Adaptation in genes, duplicates, families, functional modules and genomes

François Serra

October 2011

Contents

Nomenclatura

BWT Burros-Wheeler transform

CR Complexity Ratio

1. Introduction

1.1. Adaptive changes to evolutionary speed

1.2. Evolution, and the detection at molecular level

1.3. Grouping genes and finding evolutionary patterns

1.4. What is DNA? How genes rose?

1.5. Life in DNA, from genes to repetitive elements.

Part I.

Structure and dynamics of genomes

2. Random-like structure of DNA

2.1. Introduction

From a biological perspective it seems obvious that DNA is something else than random mix of A, T, G or C nucleotides. Genomes are composed of functional elements as can be genes or promoters but also repetitive elements that by definition can not be random when taken together. However to what extent can we state that genomes are not a random soup of 4 letters?

This question could be solved in some sense by measuring genomes entropy. This measure presents the disadvantage that extreme cases of high entropy could correspond to *a*) **a specially high content of information**, entropy-based algorithms are actually used to predict or confirm automatic detection of genes [Du *et al.*2006, Gerstein *et al.*2007], *b*) **an exact random structure**, some work in the sense of testing the random structure of DNA have been done using entropy [Loewenstern & Yianilos1999]. However this characteristic of entropy could be only a semantic problem if we use it as a measure of relative variation in DNA complexity in genomes, and try to discern statistical patterns in the DNA sequences of different genomic element such as interspersed repeats or functional element (like protein-coding genes). This kind of description of DNA sequence complexity was already done by [Holste *et al.*2001], but only in human chromosome 22.

2.2. Results and Discussion

2.3. Material and methods

2.3.1. The complexity ratio and complexity value

Complexity Ratio CR is defined by a classical formula used in data compression [Adjeroh *et al.*2008], the Burros-Wheeler transform BWT [Burrows & Wheeler1994].

2.3.2. Complexity in strings

2.3.3. Simulations

3. Life inside genomes, dynamics and predictions

3.1. Genomic elements, dispersion and abundance

3.2. Species Abundance Diversity in genomes

3.3. Neutrality of SAD

3.4. Material Methods

3.4.1. Ecolopy

Part II.

Detection of selective pressures in genomes

4. Searching for evolutionary patterns in funcionally linked group of genes

4.1. Introduction

4.2. Material and Methods

4.2.1. Dataset

Five mammals

Complete genomes of 5 mammals species (*Homo sapiens*, *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus* and *Canis familiaris*) where retrieved from *Ensembl* [Flicek *et al.*2011]. Also orthology prediction between each pair of species possibly done between human and the others was retrieved from *Ensembl Compara* [Vilella *et al.*2009] using biomart [Kinsella *et al.*2011]. Only groups of orthologs *one-to-one* with one representative of each species where kept in the final dataset. 4.1 NUMBERS

6 Drosophila

4.2.2. Alignments

Each of the group of orthologous sequences were aligned with Muscle [Edgar2004], and, once aligned sequences were cleaned with trimal [Capella-Gutiérrez *et al.*2009] keeping all sequences but trimming aligment columns with the euristic1 method.

