Informational, Ecological and System Approaches for Complete Genome Analysis

François Serra

October 2011
Contents

I. Introduction
   1. Nothing in evolution makes sense except over the denominator of neutrality
   2. What is DNA? – Defining a genome.
      1.2.1. Biological complexity versus genome size
      1.2.2. Informational content of DNA
      1.2.3. Dynamics of genetics elements
   1.3. Genomic study of selective pressures in set of genes
      1.3.1. Detection of adaptation at molecular level – single gene approach
      1.3.2. Identification of selective pressures in the genomic era
      1.3.3. Back to square one – a new methodology
   1.4. Side-products – Implementation of software and pipeline
      1.4.1. Pipeline for the detection of molecular evolution
      1.4.2. Phylemon web server

II. Material and Methods
    2.1. Measuring DNA complexity
       2.1.1. The complexity ratio and complexity value
       2.1.2. Complexity in strings
       2.1.3. Simulations
    2.2. Measuring dynamics of genetic species
       2.2.1. Genomes
       2.2.2. Mining of Genetic Species
       2.2.3. Randomization of genetic elements
       2.2.4. Ecolopy
       2.2.5. Fitting Neutral Ecological models
## Contents

### 2.2.6. Model optimization

- Page 30

### 2.2.7. Model testing – Likelihood ratio test

- Page 32

### 2.2.8. Testing UNTB

- Page 32

### 2.3. Detection of selective pressure at molecular level

- Page 34

#### 2.3.1. Orthology prediction

- Page 34

#### 2.3.2. Alignments refinement and filters

- Page 36

#### 2.3.3. Evolutionary analysis

- Page 37

#### 2.3.4. GSSA, evolutionary and statistical simulations

- Page 37

### I. Structure and dynamics of genomes

- Page 45

#### 3. Random-like structure of DNA

- Page 47

##### 3.1. Computing genome complexity

- Page 47

###### 3.1.1. Genome complexity and ploidy level

- Page 52

##### 3.2. Chromosome complexity

- Page 53

##### 3.3. Complexity in repetitive elements and genes – low and high?

- Page 54

##### 3.4. Methodological validation and interpretation – Understanding the CR

- Page 56

###### 3.4.1. Complexity in chromosome segments

- Page 56

###### 3.4.2. Polyploidy and return to maximum complexity

- Page 58

###### 3.4.3. Low CR corresponds to a simple combinatorial structure of the sequence.

- Page 60

###### 3.4.4. High CR corresponds to random-like sequences

- Page 60

##### 3.5. Discussion

- Page 61

###### 3.5.1. Universal structure of DNA

- Page 61

###### 3.5.2. Mechanisms of genome amplification and divergence

- Page 61

###### 3.5.3. Genome size reduction

- Page 61

###### 3.5.4. Limits of CR space

- Page 62

###### 3.5.5. Hypotheses

- Page 62

### 4. Life inside genomes, dynamics and predictions

- Page 65

#### 4.1. Non-random distribution of genetic species

- Page 65

#### 4.2. Counterbalanced species abundances in genomes

- Page 67

##### 4.2.1. Genetic species, dispersion and abundance

- Page 67

##### 4.2.2. Randomization of genetic species in genomes

- Page 67

#### 4.3. Diversity and chromosome length

- Page 70

#### 4.4. Neutrality of species abundances and diversity

- Page 71

#### 4.5. Discussion

- Page 73
II. Detection of selective pressures in genomes 75

5. Searching for evolutionary patterns in functionally linked group of genes 77
   5.1. Gene-set selection analysis on functional modules 77
   5.1.1. Clade specific patterns 78
   5.1.2. Species specific enrichment 79
   5.1.3. Comparison of evolutionary variables used 84
   5.2. Positively selected genes among fast and slow evolving functional modules 84
   5.2.1. Descriptive analysis 84
   5.2.2. Statistical approach 86
   5.3. Discussion 87

6. Tools, programs, methods 91
   6.1. Overview on the detection of selective pressures at genomic level 91
   6.1.1. The selection of homologous sequences 91
   6.1.2. The alignment 92
   6.1.3. Model testing and phylogenetic inference 92
   6.1.4. Testing for evolutionary scenarios in protein coding genes 93
   6.2. ETE’s Evol extension 99
   6.2.1. Implementation 99
   6.2.2. General usage 100
   6.3. Phylemon2.0 104
   6.3.1. Alignment 104
   6.3.2. Model selection 105
   6.3.3. Phylogeny 105
   6.3.4. The Pipeliner 105
   6.4. Discussion 106

7. Conclusions 107

Bibliography 1

List of Figures XIX

List of Tables XXI

Glossary XXIII

A. RepeatMasker summary output A
Acknowledgments

merci merci merci merci merci merci
Read this thesis

Some particularities about this thesis:

- **The glossary**: some words are underline with dotted lines like this. A short definition of these words can be found in the Glossary. At the end of each definition, appears the page number of all occurrences of the word defined.

- **The bibliography**: the bibliography is quite standard. In the text references appears between square brackets [like this]. In the Bibliography the number of authors is limited to 15. At the end of each reference appear the page numbers corresponding to their citation in the main text (for example, [Alonso et al. 2006] cited 3 times: pages 10, 66 and 69).
# Nomenclature

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>DNA base-pair</td>
</tr>
<tr>
<td>BWT</td>
<td>Burrows-Wheeler transform</td>
</tr>
<tr>
<td>CDS</td>
<td>DNA coding sequence</td>
</tr>
<tr>
<td>chr</td>
<td>chromosome</td>
</tr>
<tr>
<td>CNE</td>
<td>Constructive neutral evolution</td>
</tr>
<tr>
<td>CR</td>
<td>Complexity Ratio</td>
</tr>
<tr>
<td>CV</td>
<td>Complexity Value</td>
</tr>
<tr>
<td>dN</td>
<td>Rate of non-synonymous substitutions</td>
</tr>
<tr>
<td>dS</td>
<td>Rate of synonymous substitutions</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FET</td>
<td>Fisher Exact Test</td>
</tr>
<tr>
<td>GE</td>
<td>Genetic Element</td>
</tr>
<tr>
<td>GS</td>
<td>Genetic Species</td>
</tr>
<tr>
<td>GSA</td>
<td>Gene-Set Analysis</td>
</tr>
<tr>
<td>GSEA</td>
<td>Gene-Set Enrichment Analysis</td>
</tr>
<tr>
<td>GSSA</td>
<td>Gene-Set Selection Analysis</td>
</tr>
<tr>
<td>H</td>
<td>Shannon’s Entropy</td>
</tr>
<tr>
<td>LRT</td>
<td>Likelihood Ratio Test</td>
</tr>
<tr>
<td>LTR</td>
<td>Long Terminal Repeat</td>
</tr>
<tr>
<td>MTF</td>
<td>Move To Front</td>
</tr>
<tr>
<td>My</td>
<td>Million years</td>
</tr>
<tr>
<td>PSG</td>
<td>Positively selected genes</td>
</tr>
<tr>
<td>RSA</td>
<td>Relative Species Abundance</td>
</tr>
<tr>
<td>SH</td>
<td>Significantly High</td>
</tr>
<tr>
<td>SL</td>
<td>Significantly Low</td>
</tr>
<tr>
<td>TE</td>
<td>Transposable Element</td>
</tr>
<tr>
<td>UNTB</td>
<td>Unified Neutral Theory of Biodiversity</td>
</tr>
<tr>
<td>WGD</td>
<td>Whole Genome Duplication</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Nothing in evolution makes sense except over the denominator of neutrality

A key concept in the study of evolution is the definition of neutrality. “Nothing in biology makes sense except in the light of evolution” [Dobzhansky 1973] the famous quote of Theodosius Dobzhansky, certainly inspired by Pierre Teilhard de Chardin’s vision:

“(Evolution) general condition to which all theories, all hypotheses, all systems must bow and which they must satisfy henceforward if they are to be thinkable and true. Evolution is a light which illuminates all facts, a curve that all lines must follow.” [de Chardin 1955]

These quotes were many time cited to support the importance of taking into account the evolutionary history in the study of whatever biological system. However in common language, evolution is often referred to as an equivalent of natural selection, and it is there that evolutionary biologist may remind the importance of neutral processes in evolutionary history.

Since the description of neutral processes, we can mention here Charles Darwin’ works with this quote appearing in The Origin of Species, right after mentioning deleterious and favourable processes:

“Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions.” [Darwin 1872]

the contribution of these neutral processes to the whole evolutionary changes has been assumed to be minor in comparison to positive and balancing forms of selection. And in any of the biological field, from ecology to (population) genetics through behavioral biology, this inference about the relative weights of neutral process versus directional changes, was accepted in spite of the lack of statistical proof. Resulting in a feeling accurately captured by Ernst Mayr:
Chapter 1. Introduction

“When one attempts to determine for a given trait whether it is the result of natural selection or of chance (the incidental byproduct of stochastic processes), one is faced by an epistemological dilemma. Almost any change in the course of evolution might have resulted by chance. Can one ever prove this? Probably never.” [Mayr 1983]

In order to answer this question a rigorous definition of neutrality was necessary. Starting from an algorithmic point of view, the definition of a neutrality is given by Dennet in the first chapter “Natural Selection as an Algorithmic Process” of his book *Darwin’s Dangerous Idea: Evolution and the Meanings of Life*:

“substrate neutrality: The procedure for long division works equally well with pencil or pen, paper or parchment, neon lights or skywriting, using any symbol system you like. The power of the procedure is due to its logical structure, not the causal powers of the materials used in the instantiation, just so long as those causal powers permit the prescribed steps to be followed exactly.” [Dennett 1995]

In this context, only once defined the causal power underlying evolutionary changes, we would be able to explain all stochastic changes, and consequently identify by default, the effect of natural selection – or, following the terminology, the logical structure of nature.

In biology, the first field to accept the importance of neutral processes and to try to define an inherent alogical structure behind the observed changes in evolution, is ecology. Since the thirties, models lacking any biological parameters were successfully able to predict abundances and diversities of species in within their ecosystem. These models were based on statistical distributions like log-normal or log-series [Motomura 1932, Fisher et al. 1943, Preston 1948]. Later, the last significant brick added to the definition of a neutral model was finally brought by Stephen Hubbell with its definition of the unified neutral theory of biodiversity [Hubbell 2001] – theory that succeeds in predicting species distributions in almost all ecosystems.

In genetics, advances in this sense had to wait until the arrival of DNA sequences data, and the first comparative works enlightening the sixties. The relative major amount of non-advantageous changes observed between sequences led Motoo Kimura to propose the neutral theory of molecular evolution [Kimura 1968]. Thanks to the amount of data and to the fantastic “trick” of comparing synonymous and non-synonymous mutations (see Detection of adaptation at molecular level – single gene approach, page 13), neutral models of molecular evolution are today widely used to directly measure the selective pressures acting at each position of a protein.
1.1. Nothing in evolution makes sense except over the denominator of neutrality

In both fields, ecology or molecular evolution, neutral processes were found to be responsible of most of the observed evolutionary changes, unleashing up today lively debates between adaptationists and non-adaptationists [Wilkins & Godfrey-Smith 2008]. Thus, and coming back to Pierre Teilhard de Chardin’s quote above, the “curve that all lines must follow” would indeed mostly represent neutral processes, outliers being directional changes. The differentiation of events appearing by successive random walks from those favoured by natural selection, only possible after the definition of a neutral model, is now crucial in both genetics and ecology.

All this thesis and the work that was done during the years preceding its publication is grounded on the definition, and posterior detection, of the neutral processes beyond the formation of the genomes and in the phenotypic consequences of its translation into proteins. The definition of the neutral substrate constitutive of the genomes will be analysed under three perspectives:

- **Informational**: this is perhaps the most simple view of a genome. It considers only the sequence of A, T, G and C nucleotides as the fundamental and independent units constituting genomes. Under this perspective, genomes are assimilated to a simple sequence with a given informational content. Defining a common structure of genomes across the diversity of life will be our first objective that would incorporate as a first step in the main outline of the description of the genomic substrate.

- **Ecological**: while in genomes description of abundances and diversity of genetic elements (mostly transposable elements) are still in their early, ecologists already implemented neutral models able to predict ecosystems composition. In this context it may seem natural to apply ecological models [Hubbell 2001, Etienne 2005] to the different genetic elements populating our genomes in order to bring out the neutral patterns explaining their dynamics.

- **System**: this last perspective under which we are going to analyse the genomes, is phenotypical. In particular we will focus on proteins working together, to complete a biochemical pathway, or to fulfil a molecular function. In the case of protein-coding genes, neutrality is precisely defined [Kimura 1985], and we are able to test with precision for any deviation. However, at the level of functionally related genes, classical methodology fails to find natural selection’s fingerprint.
1.2. What is DNA? – Defining a genome.

A genome represents the overall hereditary information of an organism. Contrary to what suggested its etymology – the blend of the words *gene* and *chromosome* – a genome includes all genes but also non-coding sequences found along the ensemble of chromosomes. The information conveyed by a genome is encoded either in DNA or in RNA, in all cases, this biological codex is constituted of 4 nucleotide bases conventionally represented by the letters A, T, G and C.

Also the codification of biological information is universal among all living species, variation in structure and amount are broad.

1.2.1. Biological complexity versus genome size

Darwinian evolution do not encompass directional change or global adaptive improvement. And in terms of biological complexity, this statement is supported by the diversity of life remaining at the lower levels of complexity. Also it is true that among the whole range of living species, from most ancestral bacteria found in the Chloroflexi’s phylum [Cavalier-Smith 2006] to mammals (Figure 1.1), some evolutionary paths suffered directional gains in complexity. However, even in these edges of gradual acquisition of more and more complex traits, the neutrality of these changes is hardly rejectable [McShea 1996]. We can mention the active debate around the ratchet-like model of constructive neutral evolution (CNE) [Gray et al. 2010, Stoltzfus 1999] originated in the study of little systems like RNA editing. Up today, this model still suffers some criticisms, being considered as too simple [Speijer 2011] (see also the reply [Doolittle et al. 2011]).

The main problem when trying to define if natural selection is favouring or not the increment in biological diversity is precisely the definition of the complexity at organismal level. At genomic level, it would be expected that the quantity of hereditary information would be proportional to the level of complexity of organisms.

However, since the first measures of genome size [Vendrely & Vendrely 1948] it was quickly rejected that the amount of DNA or $C$-value (amount of DNA found in an haploid nucleus) correlates with neither organism complexity nor even with the number of genes [Mirskey & Ris 1951]. This contradiction was referred to as the $C$-value paradox [Thomas 1971] with its most striking example the *Amoeboae dubia* holding 200 times more DNA than humans [Friz 1968, McGrath & Katz 2004]. Recent most famous work being done by T. Ryan Gregory [Gregory 2001] with the introduction of a nuance preferring the term of $C$-value enigma to refer to it. With higher number of species measured [Gregory 2012], the paradox is nowadays more conspicuous than ever (the spectrum of $C$-values now extends between 0.002pg for parasitic microsporidium *Encephalitozoon intestinalis* and}
1.2. What is DNA? – Defining a genome.

Figure 1.1.: Overview of the tree of life.
Picture adapted from the Tree of Life Web Project ©2007 [Maddison & Schulz 2007].
In this picture we see how biological complexity can raise from unicellular Universal
Common Ancestor (bottom of the tree), to tips representing living species, with, in
some cases, the acquisition of multicellularity, division of labor, evolution of meiosis,
sexual reproduction, cell differentiation, early arrest of reproductive cells, etc.

1,400pg for the free-living amoeba Chaos chaos), that is, almost 7 orders of magni-
tude (Figure 1.2).

Questions raised by the imbalance between the C-value and the number of
genes, or the biological complexity or even clade specificity (see Figure 1.2) can
be summarized in these points [Gregory 2005]:
Chapter 1. Introduction

Figure 1.2.: C-values of the main groups of life.
Variation in genome size within and among the main groups of life, adapted from [Gregory 2005].

1. What types of non-coding DNA are found in eukaryotic genomes, and in what proportions?
2. What is the origin of non-coding DNA, how is it spread and/or lost from genomes over time?
3. What effects, or perhaps even functions, does this non-coding DNA have for chromosomes, nuclei, cells, and organism phenotypes?
4. Why do some species (for example birds) exhibit remarkably streamlined chromosomes, while others possess massive amounts of non-coding DNA (like salamanders)?

Methodologically these questions appears to be divisible in two problems, the search of patterns in the informational content of genomes along the diversity of life, and the dynamics underlying the distribution and appearance of non-coding DNA. This two questions constitute the first two chapters of this thesis.
1.2. What is DNA? – Defining a genome.

1.2.2. Informational content of DNA

From a biological perspective it is almost impossible to imagine DNA as a random
mix of A, T, G and C nucleotides. Genomes contain the entirety of our hereditary
information coded either into DNA (or RNA for some viruses). They are composed
of functional elements as can be protein-coding genes, or promoters but also by
non-functional elements like repetitive elements that by definition are the exact
opposite of a random structure in computational terms. However, and challenging
a bit established concepts, we could ask: To what extent can we state that genomes
are not a random soup of 4 letters? Intuitively we may be tended to assume that,
at least in Eukaryotes, DNA present a quite simple structure weighted by the
bulk of repetitive genetic elements (GE) populating genomes. And if we divide
sequences in functional categories, a fair assumption would be that protein-coding
gene would represent a specific selection of nucleotides with surely the highest
informational content, while introns would tend more to random assembly, and
finally simple repeats would present some biases towards 2 or 3 nucleotides (e.g.: CpG islands) surely lowering dramatically sequence complexity.

One point of the C-value paradigm is that some species with similar level of
complexity or number of genes present large differences in genome size. These
differences are usually explained by the spread of repetitive genetic-elements or
large/small-scale genomic duplications [Gregory 2005]. Thus, one could expect
that, reducing the importance of repetition in the measure of genome size would
correct the defective correlation between DNA content and organism complexity.
And methodologically, a quick way of achieving this would be to use the size
of compressed genomes as was proposed by Ryan J. Taft [Taft et al. 2007], as
data compression algorithms are precisely taking advantage of the presence of
repetitions.

Difference between compressed genome size and real genome size, could then be
viewed as genome complexity. In contrast with biological complexity, the measure
of DNA complexity is generally better accepted, and give us a tight-fitting tool
to decipher DNA structure. Here, we can mention the work been done, using
different measures of DNA complexity, by Ryan J. Taft et al. [Taft et al. 2007]. In
their publication, they proposed a relationship between biological complexity and
compressed genome size hypothesizing that simple organisms compress better than
complex ones. And an other study conducted by Zhandong Liu et al. [Liu et al.
2008] deduced from the comparison of the occurrence of n-mers across genomes of
7 species that human genome is not perfectly random as it lacks some fixed-length
sequence (mandatory in a random context). Exact opposite conclusion was raised
in the study of the statistical structure of one piece of human chromosome [Azbel’
1995] and of bacteriophage lambda. Here the main conjecture raised by the work
was that statistical structure of DNA is universal for life.

The first part of this thesis stands also ahead of calculating DNA complexity,
but, as main point it takes advantage on many recently sequenced genomes and on getting a global measure, getting one single value per genome instead of using many local estimations. The complexity value that we define to describe the statistical structure of DNA is based on a classical method for data compression [Adjeroh et al. 2008], it is able to detect regularities due to repetitions among data. Almost same measure was already used in [Holste et al. 2001] but only applied to DNA chunks of human chromosome 22.

The results presented in this thesis (chapter 3: Random-like structure of DNA) show that along all the diversity of life covered by our set of species, the ratio between the complexity and the sequence size is almost maximal in genomes and chromosomes. Notable exceptions are recent polyploids.

1.2.3. Dynamics of genetics elements

From the first analysis of the human genome [Lander et al. 2001] we have significant clues about the proportion of each of the families of genetic elements Figure 1.3.

![Figure 1.3](image)

**Figure 1.3.: Genomic components of human genome.**
Relative proportions of major families of different genomic elements (GE) in the human genome according to [Lander et al. 2001].

These proportions of genetic elements in the human genome were contrasted in the years following 2001 with others genomes sequenced, revealing an important variation among organisms coming into the picture. As an example, Figure 1.4 shows the variation in proportions of the 2 major families of transposable elements (TEs) in different eukaryotic species.
1.2. What is DNA? – Defining a genome.

Figure 1.4: Relative amount of retrotransposons and DNA transposons in diverse eukaryotic genomes.

This graph shows the contribution of DNA transposons and retrotransposons in percentage relative to the total number of transposable elements in each species. Species abbreviations: Hs=Homo sapiens, Mm=Mus musculus, Ce=Caenorhabditis elegans, Dm=Drosophila melanogaster, De=Drosophila erecta, Ag=Anopheles gambiae, Aa=Aedes aegypti, Ed=Entamoeba dispar, Eh=Entamoeba histolytica, Ei=Entamoeba invadens, Em=Entamoeba moshkovskii, Sc=Saccharomyces cerevisiae, Sp=Schizosaccharomyces pombe, Os=Oryza sativa japonica, At=Arabidopsis thaliana, Gi=Giardia lamblia, Tv=Trichomonas vaginalis. Adapted from [Pritham 2009].

In the second part of this thesis we wanted to focus on the dynamics beyond the dispersion of the genetic elements in a broad array of eukaryotic genomes. Defining as “genetic species” all kinds of non-coding DNA according to classic families of repetitive elements [Wicker et al. 2007, Kapitonov & Jurka 2008], and also coding sequences classified into biotypes [Flicek et al. 2011].

Simile with ecology

When describing ecosystems, ecologists usually focus on living species’ natural environment, and their distribution and abundances within it. One common pattern that raises when studying different ecosystems, is that whatever environment studied, at whatever trophic level, it seems to be universal the case where few species are “dominating” the ecosystem, comprising the majority of the individuals, while the majority of species are relatively rares [Preston 1948, Fisher et al. 1943]. Thus, the question that raises is:

What mechanisms control this uneven distribution of species abundance in ecological communities?
Chapter 1. Introduction

This problem, reduced to the study of species diversity and abundance, is one of the oldest and more active topic in ecology [McGill et al. 2007], or citing Charles R. Darwin again:

“When we look at the plants and bushes clothing an entangled bank, we are tempted to attribute their proportional numbers and kinds to what we call chance. But how false a view is this!” [Darwin 1872]

In order to validate this vision and to actually infer the influence of chance in the increase or decrease in abundance of some species, ecologists have been implementing more or less complex models including or not parameters related to species fitness. Roughly speaking, ecological models of species abundance are of two kinds: descriptive (statistical-based) or mechanistic (niche-based or neutrals). While many mechanistic approaches assume ecological niche differences as the main cause driving community composition, neutral models assume that niche differences are null [Magurran 2004].

The unified neutral theory of biodiversity (UNTB) [Hubbell 2001, Rosindell et al. 2011] is a neutral-stochastic theory originally inspired in population genetic [Kimura 1985, Wright 1931], it assumes that individuals among trophically similar species are ecologically identical. This provocative assumption implies that individuals, regardless of the species specificity, are controlled by a common birth, death, dispersal, and speciation rates. The model is thus able to predict the species diversity pattern according to very few parameters. Indeed, the observed values of number of species, the total number of individuals, and 2 extra parameters describing the species richness and the migration rate are sufficient to model the species abundance diversity in a neutral context. The most important being the fundamental biodiversity number ($\theta$). $\theta$ is analogous to the $4N\mu$ of population genetics, it governs species richness in spatial and temporal scale. Under neutral model an other parameter can also be estimated in order to take into account a specific migration rates $m$ (see Fitting Neutral Ecological models, page 29, for details about variations in Hubbell’s neutral model).

A community is then a group of species whose competitive interaction strengths are determined by their niche overlaps, and new species originate through adaptation to new niches. This view was challenged by Robert H. MacArthur and Edward O. Wilson with their equilibrium theory of island biogeography [MacArthur & Wilson 1967], which was finally extended by Stephen P. Hubbell [Hubbell 2001].

In ecology, neutral model is thus a useful null model against to test alternative biological hypotheses of the relative species abundance distribution [Volkov et al. 2003, Alonso et al. 2006]. In non-neutral model, species are considered to be ecologically different, with more or less differences between niches. One simple step to move away from UNTB’s definition of neutrality (while remaining within the scope of neutrality) is, for example, to assume that the fitness or death rate
of a species is dependent of its abundance (see addition of $\delta$ parameter to neutral model [Jabot & Chave 2011]).

Since the first mention of selfish DNA [Dawkin 1976, Doolittle & Sapienza 1980, Orgel & Crick 1980] the idea that TEs and genes should be considered as living entities is recurrent in bibliography – for example referring to TEs as:

“Tiny organisms [...] that survive by spreading their progeny on host chromosomes” [Leonardo & Nuzhdin 2002]

As well as for this example, the attention is usually placed only on TEs, considering incidental the interaction with the remaining repetitive sequences or even genes. Nonetheless, dynamical ecological models were successfully applied to genomes considering TEs interactions [Abrusán & Krambeck 2006, Leonardo & Nuzhdin 2002, Le Rouzic et al. 2007a]. Some of these models –the more complexes– include interactions like parasitism, competition and cooperation between different families of TEs.

The exclusive utilisation of TEs in these models is certainly a consequence of the foggy or unsolved relations between the other genetic species, and also of the difficulty to glimpse a “living organism” in each of them (as it is feasible for TEs, see Table 1.1).

However, ideal models of genomics would consider not only TEs, but also all diversity of GEs populating eukaryote genomes: satellites sequences, DNA-transposons, LTRs-retrotransposons, LINEs, SINEs (retroposons), miRNA, rRNA, tRNA, and genes among the many functional and non-functional elements. Such model, yet already imagined [Le Rouzic et al. 2007b], does not exist for genomes.

