: Lebanese University, Azm Center for Biotechnology research

[2]: Leipzig University, Bioinformatics Institute

Mitochondrial Genomes

- organelles included in the cells of eucaryotic animals, related to some disease (mt disorders, heart diseases).



» About 16 kb in size and contain 37 genes:

  » 13 protein-coding genes.

  » 22 transfer RNA genes (tRNA)

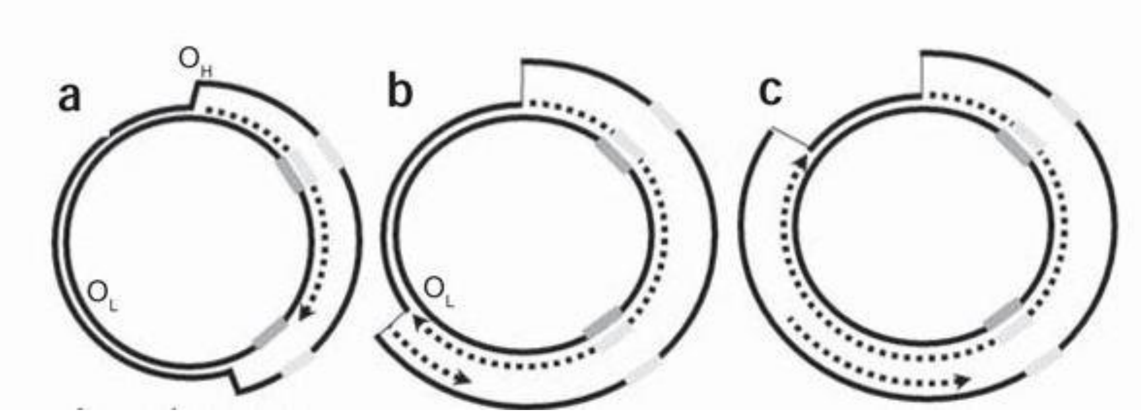
  » and two ribosomal RNA genes (rRNA) [Boore JL , Nucleic Acids Re, 1999]

- asymmetry of the nucleotide composition is a well known feature
  - Light and Heavy strand (A and G are heavier) [Perna and Kocher, J Mol Evol , 1995]

- Goal: Analysis and understanding nucleotide frequency distribution(anomalies)

Replication of Mitochondrial DNA

- replication process takes approximately 2 hours[Clayton, Int. review of cytology, 1992] and starts from the origin of H-strand replication (a) (OH).
- L-strand synthesis starts in the opposite direction.
- The H-strand is single-stranded until the L-strand replication is completed (b-c).



- Asymmetrical replication is a potential source of strand bias [Faith and Pollock , Genetics, 2003].
- the H-strand is exposed to mutation and damage:
  - hydrolytic deamination of cytosine (C->Uracil->T)
  - hydrolytic deamination of adenine (A->Hypoxanthine->G)[Reyes et al. Mol Biol Evol 1998]

Skew and biases

- Nucleotide skews between complementary strands is a remarkable feature of mtDNA and a violation of Chargaff's second parity rule called strand asymmetry.

1. AT skew $\left(\frac{A-T}{A+T}\right)$, GC skew $\left(\frac{G-C}{G+C}\right)$ and the strand bias $\left(\frac{G+C}{A+T}\right)$ [Grigoriev A, Nucleic Acids Re, 1998]
    1. + AT-skew = more A than T