4.3. open on colocalization to not random

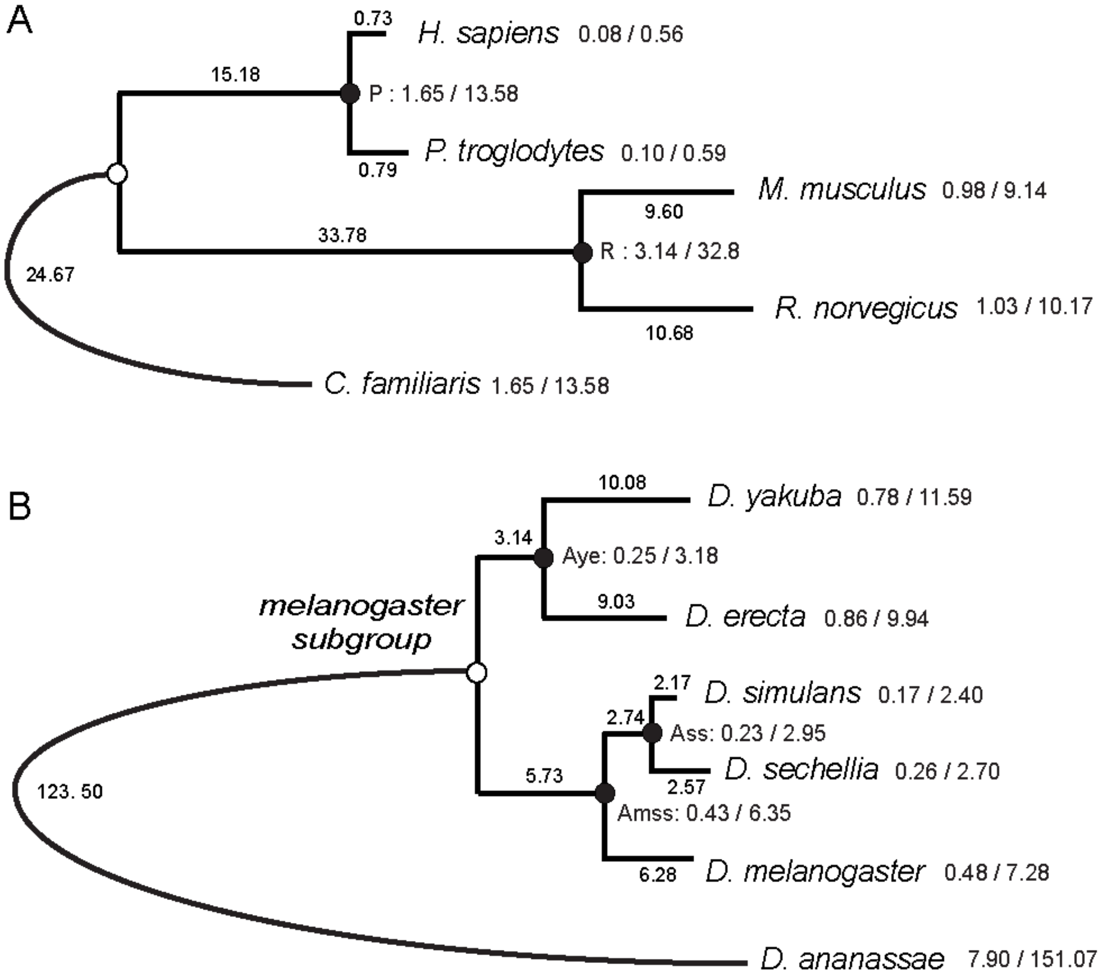

Figure 4.1.: Mammals and *Drosophila* phylogeny. blabli blob lu dkfnlskjdf

5. Tools, programs, methods

5.1. ETE-evol plugin

5.1.1. BRANCHED1

5.1.2. Protamines Rodents and Primates

5.2. Pipeline for study of adaptation at genomic scale

5.2.1. Selective pressure on duplicated genes in Drosophila

5.3. Phylemon

6. Conclusions

Bibliography

[Adjeroh *et al.*2008] Donald Adjeroh, Tim Bell, and Amar Mukherjee, The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. In *ACM SIGACT News*, vol. 41, 21–24. Springer US, Boston, MA, 2008.

[Burrows & Wheeler1994] Michael Burrows and David J Wheeler, A block-sorting lossless data compression algorithm. *Digital SRC Research Report* **124** (1994).

[Capella-Gutiérrez *et al.*2009] Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)* **25**(15) (2009), 1972–3.

[Du *et al.*2006] Jiang Du, Joel S Rozowsky, Jan O Korbel, Zhengdong D Zhang, Thomas E Royce, Martin H Schultz, Michael Snyder, and Mark Gerstein, A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics (Oxford, England)* **22**(24) (2006), 3016–24.

[Edgar2004] Robert C Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**(5) (2004), 1792–7.

[Flicek *et al.*2011] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Yuan Chen, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Leo Gordon, Maurice Hendrix, Thibaut Hourlier, Nathan Johnson, Andreas Kähäri, Damian Keefe, Stephen Keenan, Rhoda Kinsella, Felix Kokocinski, Eugene Kulesha, Pontus Larsson, Ian Longden, William McLaren, Bert Overduin, Bethan Pritchard, Harpreet Singh Riat, Daniel Rios, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sobral, Giulietta Spudich, Y Amy Tang, Stephen Trevanion, Jana Vandrovcova, Albert J Vilella, Simon White, Steven P Wilder, Amonida Zadissa, Jorge Zamora, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Richard Durbin, Xosé M Fernández-Suarez, Javier Herrero, Tim J P Hubbard, Anne Parker, Glenn Proctor, Jan Vogel, and Stephen M J Searle, Ensembl 2011. *Nucleic acids research* **39**(Database issue) (2011), D800–6.

[Gerstein *et al.*2007] Mark B Gerstein, Can Bruce, Joel S Rozowsky, Deyou Zheng, Jiang Du, Jan O Korbel, Olof Emanuelsson, Zhengdong D Zhang, Sherman Weissman, and Michael Snyder, What is a gene, post-ENCODE? History and updated definition. *Genome research* **17**(6) (2007), 669–81.

[Holste *et al.*2001] Dirk Holste, Ivo Grosse, and Hanspeter Herzel, Statistical analysis of the DNA sequence of human chromosome 22. *Physical Review E* **64**(4) (2001), 1–9.

[Kinsella *et al.*2011] Rhoda J. Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul

Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek, Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database : the journal of biological databases and curation* **2011** (2011), p. bar030.

[Loewenstern & Yianilos1999] David M Loewenstern and Peter N Yianilos, Significantly lower entropy estimates for natural DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology* **6**(1) (1999), 125–42.

[Vilella *et al.*2009] Albert J Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney, EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**(2) (2009), 327–35.

List of Figures

List of Tables