<table>
<thead>
<tr>
<th>Population genetics</th>
<th>Ecology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host species</td>
<td>Ecological niche</td>
</tr>
<tr>
<td>Host individual</td>
<td>Patch</td>
</tr>
<tr>
<td>Host genome</td>
<td>Habitat</td>
</tr>
<tr>
<td>A TE family</td>
<td>A species</td>
</tr>
<tr>
<td>Autonomous TE</td>
<td>Host</td>
</tr>
<tr>
<td>Non-autonomous TE</td>
<td>Parasite</td>
</tr>
<tr>
<td>Reproduction</td>
<td>Migration</td>
</tr>
<tr>
<td>Copy number in a genome</td>
<td>Number of individuals in a patch</td>
</tr>
<tr>
<td>Transposition rate</td>
<td>Birth (growth) rate</td>
</tr>
<tr>
<td>Deletion rate</td>
<td>Death rate</td>
</tr>
<tr>
<td>Natural selection</td>
<td>Density-dependent mortality</td>
</tr>
<tr>
<td>TE sequence</td>
<td>Genome of an individual</td>
</tr>
</tbody>
</table>

Table 1.1.: Analogies between population genetics models of TEs dynamics and inter-specific relationships in ecology.
Table reproduced from [Le Rouzic et al. 2007b]
Chapter 1. Introduction

The definition of “species”

In biology, species are defined as the basic unit of biological classification. The limit between one species and an other is classically defined from the observation of sexual capabilities, or reproductive isolation. Thus, one species is defined as the ensemble of individuals able to engender fertile offspring by interbreeding [Mayr 1942]. It is well known that this definition is almost impossible to apply literally for most of the living organisms as they reproduce asexually. Moreover, within sexual species, at both ends of the range of sex “quantity” variation from cases of “too little sex” (thelytoky, observable in arthropods or some lizards) up to “too much sex” (hybridization), the classical definition of species runs into some difficulties [Templeton 1989].

Likewise ecological communities, eukaryotic genomes contain a variable number of more or less abundant elements of different genetic classes: transposon-derived elements, satellite repetitive sequences, and their less abundant functional sequences such as RNA or genes. Here, in order to follow up our simile with ecological systems we had to decide which entity should be considered as a “species” in genomes. The decision was not trivial nor categorical, but we decided to use the last level (with lower number of individuals) of classification that allows a functional definition of the sample (with no direct description of the sequence itself). We believe that the level of hierarchy that corresponds to this condition is the “family” or “class” level according to the RepBase ontology, also referred to as “superfamily” according to the International Committee on the Classification of Transposable Elements (http://girinst.org/conference/ICCTE.html) [Kapitonov & Jurka 2008]. For simplicity we will talk of genetic species (GSs), putting together all repetitive elements and also biotypes. Finally, each of the element belonging to a GS will be referred to as an individual.

Application of ecologist’s methodology on genomic data

Here, taking advantage of the methods and models developed by ecologists we suggest 3 main questions: 1) is there a common pattern behind the relative abundance and diversity of GSs in genomes? 2) in the case that such pattern exists, is it sufficient to explain together diversities of functional and non-functional components in eukaryote genomes? and 3) To what extent abundance and diversity of genome components reflects adaptive or stochastic outcomes? Here we test the statistical adjustment of the UNTB predictions 31 eukaryotic genomes.

To achieve this objective we discuss results in three different sections.

- First we describe the shape of GSs distribution in genomes and chromosomes using the relative species abundance (RSA) curves, classical graphical tools used in ecology. We also test the role of chance in the rise of these observed
1.3. Genomic study of selective pressures in set of genes

shapes through simulation of random distributions of GSs among chromosomes.

- Second we apply an other classical methodology of ecologists, and calculate the species-area (or GSs-chromosome length) relationship.

- Third, we test the statistical adjustment of the neutral ecological theory of biodiversity to the relative abundance and diversity of GSs of eukaryote chromosomes.

We conclude that abundance and diversity of GSs in most of the chromosomes studied is predicted by the stochastic dynamics of a model for which the principle of functional equivalence among elements is the primary assumption. Additionally, we extend this observation through a test of neutrality confirming that a strong neutral component is behind the distribution and diversity of GSs in chromosomes. Finally we hypothesize that at large temporal and spatial scales across all classes of GSs, an overarching neutral or nearly neutral process governs the evolution of abundance and diversity of GSs in eukaryote genomes.

1.3. Genomic study of selective pressures in set of genes

In the past years, with the development of genomic data, evolutionary genomic researchers has been trying to detect signals of selective pressure in a growing set of genomes as they were sequenced. The methodology applied was based on the measure of significant deviation from neutrality for a given gene (methodology that have been used since [Kimura 1985]). This approach, conceived for the study of a single gene, was successfully able to detect genes escaping from neutrality ($\omega \neq 1$ see Equation 1.1 page 15), and in particular positively selected genes (PSGs) (with $\omega > 1$) [Arbiza et al. 2006, Bakewell et al. 2007, Bustamante et al. 2005, Clark et al. 2003, Nielsen et al. 2005]. However, none of these works were able to find a significant enrichment of a given functional trait, within the groups of genes detected to be under positive selection. Nevertheless, taking all these results together, assiduous readers could perceive some patterns raising (Figure 1.5). For instance, functional terms related to Sensory perception, Immune response or Regulation of transcription, were present in almost all genomic studies of positive selection conducted in primates or rodents genomes.

1.3.1. Detection of adaptation at molecular level – single gene approach

The amount of selective pressures acting on a given DNA sequence is usually inferred by comparing the number of changes able to be reflected on phenotypes,
Chapter 1. Introduction

Figure 1.5.: Cloud of Functional categories enriched in PSGs in first genomic studies.

Cloud obtained by pulling together functional categories closest from significance (e.g.: significant before correction for multiple testing) in first genomic studies conducted in primates and rodents [Arbiza et al. 2006, Bakewell et al. 2007, Clark et al. 2003, Nielsen et al. 2005].

with the number of changes observed in regions known to be escaping from natural selection.

In the specific context of coding regions of the genome, changes occurring at nucleotide level can be divided in two categories depending on whether they will be reflected in translated protein sequences or not (respectively non-synonymous or synonymous changes). Even if several works outlined the footprint of natural selection in biases of synonymous changes through codon usage (see reviews [Hershberg & Petrov 2008, Plotkin & Kudla 2011]), it is still assumed that natural selection’s stranglehold on those silent sites is weak [Yang & Nielsen 2008] and its usage as proxy for neutral mutation rate is successfully used since 1980 [Miyata et al. 1980].

On the other hand, the rate of non-synonymous mutations is assumed to be related to selective pressure as mutations occurring at those sites may have functional consequence by changing protein structures or biochemical properties. Moreover non-synonymous mutation rate is significantly lower and more heterogeneous than the rate of silent mutations. The observation of these differences in rates is indeed the typical footprint of purifying selection [Kimura 1985], as most of deleterious mutations, all appearing in non-synonymous sites, are expected to be cleaned up by natural selection.
1.3. Genomic study of selective pressures in set of genes

Thus, assuming the proxy that silent mutations are neutral, the comparison of synonymous and non-synonymous mutation rates makes codons a perfect case in point for measuring the effect of natural selection within DNA sequences. Selective pressure can therefore be directly deduced from the ratio of non-synonymous mutation rate \( (dN) \) over synonymous mutation rate \( (dS) \), namely, the \( \omega \) ratio:

\[
\omega = \frac{dN}{dS}
\]  

(1.1)

This ratio is estimated in coding regions and used to test for neutrality through different statistical methods [Nielsen 2001]. However, in the context of genomic studies, the methodology do not take into account if genes works independently or in association with others to produce a single phenotypic response. In this sense, we are applying pre-genomics concepts and methods to genomics data.

1.3.2. Identification of selective pressures in the genomic era

The current paradigm for large scale analysis of adaptation consists in a two steps framework: first, the definition of a list of genes (in a gene-by-gene framework analysis) with a statistical significant signal of positive selection \( (\omega > 1) \), and second, the search for over-represented functional classes within this list of genes. Although it is logically consistent, it has been noted that this kind of strategy causes an important loss of information due to the large number of false negatives that are accepted in order to preserve a low ratio of false positives necessary when thousands of tests are involved [Al-Shahrour et al. 2007, Al-Shahrour et al. 2005, Al-Shahrour et al. 2006, Subramanian et al. 2005].

Recently, a new methodology was proposed to improve the classical two-step analysis in an attempt to find selective signatures across species at genome-scale level. This methodology simply consists in grouping the signal observed in related species in order to raise statistical power, expecting thus that nearly significant functional terms reported by previous works (Figure 1.5) would hold enough genes to reach significance. This approach was successfully conducted in flies, mammals and a group of gamma proteobacteria [Shapiro & Alm 2008, Clark et al. 2007, Kosiol et al. 2008].

- Grouping PSGs of flies from \textit{melanogaster} group, authors [Clark et al. 2007] were able to identify functional categories showing significant deviations. These categories included \textit{defense response}, \textit{proteolysis}, \textit{DNA metabolic process}, and \textit{odorant binding}, among others.

- In mammals, the most representative functional categories found after pooling together all PSGs (400 genes) found in primates and rodents [Kosiol et al. 2008] were respectively, \textit{chemosensory perception} and \textit{defense/immunity}. 
Chapter 1. Introduction

- In a group of gamma proteobacteria, classical test for positive selection was not applied, however, The authors [Shapiro & Alm 2008] used the deviations from the expected rates of evolution for a large group of genes, to infer selective pressures. The main conclusion brought was that the coherence of selective patterns suggests that the genomic landscape is organized into functional modules independently subjected to natural selection.

Results found by way of grouping PSGs (or fast evolving genes in the case of proteobacterias) of related species together, were able to raise statistical power of the enrichment tests. Functional categories that, until then, were probable candidates (Figure 1.5) lastly reached significance, though this was achieved at the cost of species-specificity.

1.3.3. Back to square one – a new methodology

The hypothesis we aim to test in this study is not about individual genes, but about functional classes. Mutations occur on single genes but natural selection acts on phenotypes by operating on whole sub-cellular systems [Oster & Alberch 1982]. Under a Darwinian view, mutations in genes either remain finally fixed, or disappear depending on their beneficial or disadvantageous effect on individual fitness respectively. This effect on the function of individual proteins can only be understood in the context of the system (e.g. a pathway, functional roles, etc.) in which the proteins are involved. If a list of genes arranged by some parameter that accounts for their evolutionary rates is examined, it is expected that genes belonging to pathways or functional classes favored or disfavored by selection will tend to appear towards the extremes. We call this methodology the gene-set selection analysis (GSSA).

This approach circumvents the implicit assumption posed by the two-step analysis described above, assuming that the gene is specifically targeted by natural selection. However, if natural selection works by means of minor quantitative effects of many different changes distributed along different gene products, most of them working together in a few number of systems (GO functional terms, biochemical pathways), we then expect to find: 1- correlated non-synonymous rate changes associated to these functions, 2- synonymous rate changes not necessarily associated to the same functions, 3- a higher number of significant functions than those discovered in the classical two-step approach.

In the first part of the chapter: Searching for evolutionary patterns in functionally linked group of genes (page 77) we extend the classical two-step approach previously reported by several authors for human and chimp [Arbiza et al. 2006, Bakewell et al. 2007, Bustamante et al. 2005, Clark et al. 2003, Nielsen et al. 2005], to rat and mouse now considering a set of 12,453 orthologous genes of human, chimpanzee, mouse, rat and dog. In order to validate our methodology
outside mammals, we also apply the GSSA in *melanogaster* group, over the 9,240 orthologous genes between *D. yakuba*, *D. erecta*, *D. simulans*, *D. sechellia*, *D. melanogaster* and *D. ananassae*.

The objective of this work is to find functionally related group of genes subjected to a common selective pattern, either conserved or accelerated. We will identify these functional categories and discuss their similarities and differences in relation to the trends brought by the classical two-step approach (Figure 1.5). Finally, as GSSA do not involve directly PSGs, we will test their importance and localization within our results.

1.4. Side-products – Implementation of software and pipeline

1.4.1. Pipeline for the detection of molecular evolution

At the genomic Era, the integration of tools developed by the community is highly topical, bioinformatics as it emerged is gradually disappearing [Ouzounis 2012], or at least evolving. Tomorrow bioinformatics will certainly be, in a large proportion, advanced computer scientists.

In the last section of this thesis I am going to emphasize on different tools implemented in the context of improving and facilitating the different steps conducting to the detection of selective pressures in protein coding genes. More than a unique and multipurpose package as can be Galaxy [Goecks *et al.* 2010], BioPython [Cock *et al.* 2009] or PyCogent [Knight *et al.* 2007], the suite of tools presented here corresponds to specific answers to the different steps of the classical pipeline leading to the detection of selective pressures acting in genes.

This pipeline can be segregated in five main steps itemized as:

- **Species selection:** first of all we need to define a seed species or sequence. Starting from this seed, we have to select a set of sequences not to distant from our seed in order to avoid saturation of synonymous changes [Gojobori 1983, Smith & Smith 1996] (as an example, from human, we should remain within mammals). Usually it is recommended to have at least 4 sequences [Yang 2009].

- **Homologous sequences retrieval:** once selected our set of species the next step consists in retrieving homologous sequences. Most popular alternatives being Ensembl [Flicek *et al.* 2011], the database resources of the NCBI [Sayers *et al.* 2011], or the UCSC database [Fujita *et al.* 2011].

- **Alignment:** importance of hits step is often underestimated. However when measuring selective pressure, and in particular for the detection of positive
Chapter 1. Introduction

selection, misalignment have a high weight in the proportion of false positives. Most popular tools used here are MUSCLE [Edgar 2004], MAFFT [Katoh et al. 2005], Dialign [Subramanian et al. 2008] or T-COFFEE [Notredame 2010]. Alignments, once calculated, are usually trimmed with Gblocks [Talavera & Castresana 2007] or Trimal [Capella-Gutiérrez et al. 2009] in order to abnormally divergent remove columns from the alignments. Considering to pay a special attention to the generation of accurate alignments is crucial in the sense that misaligned regions may be interpreted as fast evolving regions, and bias strongly the detection of positive selection. Moreover this step being from far the fastest computationally, the cost to ensure fair alignments would be generally low.

- **Phylogenetic reconstruction**: this step is unnecessary when species tree is known (what now represents most of the cases). In any case, we will review how to construct fairly phylogenetic trees with a special attention paid to the use of model testing [Posada 2008, Abascal et al. 2005].

- **Identification of selective pressure and testing of evolutionary hypotheses**: this is the final step of our analysis. It concentrate most of the computation time, and was a real challenge to automatize. The programs used, in this work, to calculate selective pressures and fit evolutionary models are SLR [Massingham & Goldman 2005] and CodeML from PAML package [Yang 2007]. We will see how they were integrated into the ETE package [Huerta-Cepas et al. 2010], and how they can be called in order to ease automation.

We will review these steps focusing specially in the last one, the evolutionary hypothesis testing.

1.4.2. Phylemon web server

At the time of the first publication of Phylemon [Tárraga et al. 2007], no other web platform offering the possibility to conduct all the steps making up a complete phylogenetic analysis or to the test of adaptation at molecular level was available. Each one of the steps introduced above in subsection 1.4.1 are indeed part of Phylemon. And besides the presence these utilities, this online platform was proposing a complete set of features allowing to concatenate, store and even visualize a wide range of analysis.

Few years after, other online platforms for phylogenetic analysis emerged proposing Phylemon-like features, namely Datamonkey [Delport et al. 2010], MobylePasteur [Néron et al. 2009] or Phylogeny.fr [Dereeper et al. 2008]. Nuances were in the tools proposed and connections between them. Bolstered up by our experience listening users comments and teaching in phylogenetic courses, we decided
1.4. Side-products – Implementation of software and pipeline

to release a new version of Phylemon adding new tools but, and above all, emphasizing on the integration of the tools together and on developing a complete documentation illustrated by examples.

As main features Phylemon 2.0 proposes (i) an integrated environment that enables the concatenation of evolutionary analysis, the storage of results; (ii) the concatenation of the tools allowing users to follow in their analysis with the step proposed by the server; and finally (iii) an enhanced “pipeliner” that permits to graphically build, save, load or even share complete pipelines.

In the last part of this thesis, we will briefly describe these improvements, and discuss their implication in scientific community.
2. Material and Methods

2.1. Measuring DNA complexity

2.1.1. The complexity ratio and complexity value

Complexity ratio (CR) is defined by a classical formula used in data compression [Adjeroh et al. 2008]. It is the result of three transformation steps of a given sequence. First, the Burrows-Wheeler transform (BWT) [Burrows & Wheeler 1994], second the Move To Front (MTF) [Ryabko 1980] algorithm and finally the summarizing of the unexpected dispersion of the values obtained through Shannon’s entropy [Shannon 1948] (see Table 2.1 for an example of the process).

Thus the CR is Shannon’s entropy of a transformation or digestion of the sequence. The purpose of this transformation is to reveal the regularities in a sequence. Shannon’s entropy is zero –this is the minimum– only when a sequence consists just of a single repeated symbol, which is the simplest possible combinatorial structure. Conversely, when entropy is equal to one (the maximum entropy), it indicates that the sequence has a random-like combinatorial structure.

Algorithmically, the BWT of a given sequence summarizes all its permutations sorted lexicographically. The MTF transforms a given sequence into a list of numbers. The higher the number, the less the character was used in the previous part of the sequence of length equal to the number of characters found in the sequence (in the case of DNA, this stack contains 4 characters). The MTF operates from left to right. Each number returned is an index in the stack and denotes an alphabet symbol. Shannon’s entropy maps a sequence into a real number between zero and one. It weights the frequency of the alphabet symbols in a given sequence. For each symbol $i$ in the alphabet, let $p(i)$ be the probability of finding $i$ in the sequence $s$; $N_i$ the number occurrences of $i$ in $s$ and $\text{length}(s)$ the total length of the sequence $s$:

$$p(i) = \frac{N_i}{\text{length}(s)} \quad (2.1)$$

For DNA alphabet entropy is defined as:

$$E(s) = - \sum_{i=0}^{3} p(i) \times \log_4(p(i)) \quad (2.2)$$
Chapter 2. Material and Methods

Given a sequence, \( seq = \text{AACCTTCGTA} \text{GCATG} \): |

<table>
<thead>
<tr>
<th>#</th>
<th>Rotating sequence</th>
<th>( I. )</th>
<th>BWT</th>
<th>Char. list</th>
<th>MTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>AACCTTCGTAGCATG</td>
<td>0</td>
<td>G</td>
<td>G</td>
<td>a t c</td>
</tr>
<tr>
<td>1</td>
<td>ACCTTCGTAGCATG[A]</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>g T c</td>
</tr>
<tr>
<td>2</td>
<td>CCTTCGTAGCATG[AA]</td>
<td>5</td>
<td>T</td>
<td>T a g c</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>TTCGTAGCATG[AAAC]</td>
<td>7</td>
<td>C</td>
<td>C t a G</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>TCGTAGCATG[AAACCT]</td>
<td>15</td>
<td>G</td>
<td>G c t A</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CGTAGCATG[AAACCTT]</td>
<td>13</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>GTAGCATG[AAACCTT]</td>
<td>6</td>
<td>T</td>
<td>A g c T</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>TAGCATG[AAACCTT]</td>
<td>11</td>
<td>C</td>
<td>T a g C</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>AGCATG[AAACCTT]</td>
<td>12</td>
<td>G</td>
<td>C t a G</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>GATGG[AAACCTT]</td>
<td>9</td>
<td>T</td>
<td>A g c T</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CATG[AAACCTT]</td>
<td>4</td>
<td>C</td>
<td>T a g C</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>ATGG[AAACCTT]</td>
<td>3</td>
<td>G</td>
<td>C t a G</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>TGG[AAACCTT]</td>
<td>14</td>
<td>T</td>
<td>G c T a</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>GGG[AAACCTT]</td>
<td>10</td>
<td>A</td>
<td>T g c A</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>GGG[AAACCTT]</td>
<td>8</td>
<td>C</td>
<td>A t g C</td>
<td></td>
</tr>
</tbody>
</table>

\( CR(seq) = E(MTF(BWT(seq))) = E(0, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3, 3) = 0.593 \)

Table 2.1.: CR explained by example.

These 3 tables summarizes the steps followed to obtain the final sequence of number from which we will finally compute Shannon’s entropy. 1) The table on the left corresponds to the Burrows-Wheeler transform (BWT). Original sequence is rotated sequentially (so that first character moves to back) resulting in different strings, as many strings as characters in the sequence. The resulting sequences are then sorted in lexicographic order. The “\( I. \)” column corresponds to the Index of this ordering (e.g.: the third sequence here in original order “#” takes the fifth position in lexicographic order). 2) The table in the center corresponds to the result of the BWT, that is the last character of previous sequences ordered. 3) The table on the right corresponds to the application of the MTF algorithm. Starting from a sequence of all characters named here “Char. list” (four nucleotides in this case), the Move-to-front (MTF) algorithm will return the index of BWT’s nucleotide (upper case bold letter) in the “Char. list”. In a second step, for the next iteration, MTF will transform the “Char. list” bringing to front corresponding BWT character (upper case letter). Finally, we compute the Shannon’s entropy of these last values obtained (line below tables) that results in our CR (the CV is obtained by multiplying CR by the length of the sequence).

With \( i \) the index of characters used, for nucleotides from 0 to 3. Thus the CR can be factorize as:

\[
CR(s) = E(MTF(BWT(s)))
\]  \hfill (2.3)

The complexity value (CV) of a sequence is its CR times the number of characters in this sequence (here \( s \)):

\[
CV(s) = E(MTF(BWT(s))) \times length(s)
\]  \hfill (2.4)
As the CV of a sequence depends on the transformation of the MTF applied to the whole sequence, its computation impede the use of parts of the sequence independently.

2.1.2. Complexity in strings

Genomic sequences

Complete genomes of 54 species were downloaded from NCBI database resource [Sayers et al. 2009] and Ensembl Genome Project [Flicek et al. 2011]. Fourteen major groups of taxa were selected: virus, phages, bacteria, archaea, fungi, am-plicomplexa, heterokonta, amebozoa, urochordates, invertebrates, plants, fishes, birds, and mammals. Species among taxa were chosen to their interest as model species or the presence of particular biological features such as: variation in genome size, ancestral or recent polyploidy, living in extreme environments, living as intracellular parasites, gene expansion, genome reduction, RNA or single-strand DNA genomes, and synthetic genomes Table 3.1. Eukaryote genomes with coverage of 6× or greater were chosen. Sexual chromosomes were excluded from the analysis, and ambiguous “N” characters were removed from sequences, and thus, excluded in the measure of chromosome length. Eukaryote’s genome complexity was calculated over concatenated chromosomes.

Complexity in biological sequences was computed in the +1 strand. Analysis of -1 strand provided no significant differences in results.

Random sequences with different ploidy levels were also needed for the study, they were generated with Python base function: “random” [Van Rossum & Drake 2003]. Complexity value of biological and random sequences was computed with the DNA alphabet of four letters.

Annotation of repetitive elements

Interspersed repeats and low complexity DNA sequences were screened and mapped in genomes of all our 54 species using RepeatMasker program [Smit et al. 2010] (see details of RepeatMasker output for individual species in Appendix A). Libraries of genetic elements were retrieved from RepBase [Jurka et al. 2005], the last version available at time was used (20th of September 2011). An example of summary file given by RepeatMasker is shown in Appendix A.

Complexity of major families of repetitive elements such as DNA transposons, LTRs, LINEs, SINEs, satellites and exons, introns, and complete genes (considering untranslated regions) was computed after concatenation of all elements conserving their original order in chromosomes.
Chapter 2. Material and Methods

Human texts

Short stories, books and complete works in its original languages were downloaded from Project Gutenberg (http://www.gutenberg.org/). To automatically detect the alphabet size in texts (including mathematical and punctuation symbols) we run COMPL program (http://kapow.dc.uba.ar/compl) with “auto” option, that takes into account all characters found, including mathematical symbols, and different punctuation signs.

Complexity in windows

To study complexity along chromosomes, a sliding window method shifting along chromosomes in overlapping units of 1.0 Kb to 100 Mb was performed.

2.1.3. Simulations

We performed four kinds of experiments where CV and CR were computed. First: random polyploid construction of sequences of various sizes and ploidy levels (1× to 10×). Second: the evolution along 40 million generations by constant neutral mutation rate of $1.0e^{-08}$ mutations per site per generation (this value is in between the mutation rate estimated for Homo sapiens: $2.5e^{-08}$ [Nachman & Crowell 2000], and Arabidopsis thaliana: $7.1e^{-09}$ [Ossowski et al. 2010]) over random sequences, and chromosomes of Zea mays and Sorghum bicolor. Third: the evolution along 50,000 generations of random polyploid genomes of different sizes (100Kb, 1Mb, 10Mb) by 1.0 Kb transpositions between chromosomes. The number of transpositions per generation was set as a constant function of genome size (genome size over 1,000). Last: the concatenation and shuffling (computed with the Python base function: “shuffle”) of all repetition instances in chromosomes for main repetitive families, and genes were considered. CV and CR were calculated every 100 generations.

2.2. Measuring dynamics of genetic species

2.2.1. Genomes

For the study of dynamics of genetic elements, genomic sequences of 31 species from unicellular eukaryotes to mammals were used. These genomes correspond to a subset of the 54 genomes presented in previous section (see subsection 2.1.2). References can be found in and Table 3.1, page 49. The complete list of species used is: 1) Gallus gallus (Birds) 2) Taeniopygia guttata (Birds) 3) Danio rerio (Fishes) 4) Oryzias latipes (Fishes) 5) Tetraodon nigroviridis (Fishes) 6) Saccharomyces cerevisiae (Fungi) 7) Anopheles gambiae (Invertebrates) 8) Caenorhab-
2.2. Measuring dynamics of genetic species

2.2.2. Mining of Genetic Species

For this study, we define genetic species (GSs) as the sum of repetitive elements and functional elements, each of them described below.

Repetitive Elements

Repetitive elements were mapped following the methodology explained in section 2.1.2.

To measure dynamics of genetic elements we had to define a level to consider as “species” in the RepBase ontology (see The definition of “species”: section 1.2.3). We decided to consider “species” those class of repeats that can be defined functionally. What would correspond to superfamilies of transposable elements according to [Wicker et al. 2007] or, also, to RepBase classification [Kapitonov & Jurka 2008].

Complete list of superfamilies mapped is shown in Table 2.2.