2. New quotients to study the bias full filing the two properties(strand asymmetry and the strand bias):

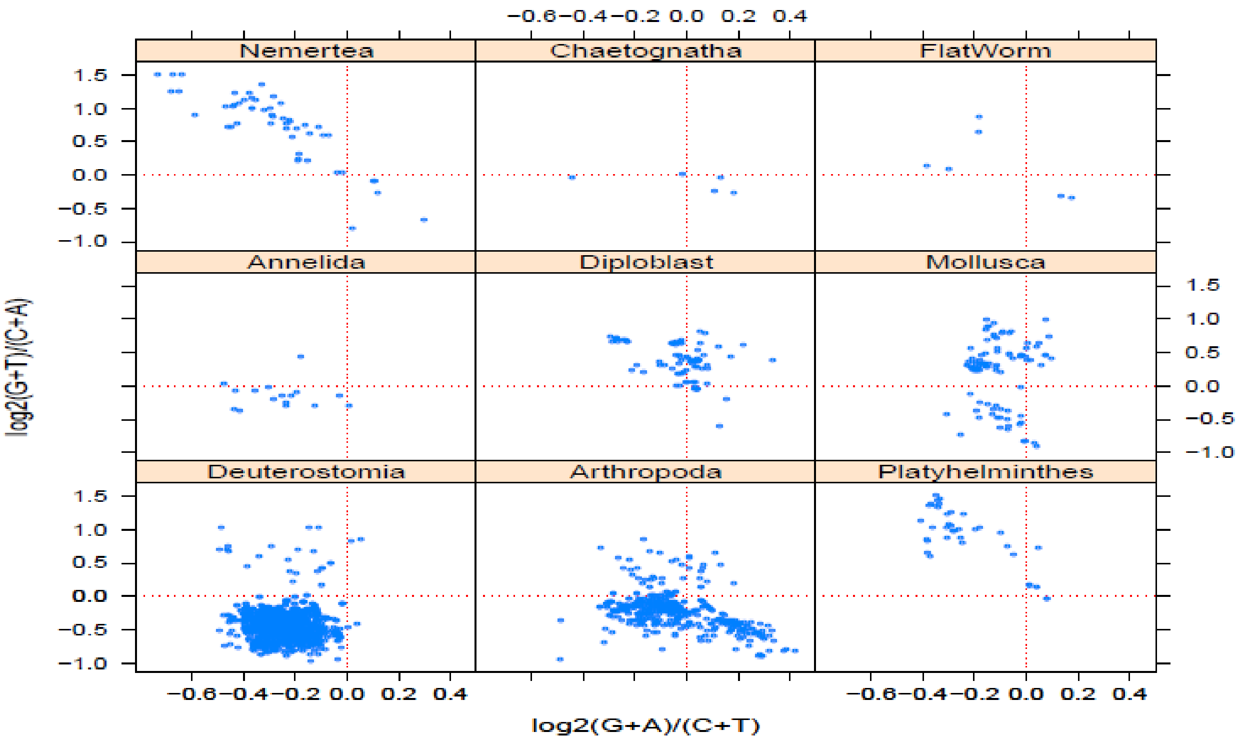
$$\log_2\left(\frac{G+T}{C+A}\right)$$

$$\log_2\left(\frac{G+A}{C+T}\right)$$

$$\log_2\left(\frac{G+C}{A+T}\right)$$

- Goals: new quotients are compared to the previously used ones

Results for Metazoa



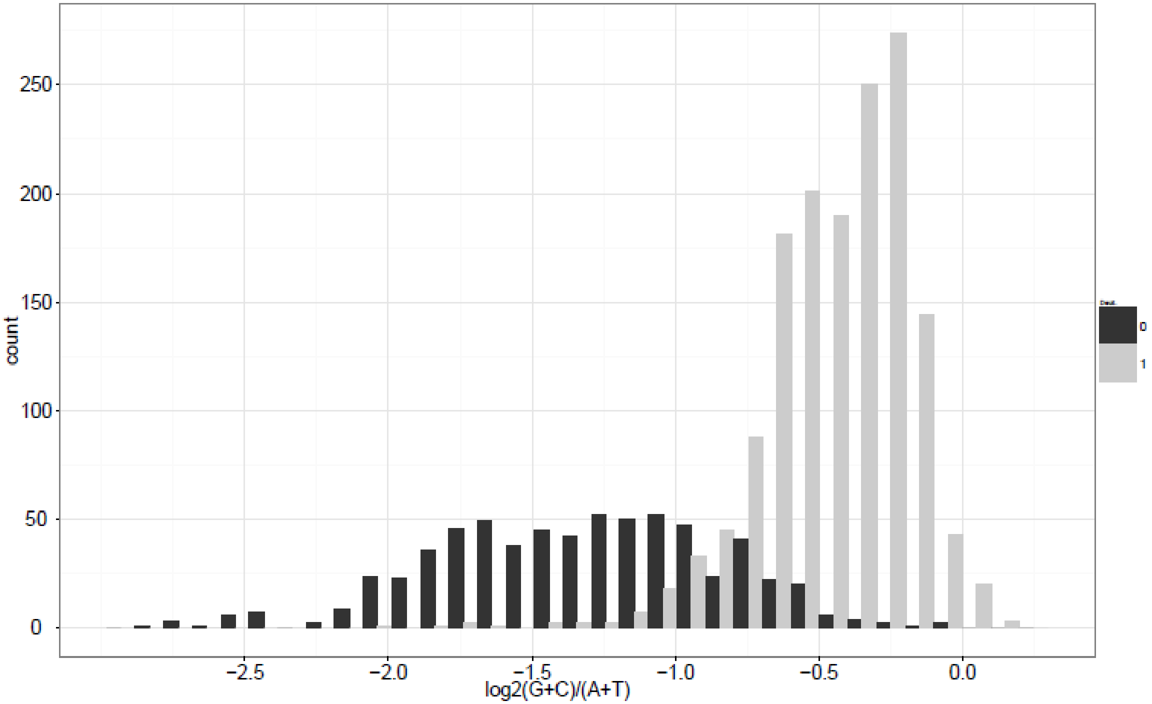
- Arthropoda and Mollusca symmetry in both dimensions