Functional Elements

Functional elements correspond to biotypes category of the genes according to Ensembl [Flicek et al. 2011] nomenclature. They were retrieved using the Biomart API [Kinsella et al. 2011]. The non-redundant list of function elements across all species is shown in Table 2.3.

Note that pseudogenes were not included in that list in order to keep the functional aspect of this family of GSs.

2.2.3. Randomization of genetic elements

In order to test for the random distribution of GSs among chromosomes of each genomes, we generated 1,000 genomes, corresponding to each species, with, for each, a random distribution of GSs.
Consequently GSs of each genome were distributed among chromosomes, according to a probability dependent of the size of the chromosome. As an example, in Human, it was around 6 times likely for a GS to belong to chromosome 1 than chromosome 22 (respective lengths are 225 megabases and 35 megabases).

Measure of chromosome size being critical for the randomization process, we discarded centromeric and not fully sequenced regions, instead of considering directly the sequence length. A way to do this is to consider only regions of the chromosome where GS can be found. We thus defined that chromosome size was the sum of all 10 kilobase windows containing at least one GS (see Table 2.4).

### 2.2.4. Ecolopy

Since the definition of neutral models in ecology [Hubbell 2001, Volkov et al. 2003] computational tools were developed in order to manipulate data collected by ecological sampling, and to apply specific statistical test over them. The main tools developed in this sense are three.

- The more complete tool takes its name directly from the neutral model it allows to test, the package untb [Hankin 2007]. This package is implemented in R language [Team 2011] and stands ahead of other R packages allowing
2.2. Measuring dynamics of genetic species

<table>
<thead>
<tr>
<th>Biotype</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IG C gene</td>
<td>Immunoglobulin constant segment</td>
</tr>
<tr>
<td>IG D gene</td>
<td>Immunoglobulin diversity segment</td>
</tr>
<tr>
<td>IG J gene</td>
<td>Immunoglobulin joining segment</td>
</tr>
<tr>
<td>IG V gene</td>
<td>Immunoglobulin variable segment</td>
</tr>
<tr>
<td>IG Z gene</td>
<td>Immunoglobulin gene found in Zebrafish</td>
</tr>
<tr>
<td>MRP RNA</td>
<td>mitochondrial RNA-processing RNA</td>
</tr>
<tr>
<td>RNase MRP RNA</td>
<td>enzymatically active ribonucleoprotein</td>
</tr>
<tr>
<td>RNase P RNA</td>
<td>enzymatically active ribonucleoprotein</td>
</tr>
<tr>
<td>SRP RNA</td>
<td>signal recognition particle RNA</td>
</tr>
<tr>
<td>TR C</td>
<td>T cell receptor constant domain</td>
</tr>
<tr>
<td>TR J</td>
<td>T cell receptor joining domain</td>
</tr>
<tr>
<td>TR V</td>
<td>T cell receptor variable domain</td>
</tr>
<tr>
<td>class I RNA</td>
<td>class of small non-coding RNA</td>
</tr>
<tr>
<td>class II RNA</td>
<td>class of small non-coding RNA</td>
</tr>
<tr>
<td>lincRNA</td>
<td>large intervening non-coding RNA (multiexonic non-coding RNA)</td>
</tr>
<tr>
<td>mRNA</td>
<td>micro RNA</td>
</tr>
<tr>
<td>misc RNA</td>
<td>miscellaneous RNA</td>
</tr>
<tr>
<td>ncRNA</td>
<td>non-coding RNA</td>
</tr>
<tr>
<td>processed transcript</td>
<td>Non-coding transcript without open reading frame (ORF).</td>
</tr>
<tr>
<td>protein coding</td>
<td>Contains an open reading frame (ORF)</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>retrotransposed</td>
<td>non-coding pseudogene produced by integration of a reverse transcribed mRNA into the genome</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>snlRNA</td>
<td>small nuclear like RNA</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNA, involved in modifications of other RNAs</td>
</tr>
<tr>
<td>tRNA</td>
<td>transfer RNA</td>
</tr>
<tr>
<td>transposable element</td>
<td>transposable element</td>
</tr>
</tbody>
</table>

Table 2.3.: Biotype and description.
Summary table of the biotypes used, and short description retrieved from Ensembl glossary [Flicek et al. 2011] and from Sequence Ontology browser [Eilbeck et al. 2005].

<table>
<thead>
<tr>
<th>Chr</th>
<th>Chr length</th>
<th>Corrected length</th>
<th>Percentage left</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>249,240,621</td>
<td>225,200,000</td>
<td>90.35%</td>
</tr>
<tr>
<td>2</td>
<td>243,188,741</td>
<td>237,670,000</td>
<td>97.73%</td>
</tr>
<tr>
<td>6</td>
<td>171,048,878</td>
<td>167,050,000</td>
<td>97.66%</td>
</tr>
<tr>
<td>10</td>
<td>135,524,747</td>
<td>131,040,000</td>
<td>96.69%</td>
</tr>
<tr>
<td>12</td>
<td>133,841,891</td>
<td>129,970,000</td>
<td>97.11%</td>
</tr>
<tr>
<td>15</td>
<td>102,521,389</td>
<td>81,520,000</td>
<td>79.52%</td>
</tr>
<tr>
<td>20</td>
<td>62,962,324</td>
<td>59,430,000</td>
<td>94.39%</td>
</tr>
<tr>
<td>22</td>
<td>51,244,541</td>
<td>34,790,000</td>
<td>67.89%</td>
</tr>
<tr>
<td>X</td>
<td>155,260,558</td>
<td>150,230,000</td>
<td>96.76%</td>
</tr>
<tr>
<td>Y</td>
<td>59,033,288</td>
<td>22,520,000</td>
<td>38.15%</td>
</tr>
</tbody>
</table>

Table 2.4.: Transformation of Chromosome size.
Example of changes in estimation of chromosome length after removing regions with no GSs for some human chromosomes.
to deal with ecological data and to execute most classical statistical analysis that ecologist have been developing [Borcard et al. 2011]. The main restriction of the package being related to the language used that may lacks of computational efficiency.

- Other suite of scripts has been implemented by Rampal S. Etienne in the context of its publication [Etienne 2005]. These scripts were implemented in PARI/GP language, in the unique context of testing the UNTB and are actually mainly taken up by the untb R package. The main advantage of this tool comes down to its simplicity of use and its speed.

- Finally the programs that allows to compute the fit of neutral models with more efficiency are Tetame and Parthy implemented by Franck Jabot and Jérôme Chave [Jabot & Chave 2011, Jabot et al. 2008]. Those programs are implemented in C++ language and are very fast, the only criticism that we could brought is about their lack of flexibility, as a counterpart of their computational efficiency.

All these packages allow to get the fit of sampling data to the UNTB plus, in some cases, the possibility to dwell on most classical ecological statistics. But at the time to test for the UNTB in genomic data, none was able to deal with the high numbers corresponding to the sampling of genomes and chromosomes. And this was the main reason why we started developing a new package “Ecolopy” in order to be able to deal with genomic data.

Ecolopy is a fully functional package in the context of testing for the UNTB, but still lacks of the whole set statistical tools available in R. However its design and the use of python [Van Rossum & Drake 2003] were though in order to provide scalable program architecture, and also the integration, if needed of some algorithms developed in R (through the RPy [Moreira & Warnes 2004] module for example). Alternatively Ecolopy takes advantage of the GNU Multiple Precision (GMP) [Granlund 2000] library, and the Multiple-Precision Binary Floating-Point (MPFR) [Fousse et al. 2007] library, through the fast multiprecision GMPY module.

Usage and source

A complete description of the functionalities as well as a quick tutorial can be found at http://bioinfo.cipf.es/ecolopy/tutorial/load_abundance.html. The source can be downloaded from https://gitorious.org/ecolopy.
2.2. Measuring dynamics of genetic species

2.2.5. Fitting Neutral Ecological models

Ewens sampling formula

Ewens sampling formula [Ewens 1972] (Equation 2.6) was originally designed in order to describe the number of different alleles expected to be observed in a given sample. However, the formula can be applied to other fields. In the context of the study of ecological communities its application was first suggested by Tavaré and Ewens [Tavaré & Ewens 1997] and finally implemented by Hubbell [Hubbell 2001]. Hubbell proposed a model defining the fundamental biodiversity parameter $\theta$ (Equation 2.5) given the speciation rate $\nu$ and $J_M$ the size of the metacommunity.

$$\theta = 2J_M\nu \quad (2.5)$$

The estimation of $\theta$ alone is sufficient to apply directly Ewens sampling formula (Equation 2.6), and to compute its likelihood for given a community (Equation 2.7).

$$Pr\{S, n_1, n_2, \ldots, n_S|\theta\} = \frac{J_M!\theta^S}{1^\phi_12^\phi_2\cdots J_M^\phi_J \phi_1!\phi_2!\cdots\phi_J!\prod_{k=1}^{J_M}(\theta + k - 1)} \quad (2.6)$$

Here $n_i$ corresponds to the abundance of species $i$ and $\phi_a$ the number of species with abundance $a$.

$$L = \frac{\theta^S}{\prod_{k=1}^{J_M}(\theta + k - 1)} \quad (2.7)$$

Etienne sampling formula

The main problem with Hubbell’s model using Ewens sampling formula is the assumption that migration is unlimited ($m = 1$). However a new sampling formula was presented recently [Etienne 2005] including cases where $m < 1$, taking into account the number of immigrants $I$ depending on the sample size $J$:

$$m = \frac{I}{I + J - 1} \quad (2.8)$$

Etienne’s sampling formula is then postulated as:

$$P[D|\theta, m, J] = \frac{J!}{\prod_{i=1}^{S} n_i \prod_{j=1}^{J} \phi_j!} \frac{\theta^S}{(I)!} \sum_{A=S}^{J} K(D, A) \frac{I^A}{(\theta)_A} \quad (2.9)$$

with $K(D, A)$ as:
Chapter 2. Material and Methods

\[ K(D, A) := \sum_{\{a_1, \ldots, a_s\} \mid \sum_{i=1}^s a_i = A} \prod_{i=1}^s \frac{\bar{s}(n_i, a_i) \bar{s}(a_i, 1)}{\bar{s}(n_i, 1)} \]  

(2.10)

Once computed \( K(D, A) \) we are able to optimize the likelihood of the model (Equation 2.9) by varying the values of the parameters \( \theta \) and \( m \) (see Model optimization subsection 2.2.6) for a given dataset.

This last step is actually the main computational bottle neck in the resolution of Etienne’s formula (Equation 2.9), and in particular the calculation the stirling numbers [Etienne 2005] in \( K(D, A) \)’s equation (see Equation 2.10). A solution was given by Franck Jabot and Jérôme Chave [Jabot et al. 2008] and implemented in the Tetame program. It consists in taking advantage of the recurrence function (Equation 2.11), that allows to build a table of values, given the dispersion of the ranked abundance of species, instead of computing them directly for each pair of values.

\[ S_{(n,m)} = S_{(n-1,m-1)} - (n-1) \times S_{(n-1,m)} \]  

(2.11)

However, at the expense of the amelioration in computation time, the size of the table of values created using this methodology was excessive in the case of genomic data. A solution, implemented in Eclopy, was to recursively remove the stirling numbers not needed for the estimation of the \( K(D, A) \).

Finally, given the order of magnitude that can reach the values of stirling numbers (> \( 10^{+1000} \) for medium size chromosomes), we needed to discard mathematics operations provided in Python. To be able to deal with this kind of data Eclopy uses the GMP [Granlund 2000] and MPFR [Fousse et al. 2007] libraries through GMPY biding [Martelli 2007].

2.2.6. Model optimization

In Eclopy models can be optimized through different strategies depending on the model selected. In the case of the Ewens’ formula, \( \theta \) is the only parameter to take into account, and this one-dimensional optimization is achieved with the “golden” optimization strategy [Jones et al. 2001]. For Etienne’s model, two parameters are optimized, \( \theta \) and \( m \). Optimization step being more complicated here, several different optimization strategies (all implemented in SciPy [Jones et al. 2001]), are proposed:

- “fmin”: the downhill simplex algorithm [Nelder & Mead 1965]
2.2. Measuring dynamics of genetic species

- “tnc”: minimize using gradient information in a Truncated Newton Conjugate-gradient [Nocedal & Wright 2000].

- “slsqp”: Sequential Least SQuares Programming [Kraft 1988].

Obviously, the best methodology consists in verifying that all strategies converge. However the use of the downhill simplex algorithm plus one of the other methodologies (that uses bounds) can be sufficient if both maximum likelihoods found converge. This last strategy was thus used in chapter 4.

Alternatively, different starting values of $\theta$ and $m$ can also be passed to the algorithm in order to ensure the finding of the global maximum.

A last step can be performed if computation time is not critical. It consist in drawing a surface of likelihood for a range of values of $\theta$ and $m$ (see Figure 2.1 for an example of likelihood surface).

![Log likelihood surface as a function of migration rate ($m$), and the fundamental biodiversity number ($\theta$) for D. rerio chromosome 19. Dark red color shows regions of the surface where parameters maximize the probability to explain abundances and diversity of genetic elements in the chromosome. Likelihood ratio tests favored Etienne in contrast to Ewens sampling formula to explain the observed data in the chromosome.](image)

Figure 2.1.: Maximum likelihood inference of neutral parameters.
Chapter 2. Material and Methods

Optimization step being critical specially under Etienne’s model, the likelihood surface of the model given a range of values of \( \theta \) and \( m \) was drawn for some of the chromosomes in our dataset. This procedure allows us to estimate graphically the best solution for both parameters. The solution found by this methodology was then compared to the optimization result, in order to validate them (see Figure 2.1 as an example of this validation step). Computation time needed to generate such likelihood contour plots prevents using it for all chromosomes, but, for the 5 chromosomes tested, results were congruent.

2.2.7. Model testing – Likelihood ratio test

In order to compare and test the fit of a given distribution in the two models computed, a likelihood ratio test [Wilks 1938] can be conducted between them. Etienne’s model has 2 free parameters (\( FP \)) while Ewens’ only one, thus the number of degrees of freedom for the chi-squared distribution is 1 (\( df = FP_{Etienne} - FP_{Ewens} = 1 \)).

2.2.8. Testing UNTB

In the last years at least two tests were developed in order to accept or reject the neutrality of a given community. Both tests are based on the comparison of a given number of random neutral communities (or replicates) to the observed distribution of abundances. Replicates being generated using the parameters estimated (see subsection 2.2.6) for the real data under a given neutral model (either Ewens or Etienne model), we expect that they would be very close to the original distribution of abundances. Distances between replicates and original data being thus a measure of how well our data fits in a neutral model.

The first of these tests [Etienne 2007] consists in comparing the values of likelihood to fit neutral model. Random neutral abundances are fitted to the neutral model, and their likelihood is used to build a distribution of likelihoods. Then, this distribution is compared to the likelihood of the observed data. The major problem of this test is technical, the computation time needed to optimize the parameters of each abundance distribution and get the likelihoods is unrealistically too high when dealing with GSs.

The second test [Jabot & Chave 2011] uses, instead of likelihood, the comparison of Shannon’s entropy [Shannon 1948] of each distribution of abundances, and is much faster as replicates do not need to be fitted into a neutral model.

Both of these methodologies were implemented in Ecolopy, however results presented here belong from the comparison of entropies.

From the neutral parameters obtained for each chromosome, we simulated 10,000 replicates and computed, for each, their Shannon’s entropy (\( H \)). Chromosomes were considered significantly non-neutral when the \( H \) of their abundances
2.2. Measuring dynamics of genetic species

Figure 2.2.: Comparing simulated and empirical evenness.
Neutrality test statistically compares simulated null distribution of $H$ with the empirical value. Here, the null distribution of $H$ values were derived from 10,000 neutral simulations of (A) *H. sapiens* chromosome 1 and (B) *A. gambiae* chromosome 2L, with parameters ($\theta$ and $m$) optimized by ML using Etienne sampling formula. Light and dark gray bars display 5% and 95% of the simulated data, respectively. Although neutrality was not rejected in B ($p = 0.291$ and $p = 0.041$ for A and B respectively), posterior correction by multiple testing favored the neutral hypothesis in both cases ($q = 0.609$ and $q = 0.159$ for A and B respectively).
was below 95% of the 10,000 random neutral $H$ values. As an example, Figure 2.2 shows the distribution of $H$ for 10,000 random neutral communities generated under Etienne’s model with $S$, $J$ fixed to the observed numbers and $\theta$ and $m$ corresponding to optimized values for 2 chromosomes.

Additionally, given the large number of test performed (one for each of the 548 chromosome), statistical significances were corrected by false discovery rate (FDR) [Benjamini et al. 2001]. For example, in Figure 2.2B, Anopheles gambiae’s chromosome 2L is brought back to neutrality after correction by FDR.

Given the lack of differences among results presented in the section: Neutrality of species abundances and diversity (page 71), we replicated this test by fixing the number of species ($S$) according to the observed value of each chromosome. No differences were observed in relation to the number of chromosomes fitting in neutral models.

### Power and specificity of neutral test

In order to validate the test of neutrality, we computed the proportion of false and true positives generating respectively random log-normal distributions and random neutral distributions. The results of the test of neutrality applied over log-normal or neutral random distributions are shown in Figure 2.3, respectively.

According to this result, we validate the power given that in the whole range of $S$ and $J$, the proportion of true positives was very high (Figure 2.3B. Nevertheless these simulations pointed out some difficulties to differentiate log-normal distributions from neutral distributions. Specifically, when $J < 100,000$ individuals, the proportion of false positives is higher than 50%. However we decided to get through as this raise in false positive rate only affect the smallest chromosomes, and also because lognormal distributions used here as alternative, are ardly distinguishable from neutral distributions [McGill et al. 2006].

## 2.3. Detection of selective pressure at molecular level

### 2.3.1. Orthology prediction

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]. Orthology predictions between seed genes and genes from other species species (seed species was H. sapiens in the case of mammals) was retrieved from Ensembl Compara [Vilella et al. 2009] using Biomart [Kinsella et al. 2011] (see Figure 2.4 to have an insight of the phylogenies and distances). From the 23,438 seed genes, only groups of orthologs annotated as “one-to-one” –with only one representative of each species– where kept in the final dataset.
2.3. Detection of selective pressure at molecular level

Figure 2.3.: Type I and Type II errors of neutral test in ranges of S and J.
Panels describe the proportion of times the test (A) rejected null hypothesis being true (red regions being dangerous areas to test), and (B) failed to reject null hypothesis being false. Numbers in both panels are chromosomes: 1-A. thaliana chr1 2-D. rerio chr1 3-D. discoideum chr2 4-C. elegans chr1 5-D. melanogaster chr2L 6-G. gallus chr8 7-G. gallus chr2 8-H. sapiens chr1 9-H. sapiens chr1 10-Z. mays chr1 11-Z. mays chr3 12-M. musculus chr0 13-M. domesticus chr1 14-M. domesticus chr3 15-M. domesticus chr5 16-P. falciparum chr3 17-R. norvegicus chr1 18-S. bicolor chr7 19-T. nigroviridis chr9 20-T. castaneum chr8 and ecosystems [Jabot & Chave 2011]: 21-BCI 22-Edoro 23-La Planada 24-Lambir 25-Lenda 26-Mudamalai 27-Pasoh 28-Sinharaja 29-Yasuni.
Figure 2.4.: Mammals and *melanogaster* group phylogeny.
Straight numbers represent the median rates of non-synonymous and synonymous substitutions ($dN/dS$) estimated from all coding sequences compared in mammals (A) and *Drosophila* (B). Branch lengths and rates were multiplied by 100. Ancestral estimations were done in primates (P), rodents (R), *D. yakuba* and *D. erecta* (Aye), *D. simulans* and *D. sechellia* (Ass), and *D. melanogaster*, *D. simulans* and *D. sechellia* (Amss). *C. familiaris* in (A), and *D. ananassae* in (B), were chosen as outgroup species.

The same procedure was applied in *Drosophila*, including 6 species namely, *D. melanogaster* as seed-species, *D. sechellia*, *D. simulans*, *D. yakuba*, *D. erecta* and, *D. ananassae* as outgroup (see Figure 2.4-B). Here, the starting number of seed genes was 14,076.

2.3.2. Alignments refinement and filters
DNA coding sequences (CDS) were aligned according to protein translation pattern using Muscle version 3.7 [Edgar 2004] embedded into the CDS-Protal utility in Phylemon 2.0 [Sánchez et al. 2011] (see section: The alignment at page 92). Poorly aligned regions were removed using TrimAl [Capella-Gutíérrez et al. 2009] keeping all sequences but checking the quality of alignment columns with
the heuristic method “-automated-1”. Additionally, alignments smaller than 100 bp were excluded from the analysis.

In mammals, the upper limit for \(dN\) and \(dS\) considered was those of the human interferon \(\gamma (dN = 3.06)\) and the relaxin protein [Graur & Li 2000] \(dS = 6.39\) substitutions per site per \(1e^{+09}\) years). Assuming the human-mouse, mouse-rat and human-chimp speciation times to be about 80, 70 and 5 million years [Blair Hedges & Kumar 2003] respectively, ortholog comparisons between primates and rodents with \(dS \geq 1\) and \(dN \geq 0.5\), rodents with \(dS \geq 0.256, dN \geq 0.122\), and primates with \(dS \geq 0.064\) and \(dN \geq 0.030\) substitutions per site were excluded.

In *Drosophila* genes were also filtered by high \(dN\) and \(dS\) values using as relaxed reference the fast evolving gene 1G5 for both \(dN\) and \(dS\) [Schmid & Tautz 1997].

The final number of orthologs kept was 12,453 for mammals and 9,240 for flies.

### 2.3.3. Evolutionary analysis

Maximum likelihood estimations of \(dN\), \(dS\), and \(\omega\) and tests of positive selection were computed using CodeML program from PAML package [Yang 2007] through the ETE program [Huerta-Cepas et al. 2010] (see section: ETE’s Evol extension, page 99 for a description of this specific methodology). Evolutionary rates were computed in orthologous sequences according to the free-ratio branch model that assumes independent \(\omega\) ratio for each branch in the trees of mammals and *Drosophila* species. Evolutionary rates \((dN, dS)\), their ratio \((\omega)\), and the difference between ancestral and descendant species \((\Delta\omega)\) were ranked along all genes of genomes and further analyzed by the gene set selection analysis (GSSA).

External branches in Figure 2.4 were marked as foreground to test for positive selection and relaxation using branch-site models in Test I and Test II [Zhang et al. 2005a] (see section: Testing for evolutionary scenarios in protein coding genes, page 96, for more complete overview of these tests). Positive results of relaxation of selective constraints (or weak signals of positive selection) were discarded [Arbiza et al. 2006]. To quantify the relative contribution of positively selected genes (PSGs) in functional modules showing significantly high values of \(\omega\) (SH\(\omega\)) and significantly low values of \(\omega\) (SL\(\omega\)), a t-test (from R package [Ihaka & Gentleman 1996]) with the mean number of PSGs per functional modules was computed in primates, rodents, mammals and *Drosophila* species. An independent set of PSGs was collected to test the robustness of our results in mammals [Kosiol et al. 2008], and *Drosophila* species [Clark et al. 2007].

### 2.3.4. GSSA, evolutionary and statistical simulations

Gene-set selection analysis works over lists of genes ranked by different evolutionary rate parameters (in this case \(dS, dN, \omega\) and \(\Delta\omega\)). Internally it uses the
Chapter 2. Material and Methods

FatiScan tool [Al-Shahrour et al. 2007]. FatiScan is a version of gene set enrichment analysis (GSEA) [Al-Shahrour et al. 2005] which can be applied to any list of ranked genes regardless of the initial experimental design [Dopazo, Huang et al. 2009]. The aim of the test is to find functional classes, namely blocks of genes that share some functional property, showing a significant asymmetric distribution towards the extremes of a list of ranked genes. This is achieved by means of a segmentation test, which consists on the sequential application of a Fisher’s exact test over the contingency tables formed with the two sides of different partitions (A and B in Figure 2.5) made on an ordered list of genes.

The two-tailed Fisher’s exact test (FET) finds significantly over or under represented functional classes (GO and KEGG) when comparing the two sides of the list ranked by an evolutionary variable (in Figure 2.5, 4 of the 5 partitions show significant differences). Previous results showed that a number between 20 and 50 partitions gives optimal results in terms of sensitivity and results recovered [Al-Shahrour et al. 2005]. Here we applied 30 partitions along all the GSSA performed. Given that multiple functional classes (C) are tested in multiple partitions (P), the unadjusted p-values for a total of $C \times P$ tests were corrected by FDR [Benjamini et al. 2001].

Originally, 1,394/1,331 GO terms, and 199/116 KEGG pathways were analyzed in mammals and Drosophila species respectively. The global GO directed acyclic graph was processed with Blast2GO [Conesa et al. 2005] to extend the annotation at missing parental nodes, keeping only GO terms between levels 2 and 8 for mammals, and between levels 2 and 12 for Drosophila. The final set of GO and KEGG terms used was also reduced to those containing at least 15 genes.

Figure 2.5. (following page): Summary of the steps developed by the GSSA.

GSSA can be described in a series of five steps (S1 to S5). S1: rank genes of a genome according to an evolutionary variable (e.g.: $\omega$), S2: assign functional categories, S3: partition the ranked list, S4: proceeds with a Fisher exact test for each partition, S5: adjust p-values by FDR. Colored boxes represent the final result: functional categories found to have values of $\omega$ a) significantly high (SH) appear in red or orange (0.1% and 5% FDR respectively), b) significantly low (SL) in blue and cyan (0.1% and 5% FDR respectively) c) not significant (NS) in white. Example shows five GO terms with significantly and NS biased distributions of $\omega$. In brackets, the number of genes annotated with the GO term. GO:0007517 was NS although, in partition 16 in human (not shown in the picture) its p-value was low, it was NS after FDR correction ($q = 0.065$). Upper (A) and lower (B) sides of the list (S3) represent both sides of the specified partition number. Remainder GO terms (GO2 to GO5) show the association of dark dots with values located at the top (SH$\omega$), and at the bottom (SL$\omega$) of the list (for GO2-GO3 and GO4-GO5, respectively). In examples, Fisher exact tests found the most significant p-value for partitions 8, 14, 22 and 27 for GO:0007186, GO:0009566, GO:0050658 and GO:0022618 in chimpanzee, human, mouse and rat genome, respectively.
2.3. Detection of selective pressure at molecular level
Chapter 2. Material and Methods

Partitioning

Some evolutionary rates presented discontinuous distributions, in particular the \( \omega \) ratio. Partitioning a list by values can be non-sense if this list scales from 0 to infinite. In order to partition more accurately we decided to use the rank order instead of the direct value (see Table 2.5).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Direct partitioning</th>
<th>Partitioning by rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000211454</td>
<td>999.0000</td>
<td>A</td>
</tr>
<tr>
<td>ENSG00000169084</td>
<td>999.0000</td>
<td>A</td>
</tr>
<tr>
<td>ENSG00000159433</td>
<td>999.0000</td>
<td>A</td>
</tr>
<tr>
<td>ENSG00000162430</td>
<td>200.2600</td>
<td>B</td>
</tr>
<tr>
<td>ENSG00000176390</td>
<td>0.2520</td>
<td>C</td>
</tr>
<tr>
<td>ENSG00000156291</td>
<td>0.0520</td>
<td>C</td>
</tr>
<tr>
<td>ENSG00000176711</td>
<td>0.0259</td>
<td>C</td>
</tr>
<tr>
<td>ENSG00000166287</td>
<td>0.0123</td>
<td>C</td>
</tr>
<tr>
<td>ENSG00000174788</td>
<td>0.0067</td>
<td>C</td>
</tr>
</tbody>
</table>

Table 2.5.: Comparison of partitioning strategies.

Given a list of genes ranked by \( \omega \), we want to divide this list in 3 partitions (A, B and C). When values are scaled from 999 to 0, “Direct partitioning” would group genes within ranges of 300 resulting in few genes with \( \omega \) between 300 and 600, while “Partitioning by rank” would lead to a more continuous distribution.

Randomization of functional categories

To test possible biases attributed to the size of the functional category or the magnitude of change in evolutionary rate, we randomized the values (\( \omega \), \( \Delta \omega \), \( dN \) and \( dS \)) assigned to the list of genes. Functional categories would point to the same set of genes, conserving thus all structural characteristic of the data, but suppressing the biological relevance of evolutionary rates (see Figure 2.6).

This methodology allows us to test the effect of functional categories sizes, and to ensure that the distribution of evolutionary rates is not affecting our experiment. Each evolutionary variable was randomized 10,000 times for each species. Proportion of false positives (GSSA significant results) were plotted along the size of functional categories (from 0 to 1,500 with intervals of 20). As these proportions never reached values higher than 0.05% FDR, we rejected the possibility that either, group sizes or rate distributions, biased GSSA results in our data set (see Figure 2.7).

Simulating evolutionary scenarios

In order to understand better the results of the GSSA, a last experiment was necessary. Indeed, at this point, we were unable to discard that results with
2.3. Detection of selective pressure at molecular level

Figure 2.6.: Randomization experiment diagram.
The diagram shows the steps followed to test possible biases attributed to the size of functional categories. Genes are randomly rearranged according to their evolutionary statistics. Finally, genes are ranked according to their newly randomly assigned values. Results of the enrichment analysis over this dataset are considered false positives.

significant high $\omega$ were brought only by positively selected or relaxed genes.

Thus, to validate the independence of the GSSA from the effects of alternative evolutionary constraints, we simulated different selective regimes (purifying selection, positive selection, and relaxation) using branch-site models. Here, we addressed the possibility of variation in the representation of significant results after GSSA. The pipeline described in Figure 2.8 shows three different areas:
Chapter 2. Material and Methods

Figure 2.7.: Randomization experiment results.
These graphics show the proportion of false positives within the results of an enrichment analysis conducted over list of genes ranked by a shuffled evolutionary variable (see main text for details). Results are segregated in i) ranges of number of genes belonging to a functional category, in order to discard the effect of size on the proportion of false positives, and ii) evolutionary variables (red, green blue and yellow) and species to discard biases due to the specific distribution of one of the variable in a given species. Randomization was conducted in mammals (A) and flies (B).
2.3. Detection of selective pressure at molecular level

Figure 2.8: Evolutionary and statistical simulation of GSSA.
The pipeline shows the steps taken along three different spaces of analysis, the real data, the simulated data and the testing block. See main text for a complete explanation.

- **Real Data:** the light yellow area (A) describes the steps of the GSSA. The orange area (B) describes the use of the CodeML program from PAML package [Yang 2007] to extract –from original set of sequences– all evolutionary parameters needed to simulate new sequences under purifying selection (PF), positive selection (PS) or relaxation of the selective constraints (RX) according to branch-site models. Human, mouse, *D. melanogaster* and *D. erecta* were used as foreground species in the corresponding models.

- **Simulated Data:** in the light blue area (C), Evolver (also from PAML package [Yang 2007]) simulates sequences evolving under the given parameters (codon frequencies and branch lengths) estimated from the empirical data. We checked the desired characteristics of PS and RX on the set of the simulated sequences Table 2.6. Evolutionary variables ($d_S$, $d_N$, $\omega$ and $\Delta \omega$) were estimated from simulated sequences using a free-ratio branch model. The complete pipeline of the GSSA was applied over the simulated data.

- **Testing simulation:** the last part of the diagram represents the calculation of the odd-ratios corresponding to a classification of the GSSA results over all dataset. On the contingency tables are counted significant categories with either SH or SL $\omega$ and belonging to 2 of the 3 simulated selective regimes (PS, RX and PF). Odd-ratios values represents the association between different selective regime simulated according to their proportions of
Chapter 2. Material and Methods

SH and SL functional categories. Statistical contribution of the simulated regimes (PS, RX and PF) to the GSSA results were tested by comparing log odd-ratios with a t-test (results in Table 2.7).

<table>
<thead>
<tr>
<th></th>
<th>PS</th>
<th>RX</th>
<th>PF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H. sapiens</strong></td>
<td>658</td>
<td>1640</td>
<td>11</td>
</tr>
<tr>
<td><strong>M. musculus</strong></td>
<td>1500</td>
<td>954</td>
<td>14</td>
</tr>
<tr>
<td><strong>D. melanogaster</strong></td>
<td>736</td>
<td>630</td>
<td>25</td>
</tr>
<tr>
<td><strong>D. erecta</strong></td>
<td>778</td>
<td>1292</td>
<td>26</td>
</tr>
</tbody>
</table>

Table 2.6.: Number of positively selected and relaxed genes in each of the simulated evolutionary scenarios.

Our results showed that in spite of the alternative evolutionary scenarios, no significant differences were observed between log odd-ratios distributions (p<0.05). The average effect of PF, and RX/PS is the proportional decrease and increase of the mean value of $\omega$ on sequences, respectively. This change has minor effects (if any) in the relative position of genes in the ranked list of genes of a genome. Accordingly, since no net differences were produced after ranking genes, no significant differences are expected after the t-test (PS-RX: $p=0.99$, PS-PF: $p=0.45$, and RX-PF: $p=0.46$). The fact that basically the same number of significant results was observed in each evolutionary scenario confirmed this prediction. We conclude that neither of the selective regimes simulated produce significant differences or biases in the GSSA of $\omega$ values.

<table>
<thead>
<tr>
<th></th>
<th>PS</th>
<th>RX</th>
<th>PF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PS</strong></td>
<td>$-\text{—}$</td>
<td>92.50%</td>
<td>98.50%</td>
</tr>
<tr>
<td><strong>RX</strong></td>
<td>91.10%</td>
<td>$-\text{—}$</td>
<td>99.00%</td>
</tr>
<tr>
<td><strong>PF</strong></td>
<td>88.90%</td>
<td>90.60%</td>
<td>$-\text{—}$</td>
</tr>
</tbody>
</table>

Table 2.7.: Proportion of significant functional categories that are still significant.

Or keeping an identical signs of odd-ratios under a different evolutionary scenario.
Part I.

Structure and dynamics of genomes
3. Random-like structure of DNA

3.1. Computing genome complexity

Among 20 major systematic groups we pick 54 species and computed the complexity value (CV) of their genomes with sizes ranging from 3.4Gb to 1.6Kb Table 3.1. The first striking observation was the degree of direct correlation observed between genome size and CV (Figure 3.1-A) with a slope of the regression equal to 0.967. This first result implies a maximum complexity for all genomes. The residual variation around the fitted regression and along the 6 orders of magnitude, was almost null \( \text{adjusted } R^2 = 0.987 \).

The slope and degree of adjustment that shows our set of species are quite surprising given the diversity of living forms used in this analysis. Our array of organism from the shortest single-strand RNA genome of *Hepatitis D* virus (size $\sim 1.69e^{+03}$ bp) to the largest double-strand DNA genome of the short-tailed opossum *Monodelphis domestica* (size $\sim 3.41e^{+09}$ bp), this, even including organisms among those with most peculiar genomes as can be:

- obligate endosymbionts bacteria with extreme reduction of genome size (*Carsonella ruddii, Buchnera aphidicola*, and *Ureaplasma urealyticum*) [Wernegreen 2002]
- parthenogenetic crustaceans with ubiquitous duplications of genes (*Daphnia pulex*)
- archean organisms living in extreme environmental conditions (*Sulfolobus islandicus, Methanocaldococcus vulcanius, Thermococcus sibiricus*)
- the first synthetic organism (*Synthetic mycoplasma mycoides*) [Gibson et al. 2010]

All of them fit the slope of the linear regression model.

In order to better contrast deviations from maximum complexity we computed the complexity ratio (CR), and the deviation to the maximum ratio (Dmax = 1 - CR) for each species. According to Table 3.1, only ten species showed Dmax > 0.05. These are:
<table>
<thead>
<tr>
<th>Feat.</th>
<th>Species</th>
<th>ACN-EV</th>
<th>GS</th>
<th>GC</th>
<th>GCR</th>
<th>Dmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>Hepatitis BV</td>
<td>NC3977.1</td>
<td>1,682</td>
<td>1,671</td>
<td>1</td>
<td>0.0016</td>
</tr>
<tr>
<td>SGS</td>
<td>Hepatitis DV</td>
<td>D01075.1</td>
<td>3,215</td>
<td>3,210</td>
<td>0.9984</td>
<td>0.0036</td>
</tr>
<tr>
<td>RNA</td>
<td>Tomato mosaicV</td>
<td>NC010836</td>
<td>5,058</td>
<td>5,040</td>
<td>0.9964</td>
<td>0.0016</td>
</tr>
<tr>
<td>SSD</td>
<td>Enterobacteria phage m13P</td>
<td>V00604</td>
<td>6,407</td>
<td>6,367</td>
<td>0.9938</td>
<td>0.0062</td>
</tr>
<tr>
<td>RNA</td>
<td>HIV 1V</td>
<td>NC001802</td>
<td>9,181</td>
<td>9,105</td>
<td>0.9917</td>
<td>0.0083</td>
</tr>
<tr>
<td>RNA</td>
<td>Sudan ebolavirusV</td>
<td>NC006432</td>
<td>18,875</td>
<td>18,842</td>
<td>0.9993</td>
<td>0.0007</td>
</tr>
<tr>
<td>DSD</td>
<td>Human herpesvirus1</td>
<td>NC008512</td>
<td>150,036</td>
<td>146,930</td>
<td>0.9854</td>
<td>0.0146</td>
</tr>
<tr>
<td>SL</td>
<td>Synthetic mycoplasma mycoidesBa</td>
<td>CP002027.1</td>
<td>1,078,809</td>
<td>1,026,444</td>
<td>0.9515</td>
<td>0.0485</td>
</tr>
<tr>
<td>EE</td>
<td>Thermococcus sibiricusAr</td>
<td>CP001463.1</td>
<td>1,242,891</td>
<td>1,237,320</td>
<td>0.9955</td>
<td>0.0045</td>
</tr>
<tr>
<td>EE</td>
<td>Methanocaldococcus vulcaniusAr</td>
<td>CP001787.1</td>
<td>1,746,040</td>
<td>1,708,968</td>
<td>0.9788</td>
<td>0.0212</td>
</tr>
<tr>
<td>EE</td>
<td>Sulfolobus islandicusAr</td>
<td>CP001731.1</td>
<td>2,722,004</td>
<td>2,692,455</td>
<td>0.9891</td>
<td>0.0109</td>
</tr>
<tr>
<td>EE</td>
<td>Bacillus subtilisBa</td>
<td>E! Bacte. 9</td>
<td>4,215,606</td>
<td>4,198,057</td>
<td>0.9958</td>
<td>0.0042</td>
</tr>
<tr>
<td>EE</td>
<td>Mycobacterium tuberculosisBa</td>
<td>E! Bacte. 9</td>
<td>4,411,532</td>
<td>4,348,606</td>
<td>0.9857</td>
<td>0.0143</td>
</tr>
<tr>
<td>EE</td>
<td>Escherichia coliBa</td>
<td>CP001396.1</td>
<td>4,578,159</td>
<td>4,551,258</td>
<td>0.9941</td>
<td>0.0059</td>
</tr>
<tr>
<td>LBG</td>
<td>Burkholderia xenovoransBa</td>
<td>NC007951-3</td>
<td>9,731,138</td>
<td>9,593,486</td>
<td>0.9859</td>
<td>0.0141</td>
</tr>
<tr>
<td>AP</td>
<td>Saccharomyces cerevisiaeFa</td>
<td>E! Fungi 3</td>
<td>12,070,898</td>
<td>11,974,342</td>
<td>0.992</td>
<td>0.008</td>
</tr>
<tr>
<td>UE</td>
<td>Plasmodium falciparumAp</td>
<td>E! Proti. 9</td>
<td>23,263,322</td>
<td>21,070,640</td>
<td>0.9057</td>
<td>0.0943</td>
</tr>
<tr>
<td>UE</td>
<td>Phaegactylym tricornutumHe</td>
<td>E! Proti. 9</td>
<td>25,805,651</td>
<td>25,676,448</td>
<td>0.9946</td>
<td>0.0054</td>
</tr>
<tr>
<td>UE</td>
<td>Thalassiosira pseudonanaHe</td>
<td>E! Proti. 9</td>
<td>31,199,234</td>
<td>31,023,020</td>
<td>0.9944</td>
<td>0.0056</td>
</tr>
<tr>
<td>UE</td>
<td>Dicyostelium discoideumAm</td>
<td>E! Proti. 9</td>
<td>33,919,934</td>
<td>30,877,496</td>
<td>0.9103</td>
<td>0.0897</td>
</tr>
<tr>
<td>UE</td>
<td>Ciona intestinalisUr</td>
<td>E! 62</td>
<td>87,649,861</td>
<td>84,674,396</td>
<td>0.9661</td>
<td>0.0339</td>
</tr>
<tr>
<td>UE</td>
<td>Caenorhabditis elegansIn</td>
<td>E! Meta. 9</td>
<td>100,272,217</td>
<td>97,720,472</td>
<td>0.9746</td>
<td>0.0254</td>
</tr>
<tr>
<td>AP RG</td>
<td>Arabidopsis thalianaPl</td>
<td>E! Plants 9</td>
<td>118,960,082</td>
<td>116,563,556</td>
<td>0.9799</td>
<td>0.0201</td>
</tr>
</tbody>
</table>
## 3.1. Computing genome complexity

<table>
<thead>
<tr>
<th>Species †clade*</th>
<th>ACN-EV</th>
<th>GS</th>
<th>GC</th>
<th>GCR</th>
<th>Dmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drosophila melanogaster†In</td>
<td>E! Metaz. 9</td>
<td>120,290,887</td>
<td>118,973,632</td>
<td>0.989</td>
<td>0.011</td>
</tr>
<tr>
<td>GE</td>
<td>Daphnia pulex†In</td>
<td>E! Metaz. 9</td>
<td>158,632,523</td>
<td>150,111,316</td>
<td>0.9463</td>
</tr>
<tr>
<td>AP</td>
<td>Arabidopsis lyrata†Pl</td>
<td>E! Plants 9</td>
<td>173,245,910</td>
<td>161,798,504</td>
<td>0.9339</td>
</tr>
<tr>
<td>AP</td>
<td>Tetraodon nigroviridis‡Pl</td>
<td>E! 62</td>
<td>208,708,313</td>
<td>207,067,712</td>
<td>0.9921</td>
</tr>
<tr>
<td>AP</td>
<td>Apis mellifera†In</td>
<td>E! Metaz. 9</td>
<td>224,750,524</td>
<td>219,278,732</td>
<td>0.9757</td>
</tr>
<tr>
<td>Anopheles gambiae†In</td>
<td>E! Metaz. 9</td>
<td>225,028,531</td>
<td>221,180,624</td>
<td>0.9829</td>
<td>0.0171</td>
</tr>
<tr>
<td>AP</td>
<td>Brachypodium distachyon†Pl</td>
<td>E! Plants 9</td>
<td>270,058,956</td>
<td>257,893,524</td>
<td>0.955</td>
</tr>
<tr>
<td>AP</td>
<td>Oryza sativa†Pl</td>
<td>E! Plants 9</td>
<td>293,104,375</td>
<td>271,137,108</td>
<td>0.9251</td>
</tr>
<tr>
<td>AP</td>
<td>Populus trichocarpa†Pl</td>
<td>E! Plants 9</td>
<td>370,421,283</td>
<td>352,063,876</td>
<td>0.9504</td>
</tr>
<tr>
<td>AP</td>
<td>Phlycomitrella patens†Fr</td>
<td>E! Plants 9</td>
<td>453,927,385</td>
<td>399,508,556</td>
<td>0.8801</td>
</tr>
<tr>
<td>AP</td>
<td>Sorghum bicolor†Pl</td>
<td>E! Plants 9</td>
<td>625,636,188</td>
<td>491,993,216</td>
<td>0.7864</td>
</tr>
<tr>
<td>AP</td>
<td>Oryzias latipes‡Pl</td>
<td>E! 62</td>
<td>582,126,393</td>
<td>562,662,192</td>
<td>0.9666</td>
</tr>
<tr>
<td>Gallus gallus †Ri</td>
<td>E! 62</td>
<td>984,855,151</td>
<td>971,359,304</td>
<td>0.9863</td>
<td>0.0137</td>
</tr>
<tr>
<td>Taeniopygia guttata†Ri</td>
<td>E! 62</td>
<td>1,013,982,659</td>
<td>996,918,996</td>
<td>0.9832</td>
<td>0.0168</td>
</tr>
<tr>
<td>AP</td>
<td>Danio rerio‡Pl</td>
<td>E! 62</td>
<td>1,354,636,069</td>
<td>1,191,452,752</td>
<td>0.8795</td>
</tr>
<tr>
<td>AP RP</td>
<td>Zea mays†Pl</td>
<td>E! Plants 9</td>
<td>2,045,697,632</td>
<td>1,197,255,904</td>
<td>0.5853</td>
</tr>
<tr>
<td>Canis familiaris †Ma</td>
<td>E! 62</td>
<td>2,309,875,279</td>
<td>2,272,374,188</td>
<td>0.9838</td>
<td>0.0162</td>
</tr>
<tr>
<td>Equus caballus †Ma</td>
<td>E! 62</td>
<td>2,335,454,424</td>
<td>2,307,202,104</td>
<td>0.9879</td>
<td>0.0121</td>
</tr>
<tr>
<td>Bos taurus †Ma</td>
<td>E! 62</td>
<td>2,466,956,401</td>
<td>2,406,743,280</td>
<td>0.9756</td>
<td>0.0244</td>
</tr>
<tr>
<td>Rattus norvegicus †Ma</td>
<td>E! 62</td>
<td>2,477,053,718</td>
<td>2,430,894,052</td>
<td>0.9814</td>
<td>0.0186</td>
</tr>
<tr>
<td>Mus musculus †Ma</td>
<td>E! 62</td>
<td>2,558,509,481</td>
<td>2,521,038,616</td>
<td>0.9854</td>
<td>0.0146</td>
</tr>
<tr>
<td>Pan trogodytes †Ma</td>
<td>E! 62</td>
<td>2,598,733,311</td>
<td>2,566,544,820</td>
<td>0.9876</td>
<td>0.0124</td>
</tr>
<tr>
<td>Macaca mulatta †Ma</td>
<td>E! 62</td>
<td>2,646,263,164</td>
<td>2,621,196,144</td>
<td>0.9905</td>
<td>0.0095</td>
</tr>
<tr>
<td>Pongo abelii †Ma</td>
<td>E! 62</td>
<td>2,722,968,487</td>
<td>2,697,592,876</td>
<td>0.9907</td>
<td>0.0093</td>
</tr>
<tr>
<td>Homo sapiens  †Ma</td>
<td>E! 62</td>
<td>2,858,658,095</td>
<td>2,841,049,052</td>
<td>0.9938</td>
<td>0.0062</td>
</tr>
<tr>
<td>LGS</td>
<td>Monodelphis domestica†Ma</td>
<td>E! 62</td>
<td>3,412,593,369</td>
<td>3,402,944,248</td>
<td>0.9972</td>
</tr>
</tbody>
</table>

### Table 3.1.: Genomes Complexity.

Genomes size (GS), genomes complexity (GC), genome complexity ratio ($GCR = \frac{GC}{GS}$), and deviation from the maximum GCR (Dmax=1-GCV) for 54 species of different taxa. NCBI accession number or Ensembl (E!) version (ACN-EV). Features: AP: Ancient Polyplid; DSD: Double-Strand DNA; EE: Extreme Environment; GE: Gene Expansion; IP: Intracellular Parasite; LBG: Largest Bacterial Genome; LGS: Largest Genome Sequenced; RG: Reduced Genome; RNA: RNA Virus; RP: Recent Polyplid; SBG: Shortest Bacterial Genome; SGS: Shortest Genome Sequenced; SL: Synthetic Life; SSD: Single-Strand DNA; UE: Unicellular Eukaryote. Notes: -1:- [http://www.hgsc.bcm.tmc.edu/ftp-archive/TcΛstanium/Trena3.0/](http://www.hgsc.bcm.tmc.edu/ftp-archive/TcΛstanium/Trena3.0/), (*) Clades are abbreviated as: Vi: Virus; Ph: Phage; Ba: Bacteria; Ar: Archaea; Fu: Fungi; Ap: Apicomplexa; Am: Amoebozoa; He: Heterokonta; Ur: Urochordate; In: Invertebrates; Pl: Plants; Fii: Fishes; Br: Bryophyta; Ma: Mammals;
Chapter 3. Random-like structure of DNA

- six ancient or recent polyploid species; the most extreme case of genome reduction in bacteria
- the explosive case of gene expansion in Daphnia [Colbourne et al. 2011]
- two unicellular eukaryotes that curiously correspond to the 2 genomes sequenced with higher proportion of A + T, *Plasmodium* [Gardner et al. 2002] (A + T content around 81%) and *Dictyostelium* [Eichinger & Noegel 2003] (A + T content around 78%)

The highest CR=1 was obtained for randomly generated sequences with uniform distribution of A C G T. To simulate events of polyploidization random sequences were duplicated up to 5 times (corresponding to a 10×). The fall in CR reaching CR=0.25 for 10× could thus be compared to CR values of real polyploids placing maize genomes at the level of a perfect triploid and sorghum to the level of a perfect diploid.

Besides genomes and random sequences, results of the computation of CR of human texts were added to Figure 3.1-B (values used can be found in Table 3.2).

All together, complexity ratios of complete genomes, random sequences of different ploidy and human language texts were computed. Maximum CR corresponds to random sequence of lengths ranging from 5 Kb to 2.5 Gb. In the case of biological sequences, non-polyploid genomes showed CR > 0.90. Conversely, polyploids showed CR below 0.95, with the lowest ratio for *Z. mays* (CR=0.58), and the next to the lowest ratio, its closest relative *S. bicolor* (CR=0.78). Overall non-random strings analyzed, the lowest CR was obtained in human language texts. CR of 11 human texts of different sizes and languages, from short scientific abstract to the complete works of William Shakespeare, are also depicted Figure 3.1-B. CR diminishes as texts size increases, due to the limited lexicon and the fixed language grammar. Complexity reached the lowest ratio in Darwin’s *Origin of Species* (≈ 0.309), which is comparable to the CR of a random polyploid sequence

Figure 3.1. (following page): Genome complexity value.

(A) Complexity values and genome size of 54 genomes. Log scales are used to display species diversity. Some relevant species are labeled (the complete list Table 3.1). (B) Most genomes have complexity ratio (CR) between 0.90 and 1.0. Four polyploid species have CR < 0.9 (in bold in the figure): *P. patens* (0.880), *D. rerio* (0.879), *S. bicolor* (0.786) and *Z. mays* (0.585). The stars with CR = 1 correspond to random [ACGT] strings of 30, 50, 100, 250 and 500 Mb length, respectively. Others stars with lower CR correspond to the 500 Mb random string repeated from 2× to 6× simulating perfect polyploids. Changes in sequence length due to polyploidy produce no change in complexity ratio. Notice the low CR of human texts (see Table 3.2). Confidence band in both plots, corresponds to the 99% certainty to contain the best regression line.
3.1. Computing genome complexity
Chapter 3. Random-like structure of DNA

<table>
<thead>
<tr>
<th>Kind</th>
<th>Author - Writings</th>
<th>Lang.</th>
<th>L</th>
<th>C</th>
<th>CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>C Venter The human genome (abstract)</td>
<td>English</td>
<td>2,662</td>
<td>1,613</td>
<td>0.6059</td>
</tr>
<tr>
<td>SS</td>
<td>J L Borges El Aleph</td>
<td>Spanish</td>
<td>28,507</td>
<td>14,991</td>
<td>0.5259</td>
</tr>
<tr>
<td>B</td>
<td>A Von Goethe Torcuato Tasso</td>
<td>German</td>
<td>152,104</td>
<td>68,187</td>
<td>0.4483</td>
</tr>
<tr>
<td>B</td>
<td>H Quiroga Cuentos de amor, locura y muerte</td>
<td>Spanish</td>
<td>293,482</td>
<td>125,552</td>
<td>0.4278</td>
</tr>
<tr>
<td>B</td>
<td>D F Sarmiento Facundo</td>
<td>Spanish</td>
<td>601,477</td>
<td>242,982</td>
<td>0.4259</td>
</tr>
<tr>
<td>B</td>
<td>D Alighieri Divina Commedia</td>
<td>Italian</td>
<td>570,480</td>
<td>301,609</td>
<td>0.3692</td>
</tr>
<tr>
<td>B</td>
<td>I Newton Principia Mathematica</td>
<td>Latin</td>
<td>817,032</td>
<td>237,558</td>
<td>0.395</td>
</tr>
<tr>
<td>B</td>
<td>B C Darwin The Origin of species</td>
<td>English</td>
<td>981,958</td>
<td>303,503</td>
<td>0.3091</td>
</tr>
<tr>
<td>B</td>
<td>B M Cervantes El Quijote</td>
<td>Spanish</td>
<td>2,097,943</td>
<td>790,702</td>
<td>0.3769</td>
</tr>
<tr>
<td>B</td>
<td>B V Hugo Les Miserables</td>
<td>French</td>
<td>3,259,269</td>
<td>1,141,378</td>
<td>0.3502</td>
</tr>
<tr>
<td>CW</td>
<td>W Shakespeare</td>
<td>English</td>
<td>5,447,165</td>
<td>2,111,425</td>
<td>0.3876</td>
</tr>
</tbody>
</table>

Table 3.2.: Human language Complexity
Work length (L), complexity (C), complexity ratio (CR), and deviations from the maximum ratio of complexity (Dmax=1- CR) for 11 human writings in six different languages. Kinds: SA: Scientific abstract, SS: Short story; B: Book, CW: Complete Work

The human genome

Observe that text sizes are contained in the range of phages, virus and bacteria genome sizes. Complexities of human writings detailed in Table 3.2.