  - maybe inverted values because of gene rearrangements

  - maybe the reverse complementary strand is used or it is related to a change in the mode of replication of these species (Mollusca and Arthropoda)

- Deuterostomia have some peculiar grouping(top left part) and they need to be analysed in more details

Deuterostomia

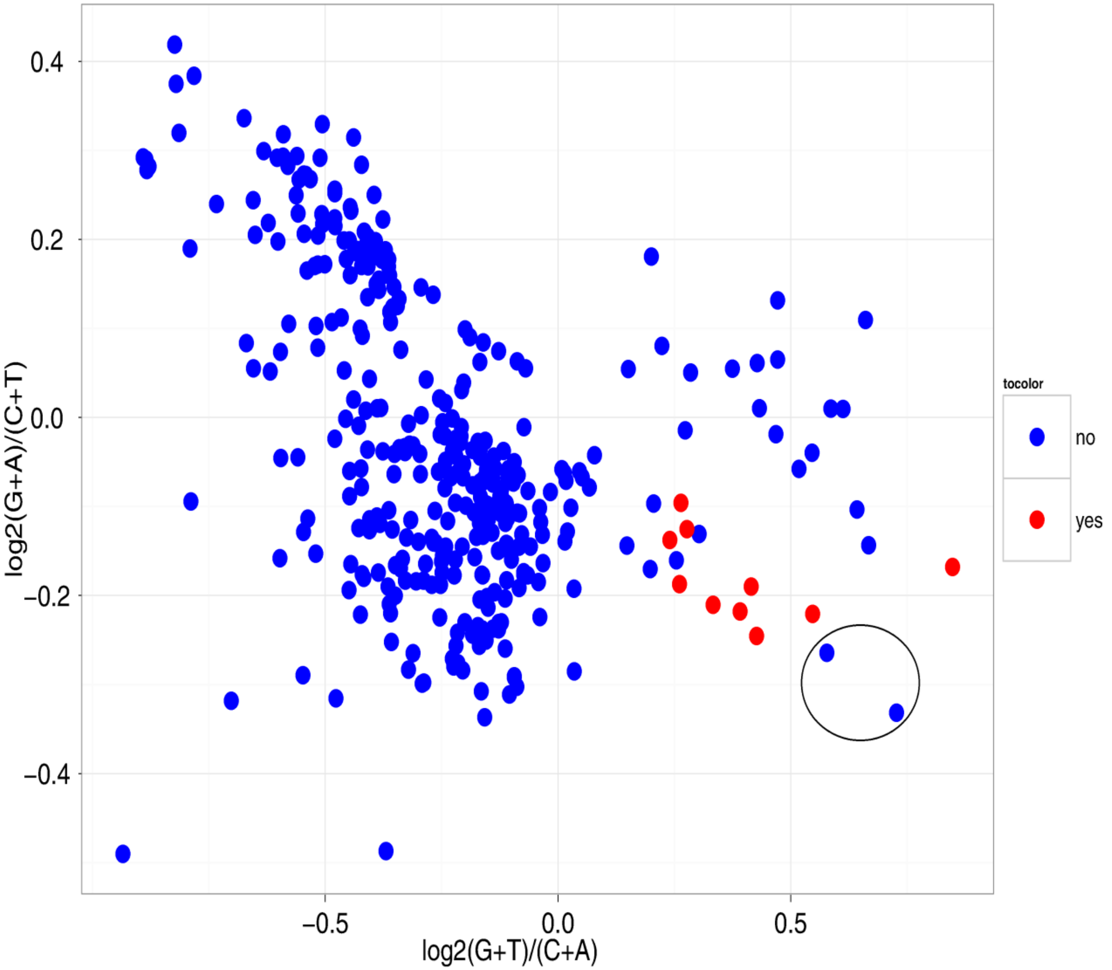


- Light gray: Deuterstomia
- Dark grey: all other metazoan

•Only Deuterostomia have few species with $\log_2\left(\frac{G+C}{A+T}\right) >= 0$

•Why do we have these values? What is special about them?