3.1.1. Genome complexity and ploidy level

Analysis of CR Figure 1.2-B reveals a clear segregation of species’ genomes by their level of ploidy, most recent polyploids like maize and sorghum exhibit lowest CR. However this trend seems to be able to be lost quickly as ancient polyploids are hardly distinguishable from non-polyploids. A very illustrative example can be found within Arabidopsis genus where the 2 close relatives thaliana and lyrata (they diverged 10 My ago [Hu et al. 2011]) seem to have followed different routes after their whole genome duplication (< 70 My ago [Proost et al. 2011]). While Arabidopsis thaliana suffered a drastic genome reduction after polyploidization (mainly due to hundreds of thousands of small deletions), its relative, Arabidopsis lyrata, remained complete [Hu et al. 2011]. And this is reflected in the differences in CR between those two species CR=0.9339 for lyrata and CR=0.9799 for thaliana that undergoes a faster increment in CR.

In order to confirm the trend observed around level of polyploidy and CR value, we tested the hypothesis that the observed genome complexity values correlate with size and ploidy level. A categorical variable divided polyploid (ancient or recent), and non-polyploid species described in Table 3.1. The size-interaction term provided significant deviations (p < 2e−16, adjusted-R2 = 0.997), while independent linear models slopes were 0.633 (p < 4.8e−07, adjusted-R2 = 0.921), and 0.988 (p < 2e−16, adjusted-R2 = 1.00) for polyploid and non-polyploid genomes.
3.2. Chromosome complexity

Following the same methodology used for genomes, we computed CR of individual chromosomes (567 autosomes of 31 species). CV obtained were normalized by chromosome size resulting in a CR and can be seen in Figure 3.2. As previously, statistics were very convincing. The slope of the relationship between chromosome size and CV was around 0.924, and could be increased excluding polyploid species to 0.951 (alone, polyploid species exhibit a low slope = 0.696). And again, the size-interaction term was indisputable (p < 2e-16).

Figure 3.2.: Chromosome complexity ratio.
Most chromosomes (96.2%) have complexity ratios ranging 0.9 to 1.0, as observed for complete genomes Figure 3.1B. Boxplot on the left shows the distribution of CR of all chromosomes (outliers are shown in red, and the yellow star corresponds to the mean value). Notice how far are standing chromosomes of Z. mays, and to a lesser degree S. bicolor ones (both recent polyploid species) from the general trend.

Notice that when considering non-polyploid species, the slope of the correlation, CV versus size, is almost common for chromosomes and genomes (slope = 0.989 and 0.988, respectively).

The boxplot inside Figure 3.2 summarizes the distribution of chromosomes’ CR. The first quartile of the full sample indicates that 75% of the data are above 0.958, while the median and mean are 0.974 and 0.964. The minimum CR value corresponds to maize chromosome 10 (0.683), and maximum to P. tricornutum chromosome 28 (0.999). Opossum chromosome 1 (the largest chromosome) has a CR of 0.942. Mean CR of maize’s chromosomes was 0.698, while maize genome CR was 0.585. The difference suggests extensive duplicated regions in maize chromosomes, which was previously described in [Weber & Helentjaris 1989, Gaut...
2001] and attributed to a tetraploid event occurred in the origin of maize 11 My ago [Gaut & Doebley 1997, Wolfe 2001].

What brought out when comparing CR of genomes and of their corresponding chromosome, is that even if the global picture was conserved, some differences raised. Namely, here are some of those differences (genomes’ CR – mean chromosomes’ CR) $S$. bicolor (0.854 – 0.786), $D$. rerio (0.924 – 0.879), $A$. lyrata (0.966 – 0.934), $P$. trichocarpa (0.971 – 0.950), $S$. cerevisiae (0.996 – 0.992), $A$. thaliana (0.986 – 0.980), $M$. domestica (0.944 – 0.997), $M$. musculus (0.959 – 0.985) and $H$. sapiens (0.960 – 0.993)

As it appears, those differences are either positive or negative. The falls in genomes’ CR, occurs generally in polyploid species and can be explained by a broader definition of “repetitive elements” as a consequence of the consideration of a wider window (see next subsection 3.4.1). In contrast, raises in genome’s CR is a more sensitive problem further discussed in the subsection 3.4.2 on polyploid and return to maximum complexity.

3.3. Complexity in repetitive elements and genes – low and high?

Eukaryote genome structure is generally sketched out by the massive presence of non-functional repetitive elements (REs) spread out all over the genome, and a tiny portion of singular functional elements covering the rest. To get insights into the statistical structure of these contrasting regions of genomes we computed the complexity ratio of genes and of each of the main families of RE’s (as DNA-T, LTR, LINE, SINE and satellite). This could be achieved, for each family, by concatenating all units in their original order in chromosomes.

Genes, characterized by a specially high content of information, showed, as expected, the highest CR among all classes analyzed independently of the species. Indeed, the typical structure of genic regions is important for entropy-based algorithms that predicts or confirm automatic detection of genes [Du et al. 2006, Gerstein et al. 2007], however it is important here to have echoes of our premiere methodology taking all genes together instead of using the typical sliding window. Going deeper in the analysis, when genes were split in their two main components, exons showed even a higher CR.

For repetitive elements the expectation was that CR would fall due to the low complex nature of the elements. And indeed this result was found in SINE and satellites. However for LINE, LTR and DNA-T (Table 3.3) unexpectedly high values were observed. This result can be explained by the larger length of these elements and their high internal variability among families.

We also noticed some interesting clade specificity regarding the relative CR of RE families. For example, in mammals DNA-T and LTR elements exhibited
3.3. Complexity in repetitive elements and genes – low and high?

Table 3.3.: Mean complexity ratio of some genetic components.

<table>
<thead>
<tr>
<th>Species</th>
<th>Satellite</th>
<th>SINE</th>
<th>LINE</th>
<th>ETR</th>
<th>DNA-T</th>
<th>Genes</th>
<th>Introns</th>
<th>Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>H. sapiens</strong></td>
<td>0.485</td>
<td>0.437</td>
<td>0.881</td>
<td>0.922</td>
<td>0.962</td>
<td>0.953</td>
<td>0.952</td>
<td>0.985</td>
</tr>
<tr>
<td><strong>P. troglodytes</strong></td>
<td>0.491</td>
<td>0.442</td>
<td>0.885</td>
<td>0.926</td>
<td>0.962</td>
<td>0.967</td>
<td>0.965</td>
<td>0.993</td>
</tr>
<tr>
<td><strong>R. norvegicus</strong></td>
<td>0.539</td>
<td>0.586</td>
<td>0.668</td>
<td>0.912</td>
<td>0.975</td>
<td>0.977</td>
<td>0.976</td>
<td>0.992</td>
</tr>
<tr>
<td><strong>M. musculus</strong></td>
<td>0.595</td>
<td>0.576</td>
<td>0.74</td>
<td>0.875</td>
<td>0.973</td>
<td>0.973</td>
<td>0.97</td>
<td>0.991</td>
</tr>
<tr>
<td><strong>C. familiaris</strong></td>
<td>0.6</td>
<td>0.487</td>
<td>0.911</td>
<td>0.974</td>
<td>0.982</td>
<td>0.982</td>
<td>0.98</td>
<td>0.993</td>
</tr>
<tr>
<td><strong>T. nigroviridis</strong></td>
<td>—</td>
<td>0.585</td>
<td>0.903</td>
<td>—</td>
<td>—</td>
<td>0.994</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td><strong>D. rerio</strong></td>
<td>0.628</td>
<td>0.43</td>
<td>0.796</td>
<td>0.791</td>
<td>0.824</td>
<td>0.942</td>
<td>0.936</td>
<td>0.988</td>
</tr>
<tr>
<td><strong>C. intestinalis</strong></td>
<td>0.644</td>
<td>0.537</td>
<td>0.836</td>
<td>0.937</td>
<td>0.801</td>
<td>0.968</td>
<td>0.957</td>
<td>0.994</td>
</tr>
<tr>
<td><strong>C. elegans</strong></td>
<td>0.52</td>
<td>0.401</td>
<td>0.93</td>
<td>0.94</td>
<td>0.827</td>
<td>0.978</td>
<td>0.957</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>A. gambiae</strong></td>
<td>0.232</td>
<td>0.438</td>
<td>0.805</td>
<td>0.902</td>
<td>0.771</td>
<td>0.992</td>
<td>0.992</td>
<td>0.9</td>
</tr>
<tr>
<td><strong>D. melanogaster</strong></td>
<td>0.548</td>
<td>—</td>
<td>0.81</td>
<td>0.744</td>
<td>0.81</td>
<td>0.985</td>
<td>0.982</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Z. mays</strong></td>
<td>0.337</td>
<td>0.531</td>
<td>0.906</td>
<td>0.495</td>
<td>0.7223</td>
<td>0.962</td>
<td>0.956</td>
<td>0.975</td>
</tr>
<tr>
<td><strong>S. bicolor</strong></td>
<td>0.345</td>
<td>0.619</td>
<td>0.966</td>
<td>0.602</td>
<td>0.757</td>
<td>0.99</td>
<td>0.991</td>
<td>0.988</td>
</tr>
<tr>
<td><strong>A. thaliana</strong></td>
<td>0.467</td>
<td>0.675</td>
<td>0.971</td>
<td>0.84</td>
<td>0.896</td>
<td>0.989</td>
<td>0.986</td>
<td>0.988</td>
</tr>
<tr>
<td><strong>A. lyrata</strong></td>
<td>0.417</td>
<td>0.457</td>
<td>0.928</td>
<td>0.772</td>
<td>0.826</td>
<td>0.994</td>
<td>0.988</td>
<td>0.996</td>
</tr>
</tbody>
</table>

Table 3.4.: Complexity ratio of genome classes concatenated and shuffled

Size and complexity ratio of different genome classes in natural order (NAT), and shuffled (SHU) for selected chromosomes. SIZE represents the length in Mb of concatenated elements.
higher CR than LINE elements, while this is not the case for fishes, some invertebrates and plants. Moreover, in plants, LINE has the highest CR after genes (see Table 3.3 for comparison among all eukaryote species analyzed).

In order to test the hypothesis that the order in which RE are placed in chromosome influences our measures, we compared the CR values of the RE in "natural" versus random order. Table 3.4 shows these values for eight selected chromosomes of different species. Curiously, CR of elements in natural order was much lower in SINEs and satellites than in the rest of the classes. Revealing a structure of identical or very similar repeats along neighbor chromosome segments. This pattern did not showed up in the other families. The notable exception was LTRs of the maize chromosome, known to have expanded dramatically in recent evolutionary times [Blanc & Wolfe 2004]. All shuffled classes (including SINEs and satellites) had a CR very close to one. This result point outs that genomes are plenty of genetic variation, even in regions where the expected pattern is the homogeneous repetition of almost indistinguishable units of RE’s.

3.4. Methodological validation and interpretation – Understanding the CR

3.4.1. Complexity in chromosome segments

In order to fully understand how CR ratios works, and how good is the idea to work with full genomes or chromosomes, we decided to apply our algorithm to windows of different sizes. Chromosomes were thus, split in overlapping windows of various sizes (from 1 Kb to 100 Mb) and CR in each of these windows was computed. Figure 3.3 shows boxplots of six selected chromosomes, at different scales, all having extreme CR.

As a first trend we observed that median values of CR over all windows. A. thaliana Chr1, C. elegans ChrI, D. melanogaster Chr2L, H. sapiens Chr1 were above 0.97 (Figure 3.3). Whereas lower values were obtained in Z. mays Chr1 and in H sapiens Chr19. The fall in CR in these last chromosome is more dramatic for large windows sizes (1Mb). The reasons for this fall are different in the two cases: while maize Chr1 is tetraploid, human Chr19 is known to contain the highest number of Alu sequences over human chromosomes [Venter et al. 2001].

Among all chromosomes presented here (Figure 3.3), but also in the rest of chromosomes analyzed (available through this link: http://bioinfo.cipf.es/das/), the observation that larger the window size, the lower the median CR value was prevalent. This pattern is explained by the existence of repeats, which can only be detected when the window size is large enough.

As a last example of how window size affects the detection of regions with low entropy stands in Figure 3.4. This figure represents the entropy-shape of D.
3.4. Methodological validation and interpretation – Understanding the CR

Figure 3.3.: Sliding window analysis in chromosomes.
Boxplots show results of sliding window analysis in 6 selected chromosomes. Most chromosomes have median CR higher than 0.975 independently of window size.

Figure 3.4.: Sliding window in a full chromosome.
All three plots corresponds to a sliding window analysis along *D. melanogaster* chromosome 2L. The first on the top is a count of genes in a window of 1e+05Kb (data retrieved from Ensembl [Flicek et al. 2011]). Next two plots represents the complexity ratio displayed at two window size scales. Ensembl annotation of the histone genes cluster with more than 100 histone genes is highlighted in gray in three plots.
melanogaster chromosome 2L for two sizes of windows. In the case of small windows (1K) the rugged pattern drawn by the values of CR across the chromosome is hardly interpretable, we can only figure that each fall in CR correspond to regions with a high number of small repetitions of one or two nucleotides. By contrast, when windows size reaches 100K, the shape of CR is much smoother, revealing only one main peak of low CR. Interestingly corresponding to an histone cluster with more than 100 genes of the family locates. The plot on the top of the figure shows Ensembl gene annotation with the location of the histone genes cluster highlighted.

These examples are revealing how the measure of entropy is affected by the size of the sequence measured. The most striking example showing the importance of selecting large windows in order to include a maximum of information being the case of maize chromosome 1 where the mean CR values is subjected to a dramatic fall when the size of the windows measured are rising.

3.4.2. Polyploidy and return to maximum complexity

Evolution erodes ancient footprints of genome polyploidy and diploidization proceeds during time [Wolfe 2001]. As shown in previous sections, CR of recent polyploids is much lower than in non-polyploid, or in ancient polyploid species. "Erosion" of polyploids can be achieved by multiple mechanisms [Wolfe 2001]. The most simple, perhaps, being the gradual disintegration of the duplicated genetic material by random mutation. Other mechanisms, more dramatic, also participate in the loss of polyploid footprints, such as massive deletion and transpositions of genetic material as was reported in A. thaliana [Hu et al. 2011]. We tested the hypothesis that the complexity ratio of polyploid genomes increases along with their “maturation”.

In order to better understand the decay of genetic redundancy after polyploidization, the action of two mechanisms –mutation and transposition– were simulated over repeated random sequences of different lengths. The first process (mutation) was also applied to 2 real chromosomes of our most recent polyploids Z. mays Chr1 and S. bicolor Chr1.

In all cases, sequences under random mutation (Figure 3.5-A) or transposition (Figure 3.5-B) reached maximum CR=1 after a number of generations large enough. A general observation is that the lower is the CR the more sensible it is to changes (either mutation or transpositions). Exactly as expected by the probability theory since each single choice (introduced by a random mutation or a transposition) in a large set is more informative than in a smaller set, because it makes a selection in a bigger space of possibilities. For real polyploids (sorghum and maize), the dynamics of CR increase was identical (Figure 3.5-A). Figure 3.5-B shows that genomes and chromosomes reached maximum CR=1 after many
Figure 3.5: Return to maximum complexity after polyploidization.

(A) Random genomes of different lengths and ploidy levels experienced increase of CR by accumulation of random mutation ($1e^{-08}$ mutations per site, per generation). Chromosomes of maize and sorghum are included in the simulation. (B) Starting from one random sequence, we simulated 2 rounds of diploidization – simulating thus a tetraploid genomes. Experiment was repeated for different lengths. Translocations of 1Kb occurred at a constant rate over 10,000 generations (plotted each 100). “○”, “△”, “+” and “×” represent 4 simulated chromosomes while lines, their concatenation.
cycles of transpositions. Using a simulated genome with tetraploid structure, transposition preserved the relation that chromosome CR is higher than genome CR, along all generations up to convergence to maximum CR=1. This property was reported above for maize and sorghum (see discussion on section 3.2).

As CR get closer to 1 through time, DNA structure of polyploids become indistinguishable from diploid genomes.

### 3.4.3. Low CR corresponds to a simple combinatorial structure of the sequence.

The combinatorial structure of a sequence is a description of the observed arrangement of the symbols among all possible permutations of the same length. Sequences with many long repeats have low CR (see subsection 2.1.2, in Material and Methods, to get a complete picture on how the CR works). Polyploid genomes of maize and sorghum have CR=0.585, and CR=0.786, respectively. Values that were respectively close to diploid and triploid simulated genomes (Figure 3.1-B). It is also possible to achieve low CR in sequences without any long repeats, but with an orderly arrangement of the symbols. Although we have not found this phenomena in natural DNA, we constructed de Bruijn sequences [de Bruijn 1946, Becher & Heiber 2011] with low CR. See Table 2.1 in Material and Methods for examples on short sequences.

### 3.4.4. High CR corresponds to random-like sequences

High complexity ratio implies a high diversity and balanced abundance of short repeats in DNA sequences. Maximum CR=1 is reached by a sequence of length \(n\) if it contains full diversity of length \(k\), for \(k \leq \log_4 n\), and each of these short sequences occurs about \(n \cdot 4^{-k}\) times.

Intuitively, a non random sequence will exhibit some significant regularity that can be used to compress the sequence. The mathematical underpinning relies on the theory of pure randomness [Chaitin 1975, Nies 2009], which states that an infinite sequence is random when its initial segments are incompressible. Up to some deviations, for finite sequences and particular compression methods, statistical randomness is the exact inverse of compressibility. Thus, high complexity ratios correspond to highly incompressible sequences, which are sequences with a random-like structure. As in statistical randomness, the number of sequences with high CR grows exponentially with the sequence length. Thus, each genome is a singular instance out of the extraordinary number of combinatorial variants of the same length with the same high complexity rate.
3.5. Discussion

3.5.1. Universal structure of DNA

Up today, no conclusive work on statistical property of DNA was conducted in full genomes. After having a broad look to our result, the most remarkable, is certainly that whatever genome taken at whatever level of magnification, genome chromosome or windows, DNA seems to be strongly attracted to maximum complexity. At genomic and chromosomal level, recent polyploid species were the only outliers observed. Aside from these exceptions, CR was stacked to 1 in the whole range of diversity of life browsed, from viruses to mammals. Even the most expected “simple” genomes like higher eukaryotes, with around a half of their genomes composed by repetitive sequences presented a CR > 0.95. And when genetic elements were analyzed separately, only SINEs and satellites were indeed presenting lower CR, other families of GEs showed high degrees of variability.

3.5.2. Mechanisms of genome amplification and divergence

From our observations, genome size enlargement by duplications result in lowering CR. However maximum CR is rapidly recovered, ancient polyploids in our dataset were all presenting high values of CR. Also the simulations conducted showed that after 30 millions generations is enough to recover a CR > 0.95, and this taking only into account single mutations at a mean rate. In this context we envision evolution of genome complexity as successive falls and growth of CR. It is though that during this process, intermediate states with genomes of constant sizes suffering mutations and rearrangements could give birth to new functional sequences, thus providing raw material of species divergence, and biological complexity growth [Lynch 2000].

3.5.3. Genome size reduction

Theoretically, genome size reduction events are not expected to lower the CR, because any region of the genome large enough showed an almost random structure. Natural selection will ultimately determine the success of genome segments losses, but as we observed, intracellular bacterial parasites seemed to moved down (along the straight line of Figure 3.1-A) from larger genomes sizes to shorter genomes by fitting almost maximum CR during this process. An other example is given by ancient polyploid species as Arabidopsis thaliana [Hu et al. 2011] that recovers a maximum CR after diploidization and genome reduction. Contrasting with its close relative Arabidopsis lyrata still tolerating high proportion of genomic redundancy.

Thus in the case of gradual sharp or genome reduction, the pattern drawn by
Chapter 3. Random-like structure of DNA

the evolution of CR values is expected to be smoother than in the case of genome amplification.

3.5.4. Limits of CR space

We speculate that the space of complexity ratios filled in human writings Figure 3.1-B, is a region neglected for life. A non-random combinatorial structure of DNA is inconceivable for organisms with small genomes like viruses, phages or prokaryotes. By the effects of natural selection, simple forms of life – with genomes dimensions varying from the size of a few text paragraphs to the complete works of William Shakespeare – are probably forced to have a random like-structure, limiting thus, their alternatives to raise in complexity.

3.5.5. Hypotheses

All together these observations lead us to hypothesize that:

- A quasi-random combinatorial structure of DNA is a universal feature of non-polyploid genomes along all diversity of life.

- Polyploid genomes’ fate is to reach almost maximal complexity in the structure of their DNA – without taking into account genome reduction, CR’s raise would be as a function of time.

- Since the DNA combinatorial structure is quasi-random, genome complexity will only increase by DNA amplification, and posterior divergence of duplicates during evolution

However, these hypotheses can be falsified in some specific cases:

- Genomes of recent polyploid species evidencing a quasi-random DNA structure (high CR).

- Genomes evidencing a non-random DNA structure (low CR), due for example to a very strong GC content bias.

In this manuscript we described the combinatorial DNA structure of genomes. Ultimately we hypothesized a universal random-like structure along all diversity of life. It is very hard to think that such a structure is adaptive in its origin. However, far from being biologically irrelevant, useful properties may freely emerge from such random-like combinatorial structure in genomes. After all, exons, the main functional pieces of genes, are the elements with the most random-like DNA structure.

A simple law controlling genome statistical design for all kind of organisms makes nature modest and beautiful. Although it is hard to argue that by rolling a
dice millions times a functional genome will suddenly emerge, perhaps the fixation of some trial by natural selection let us imagine that it is feasible.
4. Life inside genomes, dynamics and predictions

In previous chapter we have shown that the intrinsic structure of genomes is random in informational terms. Or, in other words, that whatever aleatory combination of the four nucleotides would have (nearly) the same properties as a real genome. At some point in the study of the combinatorial structure of genetic elements (GEs) (see Complexity in repetitive elements and genes – low and high? in section 3.3), we also demonstrated that the complexity of some families of GEs were slightly lower when taken alone. However, given that the complexity ratio is almost identical for all chromosomes of a given genomes, we could expect that GEs are, in some way, homogeneously distributed in chromosomes.

This hypothesis, however, would stand in contradiction with the observed proportion of GEs, in close relative species. For example, we have seen in the introduction (Dynamics of genetics elements) how the dynamic of TEs that led them to present diametrically opposed proportions of DNA transposons and retrotransposons in Eukaryotes (Figure 1.4).

As first results, in order to get a wider view of the differences in proportions of GEs in eukaryotes, we wanted to report the proportions identified repetitive elements and genic regions found in our 31 eukaryotic species (see Mining of Genetic Species, page 25). The resulting proportions are summarized in the introductory item 4.1. Also some clade specific trend can be observed in mammals for example, important differences between species were substantial.

In this context, we will start to analyze the diversity and abundance of GEs – referred to as genetic species (GSs) in this chapter – with simple statistical tests against the hypothesis of their random distribution.

4.1. Non-random distribution of genetic species

The most simple hypothesis in relation to the distribution of the GSs, is that they are stochastically distributed in the genomes. In order to test this hypothesis, we simulated a thousand random distributions of GSs of each genome among their chromosomes, and tested if these proportions were conserved in each chromosome through a one sample t-test.

The number of t-test computed here was important, for each of the 548 chro-
Chapter 4. Life inside genomes, dynamics and predictions

Figure 4.1.: Proportion of major genetic species (GSs) in 31 eukaryotes. The phylogeny is adapted from [Jaime et al 2017], with a correction for amoeba *D. discoideum* [Roger & Simpson 2009]. Pie charts represent the proportion of GSs in each genome. Terms in the legend correspond to: 1) **Protein coding** sequences, 2) **Intronic** sequences and also untranslated regions of genes, 3) **LINE**, 4) **SINE**, and 5) **LTRs** all 3 retrotransposons, 6) **DNA transposon**, 7) **Satellite long tandem repeats**, 8) **Small RNA** mostly tRNA or snRNA pseudogenes, 9) **Simple repeat** or microsatellites, 10) **Low complexity** poly-purine or poly-pyrimidine (AT or GC rich), 11) **Unclassified repeat** not yet characterized repetitive elements, 12) **Miscellaneous unique** basically what remains after the identification of all previous elements.
4.2. Counterbalanced species abundances in genomes

4.2.1. Genetic species, dispersion and abundance

Ecologists frequently use relative species abundance (RSA) curves to compare the richness, the degree of dominance, and the number of rare species among communities. The raw data used in these plots is the total number of individuals per species sampled in the ecosystem. The most interesting property of RSA curves is that species are unlabeled in the ranking order; hence ecosystems can be compared whatever the species they contain.

RSA curves were build using our full set of GSs for each of our 548 chromosomes, and also for the corresponding 31 complete genomes. These numbers, our raw data, represent censuses of GSs analogous to the one realized by ecologists in ecosystems.

Figure 4.2 display RSA curves for a selected group of genomes and their largest chromosome respectively. Curves differ in many ways although two patterns are evident: 1- RSA curves of genomes and chromosomes are very similar – the only noticeable difference being a reduction in the number of GSs in chromosomes, 2- all RSA curves (from genomes and chromosomes) display the universal S-shape also observed in ecological environments [McGill et al. 2007, Hubbell 2001]. Both observations suggest a common mechanism of distribution of GSs in genomes and chromosomes.

4.2.2. Randomization of genetic species in genomes

To what extent chromosome’s RSA curves represent the random distribution of the complete set of elements of the genome? To answer, we used the same simulated data used in section Non-random distribution of genetic species. The mean expected abundance and standard deviation of this simulated data, were used to plot random expected RSA curves for chromosomes.
Chapter 4. Life inside genomes, dynamics and predictions

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Chrm.</th>
<th>Clade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canis familiaris</td>
<td>1</td>
<td>(Mammals)</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>7</td>
<td>(Fish)</td>
</tr>
<tr>
<td>Dicyostelium discoideum</td>
<td>2</td>
<td>(Amebozoa)</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>3R</td>
<td>(Invertebrates)</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>1</td>
<td>(Birds)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>1</td>
<td>(Mammals)</td>
</tr>
<tr>
<td>Monodelphis domestica</td>
<td>1</td>
<td>(Mammals)</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>1</td>
<td>(Mammals)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Chrm.</th>
<th>Clade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mus musculus</td>
<td>1</td>
<td>(Mammals)</td>
</tr>
<tr>
<td>Oryzias latipes</td>
<td>3</td>
<td>(Fish)</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>14</td>
<td>(Ampicompexa)</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>1</td>
<td>(Plants)</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>IV</td>
<td>(Fungi)</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>2</td>
<td>(Plants)</td>
</tr>
<tr>
<td>Tetraodon nigroviridis</td>
<td>1</td>
<td>(Fish)</td>
</tr>
</tbody>
</table>

Figure 4.2.: RSA curves.
Relative species abundances for some selected genomes (A) and their corresponding largest chromosomes (B)
According to Non-random distribution of genetic species and to results just pointed out for TEs in *D. melanogaster* [Bartolomé et al. 2002, Rizzon et al. 2002], an homogeneous random process, cannot account for the observed abundances of genetic species in chromosomes. However, if observed and simulated chromosomic RSA curves are superimposed, a notable agreement is detected. Figure 4.3 shows this fit for two chromosomes. This remarkable concurrence is permitted by the loss of species’ labels in the RSA plots, and hence by the allowed compensation by shift in the ranking order of abundances. For instance, the classes of functional tRNA and satellite elements are respectively at position 43 and 23 of the overall ranking of abundances in human genome Figure 4.3-A. However, in chromosome 1, their ranking are 33 and 42 respectively. That is, tRNA and satellites elements show higher and lower abundances than the expected by random distribution.

![Figure 4.3: Relative species abundance curves for human chromosome 1 (A) and chromosome 19 (B).](image)

Red and grey lines display respectively observed and simulated values for all genetic species in chromosomes. Dotted lines are 2 standard deviations around the mean simulated values. Numbers in parenthesis depict the observed (red) and the expected value (grey) in the ranking of abundances for few genetic classes in both chromosomes. Note the higher than the expected number of SINE/Alu elements in human chromosome 19.

In order to test the adjustment of the superimposition of simulated and observed data (red and grey curves in Figure 4.3), a Kolmogorov-Smirnov test was conducted. Overall, and after the corresponding correction for multiple testing, statistical differences between observed and simulated data were detected for only
76 out of 548 chromosomes tested (KS-test, \( P < 0.05 \)). This result obtained with the same data but summarizing species identity to a local rank, contrasts greatly with previous finding that only 4% of the GSs were found in the proportions expected by chance.

Thus, given the 86% agreement of chromosomes tested here, GSs distribution finally appears to be governed by some kind of stochastic process.

### 4.3. Diversity and chromosome length

If a purely stochastic process controls the abundance and diversity of genetic elements in chromosomes – as suggested by previous results –, it is expected that \( S \), the number of GSs present in a given chromosome, will increase with chromosome length. In ecology, it is universally observed that larger areas contain more species. Does this pattern hold true for chromosomes?

**Figure 4.4.: Species chromosome size relationship**

This plot represents the correlation between the number of genetic species (GSs) and the chromosome size in a log-log transformation. As complementary information, the size of the dots is a function of the number of individuals belonging to a given species.
4.4. Neutrality of species abundances and diversity

The standard species-area relationship in ecology is the Arrhenius power law [Arrhenius 1921] $S = cA^z$, where $S$ is the species number, $A$ is the area and $c$ and $z$ are constants. Following with our simile, species areas would be analogous to chromosome length. Figure 4.4 displays the correlation between the number of GSs and the chromosome size in a log-log transformation for both, polyploids and all chromosomes.

Also the correlation between the number of GSs and chromosome length was good, the first deficiency in the analogy we have being using between genomes and ecosystems appears. Polyploid chromosomes plotted seem to follow a slower increase in GSs along the axis of chromosome length. As this observation is completely predictable given that genomes subjected polyploidization would contain theoretically the same amount of GSs for larger size, we decided to also compute the statistical fit to the species-area relationship for non-polyploid eukaryotic species.

After least square fit of the power function, we effectively observed $c = 0.28$, $z = 0.27$ ($R^2 = 0.64$, $n = 548$) for all chromosomes studied (including fishes and plant species – as polyploids), and $c = 0.50$, $z = 0.25$ ($R^2 = 0.81$, $n = 412$) (excluding them as mentioned above).

In both cases, the adjustment was statistically significant to a lineal regression model ($P << 0.001$). Thus, and likewise in community ecology, eukaryote chromosomes display the universal species-area relationship with $z$ values corresponding to regional spatial scales [Rosenzweig 1995].

At this point, we believe to have strong evidences that the distribution of GSs among chromosome is characterized by a strong stochastic component able to explain 1- the observations raised by the comparison of RSA curves of simulated and observed distributions of GSs (see previous section Randomization of genetic species in genomes) 2- and this last result showing how the number of GSs present in a chromosome is highly correlated to its length. In order to improve and test more accurately this observation, we finally ask if a neutral dynamic model can predict this shared demographic pattern of genomes.

4.4. Neutrality of species abundances and diversity

Similar to the kinetic theory of ideal gases in physics the unified neutral theory of biodiversity (UNTB [Hubbell 2001]) is a stochastic theory assuming equivalence among interacting individuals. The theory assumes that diversity in a local community of individuals is maintained by migration from the metacommunity at a constant rate ($m$). Births and deaths in the local community occur at constant rates during generation regardless the species. The metacommunity dynamics is controlled by speciation at a single constant rate ($\nu$) [Rosindell et al. 2011, Alonso et al. 2006].
Table 4.1.: Result of the fit to UNTB and test of neutrality for selected chromosomes

The table depicts parameters and statistics estimated for a selection of chromosomes of different species. Chromosomes are arranged according to p-value from the less to the most neutral. J the total number of genetic elements; S is the number of genetic species; ΔH the difference between the observed and the expected evenness (Shannon’s diversity index); ✓ the fundamental diversity number; m the migration rate; P and Q-val are statistical significances of the neutral test before and after false-discovery rate correction. The last column shows the model (Ewens or Etienne) that best fitted the empirical distribution of genetic elements in the chromosome after likelihood-ratio test (p < 0.05, df = 1). None of the 548 chromosomes over the 31 eukaryotes genomes showed significant deviations from neutrality (Q-val < 0.05).

For genomes, we realized that each chromosome is the physical arena where GSs die and are replaced by other elements of the same or different species. These GSs could come from the same chromosome, or from any other chromosome of the genome. We assume that each chromosome represents a local community of J elements and S different genetic classes (species) while the rest of the genome corresponds to the metacommunity of size J_M. Thus, we used the sample of the total number of functional and non-functional elements in each chromosome as raw-data to optimize by maximum likelihood (ML) neutral model’s parameters m and ✓ (= 2J_M nuclei) using Ewens and Etienne’s sampling formula (Equation 2.6).

Deviations from neutrality were detected in 33 out of 548 (6.0%) chromosomes. However, these deviations vanished at all after multiple testing correction (FDR < 0.05, see Table 4.1 for a summary of the results). We conclude that Hubbell’s neutral model fits abundance and diversity of GSs in all the chromosomes of the 31 eukaryotes genomes analyzed.
Almost one hundred years ago ecologists recognized the universal uneven distribution of species abundance, and the species increment in larger areas [Magurran 2004]. Just a decade ago however, neutral demographic processes emerged as the simplest mechanical explanation behind both patterns in communities [Hubbell 2001]. More recently, Michael Lynch and John S. Conery [Lynch & Conery 2003] hypothesized that complexity of eukaryote genomes emerged passively during evolution as a consequence of population size reduction. Here we demonstrated that a simple stochastic process associated to a few number of parameters fits the pattern of abundance and diversity of genetic species along a great diversity of eukaryote genomes.

An example of implementation of a neutral model to explain the distribution of genetic elements in genomes can be found very recently in the work conducted by Bart Haegeman and Joshua S. Weitz with 6 bacterial genomes [Haegeman & Weitz 2012]. In this work authors defined a neutral model for the evolution of the genomes able to predict the proportions of 1- genes shared by all the genomes, 2- genes absent from some genomes, 3- and species-specific genes. This model, that combines birth and death process among individuals of the same species, and gene transfer between species, was able to reproduce the observation that most genes are either specific to one genome or shared by all species. Also the model proposed by the authors of this work is only applicable to bacteria, we do think that their results are in the same lines as ours, as the apparition and conservation of genes in genomes are following a neutral process.

We are certainly aware that the fit of a neutral pattern does not necessarily imply the existence of a neutral process behind the pattern (see Power and specificity of neutral test section 2.2.8), but the excellent, taxonomically broad fit of neutral theory to genomic element diversity and abundance raises the unavoidable question: why is there not a stronger signature of natural selection in ecological communities or in genomes at large scales? Ecologists have recognized the existence of many kinds of trade-offs, for instance species with high dispersal rates are not good competitors. However, it is not yet known to what extent such trade-offs maintain diversity or are consistent with, and permit, neutral dynamics. For genomes, the mechanisms that maintain element diversity, and whether these involve trade-offs, are not yet understood. Which mechanisms operate will also depend on whether genome size is under strong or weak selection [Cavalier-Smith 2005]. More likely, element diversity of genomes results from some combination of neutral drift and selection on different genetic species [Mustonen & Lässig 2009]. Independent of the answer, the model tested here should be the null hypothesis to test for alternative mechanisms explaining species abundance and diversity in eukaryote genomes.
Part II.

Detection of selective pressures in genomes
5. Searching for evolutionary patterns in functionally linked group of genes

5.1. Gene-set selection analysis on functional modules

Mammals, represented by human, chimpanzee, rat and mouse, and also five *Drosophila* species were studied. For each corresponding genome, genes were ranked into four lists according to the estimation of i) synonymous (*dS*), ii) non-synonymous (*dN*) rates of substitution, iii) selective pressures (*ω = dN/dS*), and iv) the change of selective pressures between (*A*) ancestor and (*D*) descendant species (*ΔωD = ωD − ωA*) along the phylogeny Figure 2.4 (see section: GSSA, evolutionary and statistical simulations, page 37 for details on the methodology).

The application of the GSSA over the lists of genes ranked by *dS*, *dN*, *ω* and the *Δω* values yielded a large number of functional modules with rates that were significantly skewed toward the extremes of the lists (Table 5.1) in both, mammal and *Drosophila* species. For instance, 11% of the GO terms, and 15% of the KEGG pathways tested were found to be significantly enriched in genes with high *ω* rates (SH*ω, 5% false-discovery rate, FDR) in mammals. In *Drosophila*, slightly lower proportions were found, with alternatively, 4.1% and 2.6% of GO terms and KEGG pathways respectively.

<table>
<thead>
<tr>
<th></th>
<th>SH* KEGG</th>
<th>GO</th>
<th>SL* KEGG</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mammals</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>dS</em></td>
<td>15 (1.9)</td>
<td>12 (2.1)</td>
<td>364 (6.5)</td>
<td></td>
</tr>
<tr>
<td><em>dN</em></td>
<td>145 (18.2)</td>
<td>230 (28.9)</td>
<td>1,839 (32.9)</td>
<td></td>
</tr>
<tr>
<td><em>ω</em></td>
<td>123 (15.5)</td>
<td>206 (25.9)</td>
<td>1,675 (30.0)</td>
<td></td>
</tr>
<tr>
<td><em>Δω</em></td>
<td>64 (8.0)</td>
<td>107 (13.4)</td>
<td>818 (14.7)</td>
<td></td>
</tr>
<tr>
<td><strong>Drosophila</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>dS</em></td>
<td>18 (3.1)</td>
<td>26 (4.5)</td>
<td>1,263 (18.9)</td>
<td></td>
</tr>
<tr>
<td><em>dN</em></td>
<td>31 (5.3)</td>
<td>26 (4.5)</td>
<td>2,097 (31.5)</td>
<td></td>
</tr>
<tr>
<td><em>ω</em></td>
<td>15 (2.6)</td>
<td>24 (4.1)</td>
<td>1,321 (19.8)</td>
<td></td>
</tr>
<tr>
<td><em>Δω</em></td>
<td>2 (0.3)</td>
<td>7 (1.2)</td>
<td>184 (2.8)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1.: Numbers and percentages of functional modules with significant results after GSSA.

For Significantly High (SH) and Significantly Low (SL) results after correction for multiple testing (5%FDR).

Table 5.1 also reveals that functional modules with genes changing at significantly low *ω* ratios (SL*ω*), and therefore showing a distribution shifted towards
the bottom of the ranked list (see Figure 5.1), were more frequent than modules under a significantly high $\omega$ (SH$\omega$). This observation is in agreement with the fact that purifying selection is the predominant form of selection in biological systems. Moreover, in support to the neutral character of synonymous mutations, and to the effects of population size in the final outcome of selection [Lynch 2007], GSSA results show a higher number of significant deviations of $dS$ in *Drosophila* rather than in mammals.

When contrasted to the $\omega$ rate of ancestral sequences, the tendency observed is that only a minor proportion of functional terms were under significantly high or low selective pressures. Specifically, increased or decreased $\omega$ values up to the external branches (recorded by positive and negative values of $\Delta\omega$) were observed for only half of the cases where a significant increase or decrease of $\omega$ was identified in descendants. This observation points out the conservative character of the selective constraints in functionally related groups of genes.

Results of the GSSA for mammals and *Drosophila* are summarized in Figure 5.1A and Figure 5.1B respectively. This figure shows a selection of functional terms with significantly high or low rates for each of the evolutionary variable considered. The first striking point, also pointed by Table 5.1 is the amount of significant results. By considering the whole set of genes, the GSSA is able to detect functional biases with much more statistical power than considering only genes belonging to a given evolutionary scenario.

### 5.1.1. Clade specific patterns

The first pattern that stands out when looking at Figure 5.1 is the differentiation of clades, human together with chimpanzee, mouse with rat, and among the *melanogaster* subgroup, *D. yakuba* and *D. erecta* also show similar patterns. For instance, functional terms associated to neurological process and sensory perception clearly contrasted between primates and rodents (Figure 5.1-A). This segregation by clade is expected taking into account the weight of the evolutionary history since ancestral state in the final count of synonymous and non-synonymous changes. As an example we can focus on the terms related to neurological process in mammals. In both, human and chimpanzee, neurological process and sensory perception show a significant raise in $\omega$ when compared to their common ancestor—which can be seen through $\Delta\omega$ significance. In contrast, in both rodents, the values of $\Delta\omega$ are significantly low for these functional terms related to neurological process.

Alternatively, functional modules associated to *Immunity* and *Defense response* evolved at significantly higher rates than expected in rodents, but decreased significantly in relation to the ancestral rates in primates. Such functional differences between primates and rodents were previously observed when pooling groups of
5.1. Gene-set selection analysis on functional modules

species [Kosiol et al. 2008] (see section: Genomic study of selective pressures in set of genes, page 13). And besides these fast evolving categories, other functional modules such as Development, and Transcription/Transduction comparatively evolved at very low dN and ω ratio but experienced a higher relaxation of the ancestral constraints (+Δω) in primates than in rodents.

In Drosophila, most of the GO terms significantly associated to high dN and ω in Drosophila were also unevenly distributed within the two clusters of the phylogeny (Figure 5.1-B). GO terms such as sensory perception, defense response, immune response and metabolic process among others, presented a remarkable divergence in the monophyletic groups of D. erecta and D. yakuba but they were not significant in D. sechellia, D. melanogaster and D. simulans. Most of GO terms related to Development, Transcription and Translation (Figure 5.1-A and -B) were found to be constrained by purifying selection with significantly low rates of ω (5% FDR) in both taxa.

5.1.2. Species specific enrichment

Going further in the analysis of the results we are able to see species-specific functional enrichment. Following with the example of sensory perception in rodents, we can see how G-protein coupled receptor protein signaling pathway has a significantly high value of ω in rat but not in mouse.

An important result is also the significant differences observed between human and chimpanzee in few neurological processes as, the KEGG pathway Ha04360: Axon guidance or the GO term GO0007268: synaptic transmission. In both cases, genes related to these functional terms are more conserved in human.

In Drosophila, clade consistency in the group D. erecta-D. yakuba was higher. However some interesting differences raised as 1- Drug metabolism - other enzymes that appears more conserved in D. erecta or 2- neurogenesis and nervous system development that are conserved in D. sechellia not significant in D. simulans and conserved, but in a process of relaxation since ancestor (high Δω) in D. melanogaster.

Figure 5.1. (following page): GSSA of evolutionary variables.

The figure shows a selection of GO terms and KEGG pathways with significant and not significant deviations after GSSA of evolutionary rates in mammals (A) and Drosophila (B) species. Colored boxes represent functional modules with genes significantly accumulated at the corresponding extremes of the ranked list as explained in Figure 2.5. The number inside each box represents the percentage of the total number of genes of the functional module (in parenthesis) that contribute to its significance. Here we reported the numbers of the first significant partition after FET and FDR. Topologies represent the phylogenetic relationships of species.
Chapter 5. Searching for evolutionary patterns in functionally linked group of genes

| KEGG pathways                      | dS | dN | o | ω | A0
|-------------------------------------|----|----|---|---|---
| Olfactory transduction (136)       | 25 | 19 | 19| 12| 30
| Axon guidance (109)                | 71 | 64 |    |   |   
| Complement and coagulation cascades (43) | 27 | 69 | 39 | 82| 18
| Cytokine-cytokine receptor interaction (174) | 17 | 25 | 13| 24| 37| 14
| Antigen processing and presentation (41) | 29 | 63 |    |   |   
| Spliceosome (86)                   | 20 | 33 | 80| 50 | 33| 79
| Melanogenesis (69)                 | 91 |    |   |   |   
| Wnt signaling pathway (111)        | 30 | 55 | 86| 13 | 38| 90
| GnRH signaling pathway (73)        | 79 | 68 |    |   |   
| TGF-beta signaling pathway (66)    | 57 | 59 | 93| 59| 36| 46
| Drug metabolism - cytochrome P450 (28) | 53 | 32 | 50| 39| 35| 28

| GO annotation                      | dS | dN | o | ω | A0
|-------------------------------------|----|----|---|---|---
| Neurological process and sensory perception | 17 | 32 | 28| 17| 29| 77| 56| 16
| G-protein coupled receptor protein signaling pathway (795) | 56 | 74 | 70| 73 | 88| 83| 86| 84| 90| 81
| sensory perception of smell (144)  | 59 | 72 | 96| 65 | 49| 81| 27 | 31| 80| 42| 84| 91
| central nervous system development (347) | 73 | 70 | 71| 62 | 75| 86| 66 | 64| 83| 68| 64| 91
| generation of neurons (437)        | 26 | 16 | 94| 21| 29| 97| 83| 83| 86| 77| 89| 94
| neurological system process (793)  | 48 | 22 | 41| 39| 57| 74| 58| 90| 77| 93|   
| cognition (520)                    | 66 | 73 | 93| 63 | 63| 92| 75 | 58| 79| 4 | 61| 60| 90
| synaptic transmission (287)        | 29 | 71 | 72| 93| 55 | 56| 77| 43 | 66| 83|   

| Immunity and defense response      | dS | dN | o | ω | A0
|-------------------------------------|----|----|---|---|---
| Immune system process (759)        | 36 | 18 | 23 | 20| 14
| defense response (495)             | 22 | 51 | 20| 23 | 32| 13
| response to stress (1490)          | 11 |    |   |   |   
| coagulation (132)                  | 63 | 76 |    |   |   
| inflammatory response (279)        | 42 | 65 | 22| 29 | 33| 12
| positive regulation of immune response (102) | 29 |    |   |   |   

| Reproduction                      | dS | dN | o | ω | A0
|-----------------------------------|----|----|---|---|---
| fertilization (51)                | 50 | 50 | 98| 50| 19
| urogenital system development (92) | 68 | 70 | 62| 80| 69 | 34| 46| 85
| lipid metabolic process (666)     | 68 | 84 | 84| 68| 44| 84
| cellular amine metabolic process (332) |   |   |   |   |   
| protein catabolic process (481)   |    |   |   |   |   

| Development                      | dS | dN | o | ω | A0
|----------------------------------|----|----|---|---|---
| muscle system process (149)      | 88 | 59 | 88| 88 | 60| 90| 25
| embryonic morphogenesis (252)    | 74 | 88 | 89| 86| 86 | 83| 89| 74| 62| 81
| heart development (180)          | 98 | 64 | 92| 68| 68| 64| 52| 40
| eye development (102)            | 92 | 18 | 98| 88 | 81| 81| 16| 85| 82
| ear development (75)             | 77 | 26 | 78| 70| 61| 61| 44| 37|   

| Transcription and translation modulation | dS | dN | o | ω | A0
|----------------------------------------|----|----|---|---|---
| chromatin modification (225)          | 86 | 62 | 93| 88| 79 | 76| 86| 48
| transcription (1833)                  | 67 | 44 | 85| 81 | 59| 44| 86| 40
| translation (324)                    | 61 | 34 | 61| 89 | 78| 18| 17|   
| mRNA processing (244)                | 71 | 56 | 97| 67 | 51| 87| 29 | 62| 80| 24
| RNA splicing (213)                   | 73 | 56 | 84| 69 | 52| 88| 26 | 67| 85| 86
| apoptosis (765)                       | 62 | 51 | 97| 59| 12| 14| 19| 19|   

80
5.1. Gene-set selection analysis on functional modules

<table>
<thead>
<tr>
<th></th>
<th>D. melanogast</th>
<th>D. simulans</th>
<th>D. sechellia</th>
<th>D. yakuba</th>
<th>D. erecta</th>
</tr>
</thead>
<tbody>
<tr>
<td>dS/dN (Δ0)</td>
<td>100</td>
<td>70</td>
<td>31.97</td>
<td>50</td>
<td>50.63</td>
</tr>
<tr>
<td>GO annotation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurological process and sensory perception</td>
<td>24 34</td>
<td>53 59</td>
<td>33 28 92</td>
<td>37 26</td>
<td></td>
</tr>
<tr>
<td>Spliceosome (83)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensory perception (173)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensory perception of chemical stimulus (123)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensory perception of smell (43)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEGG pathways</td>
<td>Aminocacyl-tRNA biosynthesis (29)</td>
<td>Drug metabolism - other enzymes (34)</td>
<td>Limonene and pinene degradation (44)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunity and defense response</td>
<td>31 37</td>
<td>47 41</td>
<td>30 38</td>
<td>19 19 19 19 19</td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>20 23</td>
<td>40 61</td>
<td>19 10 21</td>
<td>64 30 34</td>
<td>32 40 40</td>
</tr>
<tr>
<td>Metabolism</td>
<td>47 18 35</td>
<td>20 6 12</td>
<td>16 4 11 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>42 42 42 42</td>
<td>20 6 12</td>
<td>16 4 11 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>27 42</td>
<td>24 19 28 28 28 28 28 28</td>
<td>13 15 15 15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>38 38 38 38</td>
<td>20 6 12</td>
<td>16 4 11 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>77 77 77 77</td>
<td>20 6 12</td>
<td>16 4 11 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproduction</td>
<td>26 13</td>
<td>17 17 34</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
</tr>
<tr>
<td>Development</td>
<td>26 22 45</td>
<td>17 6 26</td>
<td>25 15 33</td>
<td>10 54 54</td>
<td>10 7 7</td>
</tr>
<tr>
<td>Development</td>
<td>17 17 34</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>26 22 45</td>
<td>17 6 26</td>
<td>25 15 33</td>
<td>10 54 54</td>
<td>10 7 7</td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td>14 14 14 14 14 14 14 14</td>
<td>5 29 56</td>
<td>13 10 65</td>
<td>12 8 22</td>
<td></td>
</tr>
</tbody>
</table>

RNA processing (187) | transcription (541) | translation (264) | gene expression (1000) |           |           |
# Chapter 5. Searching for evolutionary patterns in functionally linked group of genes