Arthropoda



- most species have a strand asymmetry characterized by (A>T and C>G)

- Some arthropods that have a reverse in strand asymmetry tested using AT-skew and GC-skew (Hymenoptera, Phthiraptera, Hemiptera) (colored in red) [Wei et al. Plosone, 2010] which are spiders.

- Circled points (American House dust mite and European house dust mite [Wikipedia]) which are of the same spiders family.

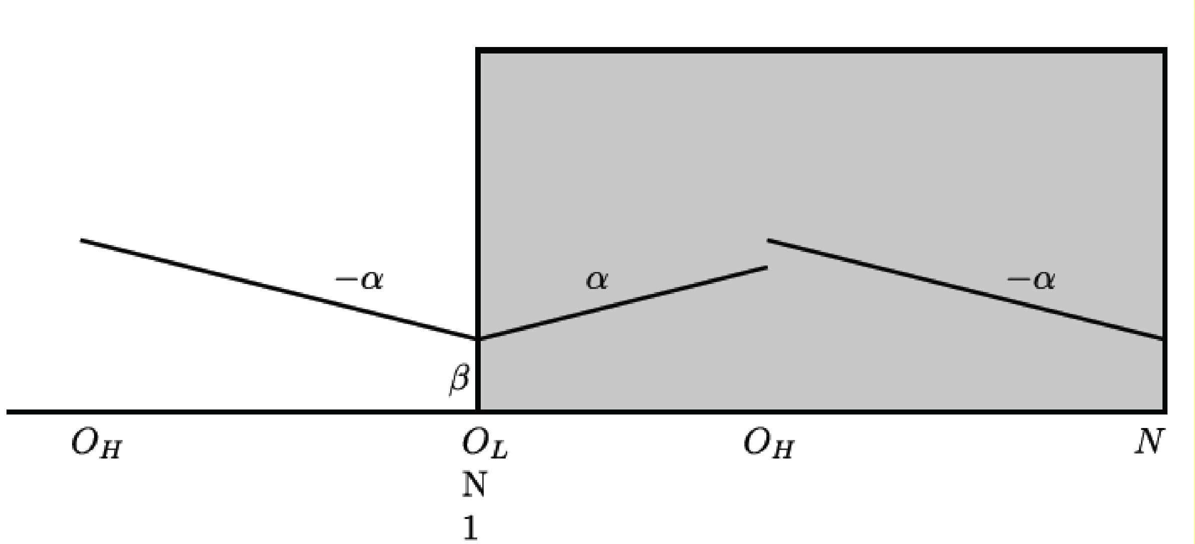
- Further analysis is needed

Linear Regression and origin of replication

- Replication process and how it’s affecting nucleotide frequencies

the distances to the replication origins and the number of mutations G -> A, T -> C [Krishnan and Pol. DNA Cell Biol, 2004]

Linear Regression



•Given the positions of OL and OH

•We are trying to test for all possible OH and OL .

1. Does the pair with the minimum error correspond to the actual position??
   If yes we can try to find the OH and OL for other species

Equations

- We base our study on the hypothesis of a linear relation between the time of being single stranded during replication (i.e. the distances to the replication origins) and the number of mutations.

- Error function as illustrated in the previous figure:

$$E = \sum_{i=0}^{O_H} [\alpha i + \beta - y_i]^2 + \sum_{i=O_H+1}^{N-1} [\alpha(N-i) + \beta - y_i]^2$$

- α and ß are the slope and intercept of the linear function.

- N is the length of the genome, Y_i gives the value of $\log_2\left(\frac{G+T}{C+A}\right)$

- OH,OL (origin of heavy and light strand replication origin)

- We assume that the data is transformed with a cyclical shift to leave OL on the first position because the single strand state will start from this position

Perspectives

1. What happens when the genomes are rearranged.?
    1. Extreme case inversion or transposition of the replication origins?
    2. Is this observable in the skew/bias values?

2. The model for replication is only verified for a few chordates.

3. In other species(e.g. Protostomia) the properties are often very different: Multiple replication origins?

4. For few species OL has been reported 97% after OH for Arthropods and around 2/3 of the genome for Mammalia.
    1. What about the Others?
    2. Is this position producing these peculiar values??

Acknowledgments

- Thank you for your attention
- Azm Association and Lebanese University for funding my PhD
- Leipzig University for also helping with the funds
- Kifah Tout
- Peter Stadler
- Matthias Bernt
- Christian and Stephanie for Driving us to Bled
- Institute staff for helping me to learn German ;)

Danke!!