<table>
<thead>
<tr>
<th>Biological process</th>
<th>Enrichment in PSGs</th>
<th>GSSA LOW ω rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensory Perception</td>
<td>H Pr† H Pr† R†</td>
<td>me† ya† er†</td>
</tr>
<tr>
<td>Cell surface receptor-mediated signal transduction</td>
<td>H Pr†</td>
<td></td>
</tr>
<tr>
<td>Cell adhesion</td>
<td>H</td>
<td>H† C† me† er†</td>
</tr>
<tr>
<td>Amino acid transport</td>
<td>Pr</td>
<td>R†</td>
</tr>
<tr>
<td>Protein amino acid glycosylation</td>
<td>Pr</td>
<td>H†</td>
</tr>
<tr>
<td>Amino acid transport</td>
<td>C</td>
<td>C†</td>
</tr>
<tr>
<td>Hearing / Perception of sound</td>
<td>H Pr†</td>
<td>M† R† ya† er†</td>
</tr>
<tr>
<td>Neurological process</td>
<td>Pr†</td>
<td>H† M† R† me† se† er† ya†</td>
</tr>
<tr>
<td>Synaptic transmission</td>
<td>Pr</td>
<td></td>
</tr>
<tr>
<td>Signal transduction/intracellular signaling cascade</td>
<td>H Pr†</td>
<td>Dr H† C† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Ion transport</td>
<td>H</td>
<td>H Dr H† M† R† me† se† er†</td>
</tr>
<tr>
<td>Potassium ion transport</td>
<td>Pr</td>
<td>H† C† M† R†</td>
</tr>
<tr>
<td>Inorganic anion transport</td>
<td>Pr</td>
<td>M† R†</td>
</tr>
<tr>
<td>Intracellular protein traffic</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>Transport</td>
<td>Dr</td>
<td>me† se† er† ya†</td>
</tr>
<tr>
<td>Protein transport</td>
<td>H Dr</td>
<td>H† C† M† R† me† si† se† er† ya†</td>
</tr>
<tr>
<td>Metabolism of cyclic nucleotides</td>
<td>H</td>
<td>M† R†</td>
</tr>
<tr>
<td>Protein metabolism &amp; modification</td>
<td>H C Dr</td>
<td>H† C† M† R† er† ya†</td>
</tr>
<tr>
<td>Phosphorus metabolism/phosphorylation</td>
<td>H C Dr</td>
<td>H† C† M† R† me† ya† er†</td>
</tr>
<tr>
<td>Purine metabolism</td>
<td>C</td>
<td>M† R† se†</td>
</tr>
<tr>
<td>Carbohydrate biosynthesis</td>
<td>Pr</td>
<td>M† R†</td>
</tr>
<tr>
<td>Cation transport</td>
<td>H</td>
<td>H† M† R†</td>
</tr>
<tr>
<td>Nervous system development</td>
<td>Dr</td>
<td>H† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Skeletal development</td>
<td>C</td>
<td>M† R†</td>
</tr>
<tr>
<td>Organ development</td>
<td>Dr</td>
<td>H† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Post-embryonic development</td>
<td>Dr</td>
<td>M† me† ya† er†</td>
</tr>
<tr>
<td>Embryonic development</td>
<td>Dr</td>
<td>H† C† M† R† ya† er†</td>
</tr>
<tr>
<td>Ectoderm development</td>
<td>H C Dr</td>
<td>C† M† R† me† ya† er†</td>
</tr>
<tr>
<td>Cell proliferation and differentiation</td>
<td>C Dr</td>
<td>H† C† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>Dr</td>
<td>H† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Cell structure/morphogenesis</td>
<td>C Dr</td>
<td>H† C† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Cell structure and motility</td>
<td>C</td>
<td>H† M† R† se†</td>
</tr>
<tr>
<td>Inhibition of apoptosis</td>
<td>Pr†</td>
<td>H† ya†</td>
</tr>
<tr>
<td>Cell-cell signalling</td>
<td>Dr</td>
<td>H† C† M† R† me† se† ya† er†</td>
</tr>
<tr>
<td>Regulation of nucleobase</td>
<td>H C</td>
<td>H† C† M† R† er†</td>
</tr>
<tr>
<td>Translation</td>
<td>Dr</td>
<td>M† R† me† si† se† ya† er†</td>
</tr>
<tr>
<td>Transcription</td>
<td>H C Dr</td>
<td>H† C† M† R† er†</td>
</tr>
<tr>
<td>Protein catabolism</td>
<td>H C</td>
<td>H† C† M† R†</td>
</tr>
<tr>
<td>Interferon-mediated immunity</td>
<td>Pr†</td>
<td></td>
</tr>
</tbody>
</table>

82
5.1. Gene-set selection analysis on functional modules

<table>
<thead>
<tr>
<th>Biological process</th>
<th>Enrichment in PSGs</th>
<th>GSSA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5 6 7</td>
<td>HIGH ω rates</td>
</tr>
<tr>
<td>Olfaction/Sensory perception of smell</td>
<td>H Pr†</td>
<td></td>
</tr>
<tr>
<td>Chemosensory perception</td>
<td>H Pr†</td>
<td></td>
</tr>
<tr>
<td>G-protein-mediated signaling</td>
<td>H H Pr†</td>
<td></td>
</tr>
<tr>
<td>DNA/nucleic acid metabolism</td>
<td>C Dr C† M† R† me† se†</td>
<td></td>
</tr>
<tr>
<td>Amino acid metabolism</td>
<td>H C</td>
<td></td>
</tr>
<tr>
<td>Proteolysis</td>
<td>Dr M† R† me† si† se† ya†</td>
<td></td>
</tr>
<tr>
<td>Fatty acid/Lipid metabolism</td>
<td>H Dr M† R†</td>
<td></td>
</tr>
<tr>
<td>Carbohydrate metabolism</td>
<td>Dr se† ya† er†</td>
<td></td>
</tr>
<tr>
<td>Adult reproduction and gametogenesis</td>
<td>Dr se†</td>
<td></td>
</tr>
<tr>
<td>Spermatogenesis and motility</td>
<td>Pr† Pr</td>
<td></td>
</tr>
<tr>
<td>Immune response</td>
<td>Pr† Pr† H</td>
<td></td>
</tr>
<tr>
<td>Inflammatory response</td>
<td>Ro†</td>
<td></td>
</tr>
<tr>
<td>Defense response</td>
<td>Ro†</td>
<td></td>
</tr>
<tr>
<td>Response to wounding</td>
<td>Ro†</td>
<td></td>
</tr>
<tr>
<td>Hummoral imm. resp. mediated by circulating Ig</td>
<td>Ro†</td>
<td></td>
</tr>
<tr>
<td>T-cell-mediated immunity</td>
<td>Pr†</td>
<td></td>
</tr>
<tr>
<td>Natural killer-cell-mediated immunity</td>
<td>Pr†</td>
<td></td>
</tr>
<tr>
<td>B-cell- and antibody-mediated immunity</td>
<td>Pr†</td>
<td></td>
</tr>
<tr>
<td>Response to pest pathogen or parasite</td>
<td>H C</td>
<td></td>
</tr>
<tr>
<td>Stress response</td>
<td>C Ro†</td>
<td></td>
</tr>
<tr>
<td>Response to external stimulus</td>
<td>Ro†</td>
<td></td>
</tr>
<tr>
<td>Sensory Perception</td>
<td>H Pr† H Pr†</td>
<td></td>
</tr>
<tr>
<td>Cell surface receptor-mediated signal transduction</td>
<td>H Pr† C†</td>
<td></td>
</tr>
<tr>
<td>Cell adhesion</td>
<td>H Pr†</td>
<td></td>
</tr>
<tr>
<td>Amino acid transport</td>
<td>Pr</td>
<td></td>
</tr>
<tr>
<td>Protein amino acid glycosylation</td>
<td>Pr</td>
<td></td>
</tr>
<tr>
<td>Amino acid transport</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2.: Functional enrichment results using gene-by-gene and gene-set approaches.

The table depicts selected biological functions enriched by PSGs as cited in references 1 to 7, and the corresponding significant result observed after GSSA of ω values. References 1 to 7 correspond to cites [Clark et al. 2003], [Nielsen et al. 2005], [Mikkelsen et al. 2005], [Arbiza et al. 2006], [Bakewell et al. 2007], [Kosiol et al. 2008] and [Clark et al. 2007] in the manuscript, respectively. Abbreviations: SHv: statistically significant high v values; SLv: statistically significant low v values; H: H. sapiens; C: P. troglodytes; Pr: primates; M: M. musculus; R: R. norvegicus; Ro: rodents; me: D. melanogaster; si: D. simulans; se: D sechelia; ya: D. yakuba; er: D. erecta; Ds: Drosophila species. † p=0.05; ‡ p=0.001
Chapter 5. Searching for evolutionary patterns in functionally linked group of genes

5.1.3. Comparison of evolutionary variables used

The fact that most of the functional modules with SHω and SLω correlate with changes in dN, suggests that selective pressures are mainly driven by non-synonymous rather than by synonymous substitutions. Moreover, according to the expectation of the nearly neutral theory, a low but still considerable number of significant associations of functional modules to dS were found in Drosophila (19.5%) and rodents (11.3%), while in primates (6.4%), where population sizes are known to be smaller, the number of significant modules was smaller [Petit & Barbadilla 2009].

5.2. Positively selected genes among fast and slow evolving functional modules

We have seen how GSSA was able to find significant functional enrichment towards the extremes a list of genes ranked by an evolutionary variable. However this result stands, in some way, in contradiction with the lack of significance observed when analysing only set of genes under positive selection, leading us to the question: To what extent genes under positive selection contribute to the significance of functional modules in mammals and Drosophila species after GSSA?

To answer this question, a branch-site test of positive selection –the most sensitive and widespread– was conducted on terminal branches of both phylogenies (Figure 2.4). Overall, 715 positively selected genes (PSGs) were detected in mammals and 626 in Drosophila. In this last section, we are going to see how these PSGs fitted into previous results.

5.2.1. Descriptive analysis

The idea here, was to describe the distribution of PSGs among our functional modules with either SH, SL or NS rates of ω. Thus, we first plotted the distribution of all functional categories (putting together KEGG pathways and GO terms of both mammals and Drosophila) according to their mean values of dN and dS. In such representation, functional categories detected to have a SHω are expected to be above a relative line passing through the origin with a slope representative of a given cutoff value of ω (\(\frac{dN}{dS}\)). And this tendency is indeed appreciable in Figure 5.2-A between functional modules providing significant and not significant results after GSSA of the ω ratio.

When combining the dataset formed by PSGs with all functional categories analysed, we found that the total number of functional modules containing at least one PSG represents respectively 55%, 53%, and 42% of the functional categories with SH, SL and NS results after GSSA (for ω values). Figure 5.2-B is a graphical representation of this result, it is the exact replicate of Figure 5.2-A but only
5.2. Positively selected genes among fast and slow evolving functional modules

Figure 5.2.: Positive selection in functional modules’ evolutionary scenarios. Circles and triangles represent the median values of $dN$ and $dS$ for, respectively, functional categories in mammals, and *Drosophila* species. SH$\omega$, SL$\omega$ and NS$\omega$ results after GSSA are shown in red, blue and grey. Yellow dots depict the genomic median for *H. sapiens* (1), *P. troglodytes* (2), *M. musculus* (3), *R. norvegicus* (4), *D. simulans* (5), *D. sechellia* (6), *D. melanogaster* (7), *D. yakuba* (8) and *D. erecta* (9). (A) represents all functional categories, while (B) shows only modules containing at least 1 PSG. Note that they are distributed along a wide range of $dS$ and $dN$ and in functional categories with significant (red/blue), and even NS (gray) results after the GSSA ($\omega$ ratio).
Chapter 5. Searching for evolutionary patterns in functionally linked group of genes

keeping functional categories that include at least one PSG.

The proportion of functional categories “kept” suggests that:

- The accumulation of PSGs is not the main driver in the evolution of functional modules changing at SHω ratios in the genome. Functional modules such as complement and coagulation cascades in human, gonad development in chimpanzee, regulation of innate immune response in mouse, primary immunodeficiency in rat, or spermatid differentiation in D. melanogaster are examples of functional categories evolving at significantly elevated ω values without any PSG.

- Molecular adaptation takes also place in functional modules under strong selective constraints with SLω (see first part of Table 5.2). For instance, apoptosis in human, generation of neurons in chimpanzee, tissue development in mouse, Wnt signaling pathway in rat, eye development in D. melanogaster, wing disc development in D. yakuba, and generation of neurons in D. erecta are some of the functional modules that are, at the same time, evolving at SLω, and carrying PSGs.

- An important number of functional modules without significant differences in ω ratios (grey dots in Figure 5.2-B) still contain genes under positive selection. For instance, homologous recombination in humans, brain development in chimpanzee, female or male sex differentiation in mouse, regulation of mitotic cell cycle in rat, chromatin modification in D. sechellia, and oogenesis in D. melanogaster.

These results are in agreement with previous observations in Drosophila were it was emphasized that not every mutation under positive selection responds to a change in selection [Mustonen & Lässig 2009]. Beneficial changes could occur at evolutionary equilibrium, repairing previous deleterious changes and restoring affected function [Mustonen & Lässig 2009].

5.2.2. Statistical approach

Finally, we ask if PSGs preferentially concentrate in functional modules evolving at faster rates in different genomes. In this sense, we computed the mean number of PSGs in functional modules with SHω and SLω results (red and blue dots in Figure 5.2-B). The expectation that functional modules evolving at high ω ratio are carrying a higher numbers of PSGs was confirmed in rodents (p < 0.001), all mammals together (p < 0.001) and Drosophila (p < 0.001). However in primates the result was far from significance (p = 0.47) indicating an random distribution of PSGs among functional categories with SHω and SLω. Note that, in consequence
of the larger number of PSGs in rodents, the result of primates is not sufficient to lower the reported significance of the test among mammals.

To contrast these results, the same analysis was conducted with PSGs detected in previous works (Kosiol et al. 2008 for mammals and Clark et al. 2007 for Drosophila). The pattern of distribution of PSGs in functional modules was in exact agreement with the above mentioned results: significantly skewed (p<0.001) towards higher numbers of PSGs in mammals, rodents, and Drosophila species, but showing no differences in primates (p = 0.73).

In summary, PSGs are frequently observed in functional modules evolving under a wide range of evolutionary scenarios; however, they concentrate more frequently in functional groups of genes changing at elevated rates in rodents and Drosophila species. Alternatively, PSGs were evenly distributed in functional modules changing at the extreme rates of evolution in primates. This observation suggests that a more complex scheme than the cumulative differences of PSGs must rely on the reported adaptive differences in human and chimpanzee genomes. The search for integrative factors taking into account the action of multiple genes other than only those which have been targeted by positive selection [He et al. 2010], could provide a more accurate view for the analysis of the integrated framework underlying adaptation in complete genomes.

5.3. Discussion

Evolutionary biologists recognize that natural selection works on phenotypes indirectly by changing the frequency of genes in populations [Lewontin 1974]. Since the revolution of molecular techniques and its use in evolutionary genetics, the statistical search for adaptation at a gene level has superseded the complexity of measuring fitness in nature [Endler 1986]. Nowadays, to measure the influence of natural selection on phenotypes we typically look for adaptive evidences on genes and afterwards look forward over-represented functional modules PSGs found in the genomes. This approach consists thus, in two independent steps, and disregard the cooperative action of the network of genes underlying phenotypes [He et al. 2010, Alvarez-Ponce et al. 2009].

The aim of the GSSA is not to test for evolutionary constraints on individual genes as has been addressed in several previous studies. GSSA tests for significant differences in rates over functionally related groups of genes and therefore, the relative contribution of a gene inside a functional category is dependent of the genomic distribution of evolutionary rates.

The results brought out in this work confirm completely the trends observed in previous works focusing on PSGs [Clark et al. 2003, Shapiro & Alm 2008, Kosiol et al. 2008] (see Figure 1.5). The fact that most of these candidate functional categories were found among our results with significantly high values of ω can be
explained by the fact that the amount of PSGs needed to approach significance in previous work, is certainly contributing to the raise of the \( \omega \) value of the whole functional class. Moreover, this functionally based approach was able to identify as the main targets of adaptive changes, biological functions in individual species Table 5.2.

By defining functional modules submitted to specific selective pressures, this study represents a clear step forward in the demonstration of the hypothesis that phenotypes change during evolution by the coordinated action of genes. Although GSSA is not a test for positive selection, it is evident that functional modules containing PSGs can be significantly detected by this method. Results and trends brought by previous works (see Figure 1.5 and Table 5.2) are fitting perfectly in our positive outcomes.

With respect to the presence of PSGs in functional categories with either significantly low or not significant in rates of \( \omega \) some comment need to be done, at the risk of sounding like devil’s advocate. It is important here to keep in mind that genes are not usually annotated with only one functional category, thus, in some sense, it is expected to find SL\( \omega \) or even NS\( \omega \) functional categories in Figure 5.2-B. Nonetheless the biological meaning of this multiple annotation is important to consider. A protein involved in both a conserved and an accelerated pathway would intuitively be submitted to strong selective pressure. Taking this into account, the presence of PSGs in conserved functional modules can not be just a methodological artifact. Moreover, the statistical test conducted shows that–even assuming that multiple annotation would contribute to lower its power– the significant tendency, either in the case of significant bias toward SH\( \omega \) or when no significance was detected in primates, is beyond doubt. Thus, the existence of many PSGs in functional modules evolving at significant low (or no-significant) \( \omega \) ratios does not represent false positive results in the analysis of molecular adaptation. This result, registered in our data and detected in previous publications, simply suggests that PSGs are frequently recruited in the genomes for other purposes than the classical increase of rates of functional set of genes compromised in adaptive processes such as evolutionary arm-races. A possible explanation is that many of the PSGs in the genomes are changing in association with the constraints imposed by the architecture of the network [Alvarez-Ponce et al. 2009], or adjusting deleterious mutations of other genes of the network, just for the maintenance of its phenotypic function. In this sense, it is true that adaptation will requires positive selection, but the reciprocal, that would suggest that every mutation under positive selection contributes to the adaptive dynamical process of evolution of species is a dangerous approximation [Mustonen & Lässig 2009].

Currently, with the possibility of conducting analysis at the level of the genome, evolutionary biology cannot disregard major aspects of systems biology approaches that consider the modular organization of genes. With the testing
strategy presented here, we increased the statistical power for the evolutionary analysis on individual genomes and suggest that PSGs could have additional roles in the genome than the adaptive evolutionary change of phenotypes.
6. Tools, programs, methods

6.1. Overview on the detection of selective pressures at genomic level

In the pipeline leading to the identification of selective pressures acting on genes (see section: Pipeline for the detection of molecular evolution, in the Introduction), the most decisive steps are surely the first ones, the definition of homologous groups of transcripts, and their alignment. However the essentially technical aspect of this steps is often underestimated. In the sections above, we are going to review the classical methodology and propose some solutions in order to improve its accuracy.

6.1.1. The selection of homologous sequences

The first step in an analysis of selective pressure acting in a gene, is the definition of a set of homologous genes. Two parameters have to be taken into account here, the number of sequence to compare and the degree of similitude the comparison should accept. From the recommendations of Ziheng Yang in PAML’s “FAQs” [Yang 2007]:

- The number of sequences: a minimum of 4 to 5 sequences in optimal conditions optimal sequence divergence.

- The optimal divergence: The sum of $dS$ over all branches in a tree sums more than 0.5. At the other edge, the maximal divergence that can be handled to reconstruct an accurate phylogenetic tree, is assumed to be limited by the saturation of synonymous sites. However since, the emergence of maximum likelihood in phylogenetic analysis, synonymous saturation is considered to be problematic from 30-40% sequence divergence. And this degree of divergence basically moved the limitation to other problems, as the accuracy of the alignment, or the heterogeneity of nucleotide frequencies that would bias the substitution process in some species in relation to others [Yang 1998b].

In other words, being able to align fairly a set of at least 4-5 sequences summing a $dS$ over 0.5 is generally sufficient to start building a phylogenetic trees and estimate selective pressures over it.
Chapter 6. Tools, programs, methods

Over these classical considerations, another point that may be taken into account concerns the selection of the “best” transcript. The accepted approximation that consists in using the longest transcript of each gene in comparative studies, may be reconsidered in specific cases.

6.1.2. The alignment

The first aspect to consider when aligning coding sequences is its intrinsic structure, the three nucleotides constituting a codon should not be separated in the alignment process. This consideration plus the fact that alignments realized based on amino-acid sequence are generally more accurate, leads to the conclusion that a good solution to align coding-sequences is to use their translation onto protein.

When one is called upon to align sequences the variety of choice in software might be bewildering. This, specially knowing that misaligned columns have high chances to pass positive selection tests. Even if some multialignment tools seem to stand out in terms of accuracy and computation time (we can cite Muscle [Edgar 2004], MAFFT [Katoh et al. 2005], DIALIGN-TX [Subramanian et al. 2008], ProbCons [Do et al. 2005] or T-COFFEE [Notredame 2010]) the congruence between these may vary.

Some works were made in order to provide a classification of the accuracy of these tools [Plyusnin & Holm 2012] using the BAliBASE [Bahr et al. 2001] dataset. However, even if some aligners seem to globally stand out above others (ProbCons or T-Coffee), the accuracy may be highly dependent on the data used [Edgar & Batzoglou 2006].

In the pipeline leading to the identification of selective pressures among genes, the alignment step, even if crucial, is usually the fastest in terms of computational time. Thus, the solution proposed in [Edgar & Batzoglou 2006] may be taken into consideration. This solution consists in using the consensus solution of several aligners that are based in different algorithms (the example proposed in the review is T-COFFEE, ProbCons and MUSCLE). The consensus alignment can be obtained with the M-COFFEE tool for example. This strategy would lead to discard, as to variable, those sites which placement in the alignment is varying depending on the tool used.

Finally, and once obtained a fair alignment, a last step is necessary in order remove columns (or rows if possible) of the alignment that presents unrealistic variation. This step can be achieved, for example, with TrimAl [Capella-Gutiérrez et al. 2009].

6.1.3. Model testing and phylogenetic inference

In the pipeline leading to the detection of selective pressures at molecular level, the phylogenetic relationship between sequences may be known. This step however
may be skipped in the case we are working with one-to-one orthologs, and we are confident about a known species tree. Otherwise, the phylogeny must be constructed.

As with the alignment step, the reconstruction of a phylogeny is generally more accurate when dealing with amino acid sequences. The only exception being the case where sequences are too close to present differences in their amino-acid sequences.

Distance between sequences are calculated according to a model. In the case of nucleotidic models, from the most simple model Jukes and Cantor (JC) \cite{Jukes & Cantor 1969} –that assumes equal transition and transversion rates as well as equal base equilibrium frequencies–, to the most complex model, the General Time-Reversible (GTR) model \cite{Rodríguez et al. 1990} –that accounts for different rates for each kind of possible substitutions for each nucleotid–, a whole range of models can be used to weight the changes occurring between two sequences. In the case of amino-acid sequences several evolutionary models are also available, this time based on empirical data.

Whatever the case, nucleotide or amino-acid sequences, we have to decide which model is able to explain better the substitutions observed between sequences. For now, the most accepted methodology consists in comparing model’s likelihood through LRT if model are nested or their Akaike (or AIC) score \cite{Akaike 1974} otherwise (for amino-acid models).

Although this methodology is successfully finding the most likely model explaining the successive substitutions differentiating sequences, the cost in terms of computational time is high. Some approximations, however can be used in order to accelerate the process, using fixed topology build by neighbor-joining (NJ) are believed to be good approximations in order to estimate the best model \cite{Posada & Crandall 2001}.

Once identified the evolutionary model with best fit, the phylogeny can be constructed. We are not going to enter deep into this topic, however we would like to point out that, if the reconstruction fails to solve, with enough confidence, each node of the phylogeny, a good option is to keep the tree “multi-forked”. Using an approximate solution is better than betting for a solution with poor statistical support.

### 6.1.4. Testing for evolutionary scenarios in protein coding genes

Once formulated, an evolutionary hypothesis can be formulated and tested. Classical methodology for testing evolutionary hypothesis involves the use of programs like CodeML from the PAML package \cite{Yang 2007} or the SLR program \cite{Massingham & Goldman 2005} (among others). However these tools are designed to be used for the study of one or very few trees. The preparation of the data, the
configuration file and the successive test needed in order to find with confidence the optimal values of each of the parameters. At the other edge, the interpretation and integration of the results is also complicated given the quantity of information usually produced.

This section is dedicated to a solution that we brought out in order to ease the analysis of selective pressures at genomic scale, with examples of the classical evolutionary tests in protein-coding regions.

Codon substitution model

The detection of selective pressures in protein-coding regions of the genomes consists basically in the count of synonymous and non-synonymous changes between pairs of sequences. More precisely the distance between two codons is calculated according to Markov-chain model with, as state of space, the instantaneous proportion of each codon, and as substitution-rate matrix, $Q = \{q_{ij}\}$. $q_{ij}$ being the instantaneous rates from codon $i$ to codon $j$. The model commonly used is this simplification of Goldman and Yang [Goldman & Yang 1994]:

$$q_{ij} = \begin{cases} 
0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 codon positions.} \\
\pi_j, & \text{if } i \text{ and } j \text{ differ by 1 synonymous transition.} \\
\kappa \pi_j, & \text{if } i \text{ and } j \text{ differ by 1 synonymous transition.} \\
\omega \pi_j, & \text{if } i \text{ and } j \text{ differ by 1 non-synonymous transversion.} \\
\omega \kappa \pi_j, & \text{if } i \text{ and } j \text{ differ by 1 non-synonymous transition.}
\end{cases} \quad (6.1)$$

With $\pi_j$ the equilibrium frequency of codon $j$, $\kappa$ the ratio of transition over transversions (see Figure 6.1), $\omega$ the ratio of non-synonymous over synonymous mutations.

![Figure 6.1: Transition and transversion.](image)

Schema defining the different kinds of substitutions between each nucleotide.

The different codon models corresponds to different assumptions made on the distribution of equilibrium codon frequencies $\pi_j$. Most common codon models assume either 1- that each codon has the same frequency (“F1×4” in CodeML...
6.1. Overview on the detection of selective pressures at genomic level

- 1 degree of freedom) 2- codon frequency are estimated based on the observed frequencies of nucleotides (“F3×4” in CodeML – 3 degrees of freedom) 3- codon frequencies is different for each codon (“F61” in CodeML – 59 or 60 degrees of freedom).

To relate the model to real data over time (\(t\)), we need to define a transition-probability for any time, and for all possible \(i\) and \(j\) codons (\(p_{ij}(t) = Pr\{X(t) = j|X(0) = i\}\)). The matrix of this transition-probability (\(P(t)\)) can be calculated as:

\[
P(t) = \{p_{ij}(t)\} = e^{Qt}
\]

Equation 6.1 gives us the major parameters that are affecting the count of changes between 2 codons (plus the time Equation 6.2). For real data, within a phylogenetic tree, this estimation has to be done for each codon, and at each internal node. Solutions to solve this problem are, today, mainly centered around maximum likelihood methods, using Felsenstein’s pruning algorithm [Felsenstein 1981].

Overview of major/classical evolutionary models

We have seen in previous section the parameters to be considered in order to compute distances between sequences. However, the quantity of parameters needed to be estimated for each codon and at each internal node may lead to an over-fit of the model. In order to estimate the importance of the optimization of a given parameter, a likelihood-ratio test can be performed. For example, in previous section we mentioned the most frequently used codon models (F1×4, F3×4 and F61), a solution to decide which one of this model would explain with more accuracy the changes observed between codons, we can compute their likelihood over a given dataset.

An other example lies in the assumption of different selective pressures, either over branches or over sites. Selective pressures (related to the \(\omega\) ratio) can be optimized according to different hypotheses of heterogeneity. In this section we will focus on three main groups of models, site, branch and branch-site models.

Site models Site models assume that all branches of a given phylogenetic tree are evolving at the same \(\omega\) rate but allow different selective pressures to occur along the alignment. These models are useful in order to quickly see which parts of a sequence are under strong selective constraints in the studied group of species. Several strategies are available in order to obtain the shape of \(\omega\) values among sites, among them we will mention two of the most used. First, the one using a “sitewise likelihood-ratio” methodology implemented in the SLR program [Massingham & Goldman 2005] that computes a LRT at each site \(i\): \(\Lambda_i = 2 \times (l_i(1) - l_i(\hat{\omega}_i))\), with \(l_i(1)\), the likelihood of the null model assuming that \(\omega = 1\), and \(l_i(\hat{\omega}_i)\) the
likelihood of the alternative model letting the estimate of $\omega_i$ free. **Second**, the site models implemented in CodeML [Yang 2007] that are based on the definition of a prior, that segregates different categories of sites based on the distribution of the random variation of $\omega$ values (e.g.: category of sites with $\omega > 1$), followed by the assignation of a value of probability, for each site, to belong to one of these categories.

**Branch models** These models are orthogonal to site models in the sense that $\omega$ rate is not allowed to vary along the alignment, while it can be estimated independently for each branch of the phylogeny. Most simple (and most unrealistic) branch model, the “$M0$” model, assumes that all branches are evolving at the same rate (and same for sites as it is a branch model), calculating a unique value of $\omega$ for the whole phylogeny. At the other edge the most complex model, the “free-ratio” model estimates a different value of $\omega$ for each branch of the phylogeny. Also, both $M0$ and free-ratio models might be either unrealistic or a classical case of over-fitting model, they are useful as respective null or alternative model when testing for a given evolutionary scenarios.

Overall, the assumption made by branch models are dangerous as high values of $\omega$ detected are usually due to the counterbalanced effect of conserved and accelerated sites. However, it is a useful for detecting differences between the evolutionary rates of sequences taken either individually, or grouped together (e.g. in clades). A classical example of the use of these models, is the detection of different selective pressures occurring between colobines and hominids’ lysozyme protein [Yang 1998a].

**Branch-site models** As the name suggests, these models represent a compromise between previous two groups of models. More precisely, phylogenetic information is used to contrast differences in rates of a given site. This model is usually used to detect, or contrast, sites under a characteristic selective pressure. The contrast being done between to parts of a phylogenetic tree, generally referred to as foreground and background branches. Foreground branches having an extra class of sites allowing $\omega$ to be higher than 1. In a classic branch-site model (branch-site A in CodeML), foreground branches are fitted to a model similar to the $M2$ model, while background branches to a model similar to $M1$ model. These models are more realistic then sites or branch models, and are the most reliable for tests of positive selection.

**Testing for the best evolutionary model**

Beyond the descriptive use of the models described above, models can be compared through LRT in order to evaluate –with statistical significance– the importance of
the optimization of a given parameter. This methodology, for example, has been extensively used in each of the groups of models described above in the context of positive selection test. Here are listed some examples of the most popular tests that can be done by comparing the fit of different evolutionary models.

The comparison of the fit of the evolutionary models reviewed above to a given dataset by mean of their likelihood.

**Test of positive selection and relaxation in sites** As we have seen above, in CodeML, a site analysis basically consists in classifying all the sites in an alignment into the categories of sites defined by the model. In the context of the identification of positively selected sites, model can be classified in 2 categories, neutral models and positive-selection models. The difference between them being that positive-selection models have an extra category of sites that allows \( \omega \) to be higher than 1. Models, being nested, a LRT can be conducted between them (with usually 1 degree of freedom that corresponds to the estimation of the extra class of sites). If positive-selection model wins, than sites that belongs to last category (with \( \omega > 1 \)) are considered to be truly under positive selection.

A classical positive selection test is done by comparing model M1a (neutral, with 2 categories of sites: \( 0 < \omega_0 < 1 \) and \( \omega_1 = 1 \)) with model M2a (positive selection, with same categories of M1a plus \( \omega_2 > 1 \)).

**Test of different selective regimes in branches** These models are perhaps the most unrealistic at the time to test for positive selection, however they still are useful to compare and differentiate selective regimes in branches. For example, one can formulate the hypothesis that a given clade is undergoing an acceleration in mutation rate compared to the rest of the phylogeny, and test this hypothesis.

![Figure 6.2.](image)

Figure 6.2.: Simple example of allowing / disallowing different \( \omega \) rates in tree.

A simple phylogeny is represented here, the colors of the branches represents the different estimations of \( \omega \). In (A), the \( \omega \) value of all the branches is the same \( (\omega_O = \omega_C = \omega_H) \), in (B) the \( \omega_O \) of the Orangutan, is different from the rest of the tree \( (\omega_O \neq \omega_C = \omega_H) \), and in (C) each branch has a different value of \( \omega \) \( (\omega_O \neq \omega_C \neq \omega_H) \). Figure adapted from [Yang 2006].
Chapter 6. Tools, programs, methods

Figure 6.2 is an example of methodology over a simple phylogeny representing 3 sequences of Human, Chimp and Orangutan. Each of (A), (B) and (C) in the figure, represents different branch models. (A) is the simplest branch model, one $\omega$ is estimated for all branches, in (B) 2 values of $\omega$ are estimated (we thus have one extra parameter in comparison with (A)) and in (C) 3 values of $\omega$ are calculated (again one parameter more than previous model). The comparison by LRT between models (A) and (B) would thus determine if the evolutionary rate calculated for Orangutan sequence is significantly different from the rest of the tree. And, in the same way, the comparison between (B) and (C) is indeed testing if each branch is evolving at different rates.

Test of positive selection and relaxation in sites of a given set of branches
This test is perhaps the most sensitive for detecting positive selection in protein-coding genes. We have seen that, in branch tests, one protein would be detected to be under positive selection only if the average $\omega$ over all sites is higher than 1, however we can easily imagine the case where only few sites are evolving fast in a context of global purifying selection for the rest of the protein. In the same way, in the case of site tests, the $\omega$ value of a given site is averaged over all branches. However both of these tests were successful to detect positive selection in protein-coding genes, we could expect that in most of the cases significant accelerations of $dN$ in relation to $dS$ would only affect some sites in a given lineage. This is the main reason that pushes Yang and Nielsen [Yang & Nielsen 2002] to implement a new test for positive selection able to detect few sites in a particular lineage. Originally the test consists in comparing the branch-site model A (bsA) to the model M1a – this test is often referred to as “test I”. From the specification of bsA model (see Table 6.1), we see that the only difference with model M1a, is the presence of sites evolving at $\omega_2$ rate in foreground branches.

<table>
<thead>
<tr>
<th>Site class</th>
<th>Proportion</th>
<th>Background $\omega$</th>
<th>Foreground $\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$p_0$</td>
<td>$0 &lt; \omega_0 &lt; 1$</td>
<td>$0 &lt; \omega_0 &lt; 1$</td>
</tr>
<tr>
<td>1</td>
<td>$p_1$</td>
<td>$\omega_1 = 1$</td>
<td>$\omega_1 = 1$</td>
</tr>
<tr>
<td>2a</td>
<td>$(1 - p_0 - p_1)p_0/(p_0 + p_1)$</td>
<td>$0 &lt; \omega_0 &lt; 1$</td>
<td>$\omega_2 &gt; 1$</td>
</tr>
<tr>
<td>2b</td>
<td>$(1 - p_0 - p_1)p_1/(p_0 + p_1)$</td>
<td>$\omega_1 = 1$</td>
<td>$\omega_2 &gt; 1$</td>
</tr>
</tbody>
</table>

Table 6.1.: $\omega$ ratios assumed in model branch-site A.
In a phylogenetic tree where branches are divided in 2 categories, “background” and “foreground”, in order to allow over them different selective pressures, sites are allowed to evolve at $\omega$ rates higher than 1 only in foreground branches. $p_0$ and $p_1$ are the proportion of sites evolving respectively at rates $\omega_0$ and $\omega_1$.

Even if theoretically the test between bsA and M1a (test I) seems robust, simulations study found high proportions of false positive [Zhang 2004], and conducted to the implementation of new test [Zhang et al. 2005b]. This new branch-site tests (referred to as “test II”) consists in the comparison of 2 branch-site models, one
with \( \omega_2 > 1 \) (bsA) and the other, namely branch-site A1 (bsA1) with \( \omega_2 = 1 \). bsA1 being here the null model. A gene would be considered to be under positive selection if the fit of bsA model is better than bsA1.

As, the first branch-site test (test I) is considered to bring high proportion of false positives, it is often used in order to detect relaxed constraints in genes (see Figure 6.3). Thus, a protein-coding gene should be considered to be under positive selection if it passes test I, and considered to be under relaxed constraints if it passes test II but not test I.

![Branch-site tests for the detection of positive selection and relaxation.](image)

Figure 6.3.: Branch-site tests for the detection of positive selection and relaxation.

Circles here represent positive results of the LRT between branch-site model A (bsA) versus branch-site model A1 (bsA1) ("Test II") on the left, and bsA versus site model M1a on the right ("Test I"). The color represents the set of genes that should be considered to be under positive selection (light gray–left) or under relaxation (dark gray–right).

### 6.2. ETE’s Evol extension

In previous sections we have seen different evolutionary models and tests, mainly defined in the CodeML program, however it is important to contrast that these models are standards and other programs or packages offers different implementations to achieve similar tests [Knight et al. 2007, Pond et al. 2005].

This section is dedicated to what can be either a genomic solution or at least a simplification of the methodology for the determination of selective pressures and the test of evolutionary hypotheses. This solution passes through the use of ETE [Huerta-Cepas et al. 2010], and more particularly of the “Evol” extension, that allows to call CodeML and SLR programs. Results of such analysis are embedded in ETE’s Tree objects and from there can be contrasted (evolutionary model comparison), visualized or summarized.

#### 6.2.1. Implementation

ETE is a python package originally designed for the manipulation, analysis and visualization of phylogenetic trees. Ahead of the most general Tree class, stands,
Chapter 6. Tools, programs, methods

among others, the PhyloTree (from the “Phylo” extension) that implements specific algorithms to deal with phylogenetic trees. As most useful functions, we can cite the abilities to:

- link a tree to an alignment
- infer evolutionary events (speciation or duplication) through different algorithms
- relatively date nodes of a tree
- automatically root a gene tree (according to a given species tree)

In order to take advantage of these functions, and also for coherence reasons, the Evol extension was implemented as a specific case of the Phylo extension. Or in a more computational language, the EvolTree `class` inherits from PhyloTree `class` (note that in the same way, PhyloTree `class` inherits from the main Tree `class`).

The Evol extension contains two main `classes`, the EvolTree that as we have seen inherits from the PhyloTree `classes`, and the Model `classes` that represents a given evolutionary model (for now, it can be either one of the models proposed by CodeML, or by SLR).

The Evol extension, is available at https://github.com/jhcepas/ete/tree/evoltree, for now, as a branch of ETE. Some documentation can be found at http://bioinfo.cipf.es/fransua/ete-evol/tutorial/tutorial_adaptation.html.

6.2.2. General usage

In this section we are going to quickly overview how to use the Evol extension. Evol trees and alignments are loaded just as PhyloTrees. Here is a short example on how to load a tree together with its alignment and run the free-ratio model:

```python
from ete2 import EvolTree
from ete_dev import EvolTree
tree = EvolTree("(Orangutan,Human,Chimp);")
tree.link_to_alignment(""
>Chimp
GCC GCA CGA TGG CTC AAT GTA AAG TTA AGA TGC GAA TTG AGA ACA CTA AAA AAA
TTG GGA CTG GAC GCC TAC AAG GCA GTA AGT CAA TAC GTT AAA GGT CGT GCC ATT
>Orangutan
GAT GCA CGA TGG ATC AAT CGA AAC TTA AGA TGC GAA TTG AGA ACT CTG AAA AAA
TTG GGA CTG GAC GCC TAC AAG GCA GTA AGT CAA TAC GTT AAA GGT CGT AGC TCT
>Human
TAC GCA CGA TGG CTG AAC GTA AAA TTA AGA TGC GAA GTA TTA ACG TCT AAA AAA
TTG GGA CTG GAC GCC TAC AAG GCA GTA AGT CAA TAC GTT CAA GGT CGT GCC AGT """)
tree.run_model("fb")
tree.show()
```
The “show” command here works as for PhyloTrees but, additionally, it displays a summary of the selective pressures acting on branches (colored circles appearing at each node in Figure 6.4 which colors and sizes are a function of the corresponding values of $\omega$ estimated by the model).

![Figure 6.4](image)

**Figure 6.4:** Sample representation of the free-ratio model in an EvolTree.

Default representation of an EvolTree. Node’s sizes and colors are function of the values of $\omega$ for the given branch. For this example, the values found for the sequences under free-ratio model are, for Orangutan $\omega = 1.29$, Chimp $\omega = 0.28$ and Human $\omega = 0.08$.

Also the example above may be “too simple” when compared to the number of parameters that have to been set when using directly the CodeML program, the extension allows to modify each one of the parameter proposed by CodeML. As an example, this is the list of default parameters set in the Evol extension for the free-ratio model:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>aaDist</td>
<td>0</td>
</tr>
<tr>
<td>fix_alpha</td>
<td>1</td>
</tr>
<tr>
<td>alpha</td>
<td>0.0</td>
</tr>
<tr>
<td>fix_blength</td>
<td>0</td>
</tr>
<tr>
<td>cleandata</td>
<td>0</td>
</tr>
<tr>
<td>CodonFreq</td>
<td>2</td>
</tr>
<tr>
<td>getSE</td>
<td>0</td>
</tr>
<tr>
<td>icode</td>
<td>0</td>
</tr>
<tr>
<td>fix_kappa</td>
<td>0</td>
</tr>
<tr>
<td>kappa</td>
<td>2</td>
</tr>
<tr>
<td>Malpha</td>
<td>0</td>
</tr>
<tr>
<td>method</td>
<td>0</td>
</tr>
<tr>
<td>Mgene</td>
<td>0</td>
</tr>
<tr>
<td>model</td>
<td>1</td>
</tr>
<tr>
<td>ncatG</td>
<td>8</td>
</tr>
<tr>
<td>noisy</td>
<td>0</td>
</tr>
<tr>
<td>NSsites</td>
<td>0</td>
</tr>
<tr>
<td>fix_omega</td>
<td>0</td>
</tr>
<tr>
<td>fix_omega</td>
<td>0</td>
</tr>
<tr>
<td>omega</td>
<td>0.7</td>
</tr>
<tr>
<td>RateAncestor</td>
<td>0</td>
</tr>
<tr>
<td>runmode</td>
<td>0</td>
</tr>
<tr>
<td>seqtype</td>
<td>1</td>
</tr>
<tr>
<td>Small_Diff</td>
<td>1e-6</td>
</tr>
</tbody>
</table>

Changing one of this parameter can be done easily, suppose we want to set a different starting value for the optimization of $\omega$ (as shown in previous box, by default it is set to 0.7):

```
tree.run("fb", omega=1.2)
```

In the context of site analysis, the Evol extension also have some features in order to summarize visually the shape of selective pressure along sites. In order to evaluate the selective pressures among sites, we first need to validate positive selection model as being the model with the best fit:

```
tree.run_model("M2")
tree.run_model("M1")
tree.get_most_likely("M2", "M1")
```

Following with the example, the result of the “get_most_likely” –the p-value of the LRT between the models M1a and M2a– is below 0.05 (P=0.014). Thus “M2a” has a better fit, and we can accept the assumption that some site are
evolving at $\omega > 1$. As previously mentioned, an other advantage of the Evol extension, lies in the possibility to get an overview of the $\omega$ shape varying along the alignment according to a given evolutionary model (see upper plot in Figure 6.5).

![Omega shape over sites, model M2a](image1)

![Omega shape over sites, model SLR](image2)

**Figure 6.5.: Representation of model M2a with CodeML and SLR result.**

Same representation as in Figure 6.4, with two additional bar plots representing the values of $\omega$ for each site of the alignment. The plot on the top represents the values of $\omega$ computed under M2a model with the CodeML program, while the plot above sequences represents the values of omega calculated by the SLR program. Colors of the bars represent the significance of belonging to a given class of sites (red: belongs significantly (probability $> 0.99$) to class of sites with $\omega > 1$, orange: belongs significantly (probability $> 0.95$) to class of sites with $\omega > 1$, blue/cyan: belongs significantly (probability $> 0.95$) to class of sites with $\omega < 1$).

In order to confirm the positively selected sites detected by the CodeML program, a good option is to use an other methodology, as the sitewise likelihood-ratio implemented in the SLR program. Through the Evol extension, running SLR program, and displaying a summary result of both, M2a model of CodeML and SLR, would be something like this:

```python
  tree.run_model("SLR")
  tree.show (histfaces=['M2', 'SLR'])
```

Thus, according to the p-value of the LRT between models “M2a” and “M1a”, and to Figure 6.5, we can deduce that for our sample alignment we have 2 sites under positive selection, of which 1 is confirmed by the SLR program.
6.2. ETE’s Evol extension

More complex branch models can also be easily defined in order to test different evolutionary hypotheses. For example, if we want to test for significant differences in \( \omega \) rates between branches, we will have to mark the tree and compute specific branch models as explained in previous section about Branch models page 96.

Each node of an Evol trees is label with an ID that corresponds to the identifier given by CodeML (namely the “node_id”). This label are useful to mark a tree. Marks are also inspired in CodeML methodology, that is a hash symbol (#) followed by a number. To test the hypothesis that Orangutan is significantly evolving at higher rate than Human and Chimp together:

```plaintext
  tree.mark_tree([(tree & 'Orangutan').node_id], marks=['#1'])
  tree.write()
```

The “write” command is here to check that our tree has been actually marked (for this example it should be: “(Chimp,Orangutan #1,Human);”). The next step consists in fitting this marked tree to a branch model, and compare it to a null model that can be the “M0” model, were all branches have the same value of \( \omega \):

```plaintext
  tree.run_model("b_free")
  tree.mark_tree([(tree & 'Orangutan').node_id], marks=['#'])
  tree.run_model("M0")
  tree.get_most_likely("b_free", "M0")
```

The p-value of this LRT would be in this case below 0.05 (P=0.021), thus Orangutan seems to be evolving at a different rate than the other species. In order to see quickly the values estimated for each part of tree, a summary can be displayed by printing (through the “print” function) our branch model:

```plaintext
  print tree.get_evol_model("b_free")
```

Output being:

```
Evolutionary Model b_free:
  log likelihood : -244.804192
  number of parameters : 6
  sites inference : None
  sites classes : None
  branches :
    mark: #1 , omega: 1.39095890112 , nodes paml_ids: 3
    mark: #0 , omega: 0.123723533087 , nodes paml_ids: 1 2
```

According to this summary, and to the result of LRT, branches marked with “#1” have a significantly higher value of \( \omega \).

Finally, the branch-site test –usually considered the ultimate test of positive selection– can be computed in a similar way. In this case, we also have to mark the tree in order to differentiate which branches should be considered as foregroud
or background branches in branch-site model A (bsA) and branch-site model A1 (bsA1) (see section 6.1.4). Lastly we can perform a LRT between these optimized models. As an example, we can apply the branch-site test (test II) in order to see of some sites of our Orangutan’s sequence, are under positive selection.

```python
tree.mark_tree([(tree & 'Orangutan').node_id], marks=['#1'])
tree.run_model("bsA")
tree.run_model("bsA1")
tree.get_most_likely("bsA", "bsA1")
```

The p-value of this LRT is 0.005, allowing thus, the presence of sites evolving at $\omega > 1$ specifically in Orangutan’s sequence.

### 6.3. Phylemon2.0

More than a side-product Phylemon raises from the need of evolutionary biologists to easily run, store and compare computational analysis over their data. However, in its last version [Sánchez et al. 2011], some improvements were brought based on, 1- the feedback from users and fellows, 2- the advent of new tools in the fields of phylogenetics and study of adaptation, 3- our experience as advanced users of the whole pipeline leading to the detection of selective pressures in protein-coding genes.

As with previous version, Phylemon is organized in sections, that in their general lines, are following the steps leading to the detection of selective pressures in protein-coding genes we described above (section: Testing for evolutionary scenarios in protein coding genes, page 93). In this final part of the thesis, we are going to review quickly some of the advances brought by this new version of Phylemon.

#### 6.3.1. Alignment

The main improvements in Phylemon2.0 around the process of aligning sequences consists basically in the addition of the tools Lagan and M-Lagan [Brudno et al. 2003] particularly useful when dealing with long genomic sequences.

Also an effort was made in the sense of the recommendations mentioned in section: The alignment, page 92. In order to ease the achievement of this step we made available a new version of the CDS-protAl tool, that was completely re-written in order to be more efficient and stable, it also offers more options to users. Through the user-friendly web interface, the user can align coding sequences according to their translated amino-acid sequences. It is important to notice that the options proposed in Phylemon2.0 are few in comparison of what the program can do through command-line. CDS-protAl is indeed able to apply the full methodology described bellow, allowing to align a given set of sequences
6.3. Phylemon2.0

with different tools, merge the resulting alignments, and finally clean them using TrimAl. It is available at.

6.3.2. Model selection

In the context of model testing Phylemon2.0 made available several tools. Most famous being ModelTest [Posada & Crandall 1998] direct successors: jModelTest [Posada 2008] for nucleotide sequences, and ProtTest [Abascal et al. 2005] in the case of amino-acids. Additionally a new tool, PhyML-Best-AIC-tree, was implemented paying a special attention to the needs of genomic studies. PhyML-Best-AIC-tree is basically a simplification of jModelTest and ProtTest together, it can deal either with nucleotide or amino-acid sequences but only computes AIC scores. It is specifically designed thinking of its integration in a pipeline. As main feature it has the option to search for the best substitution model in an “clever” way. As we have seen in section: Model testing and phylogenetic inference page 92, a good approximation in the search for the best substitution model, is to test the fit of each model over a fixed tree. PhyML-Best-AIC-tree can be used in order to compute first this fast approximation for all models, and in second time, run a more precise analysis that involves the optimization of the tree topology but only for the models summing a given weight. The weight of a model being defined as in jModelTest or ProtTest. As for CDS-protAl, more options are available through command line, and moreover, in the case of PhyML-Best-AIC-tree, a special attention was paid to ease its integration in a pipeline, and all functions can be called independently from another python program – skipping thus the command line. The program source code is available at.

6.3.3. Phylogeny

In this section, the main advances were done around PhyML [Guindon & Gascuel 2003] and MrBayes [Ronquist & Huelsenbeck 2003] and consists mainly in an extension of the forms with more options to run these tools and a better integration of their output files in Phylemon framework. MrBayes in particular has now the option to be run non-interactively, and also to build a command block through Phylemon’s form. Output trees found can be now browsed directly either in ETE [Huerta-Cepas et al. 2010] or in Archaeopteryx [Zmasek 2012].

6.3.4. The Pipeliner

This is perhaps one of the greatest improvement in this new release of Phylemon. Phylemon’s Pipeliner is a tool designed for enabling users to develop their own pipelines in a friendly modular environment allowing to run multiple gene analyses. The pipeline covers, with a selection of tools, all the steps needed to transform
Chapter 6. Tools, programs, methods

from a given set of sequences to one, or several (as it accepts multiple input files) worthy phylogenetic tree. Moreover, any pipeline created by placing and linking different tools on the Pipeliner “playground” can be saved for further analysis, or even exchanged with a collaborator.

6.4. Discussion

In this last chapter we have overviewed some of the contributions brought as side-products of different works in the context of testing specific evolutionary hypotheses [Lavagnino et al. 2012, Martín-Trillo et al. 2011, Serra et al. 2011, Lüke et al. 2011, Gonçalves et al. 2011]. These contributions lie along the pipeline leading to the detection of selective pressures in coding sequences and, most of them consists in very little changes in the classical approaches, nevertheless an effort was made in order to fully integrate them into functional programs and in some cases into a web server.

In all cases, a special attention was paid to the scalability, and to facilitate the use for both, scientists with or without experience in computer sciences. The programming language mainly used was Python [Van Rossum & Drake 2003], and the choice of this programming language is the result of the balance between 4 major points i) the interaction with packages implemented in R [Team 2011] – python allows to call easily R function [Moreira & Warnes 2004], ii) computation speed versus flexibility – python a good compromise between the performance of C or java and the flexibility of R or perl [Fourment & Gillings 2008], iii) interactivity, python shell as R shell allows to build an analysis line by line skipping the need of writing a complete working script, iv) its raising popularity among the bioinformatics community [Bassi 2007].

Finally, I would like to highlight an extra aspect that was considered at the time of implementing these tools, the documentation. A special attention was paid to make these tools scalable and usable, and this effort passes necessarily through the development of comprehensive tutorials and, for advanced users annotation inside the source code. In Phylemon, a full help page is, since version 2.0, describing all the tools and proposing simple exercises in order to better understand the interest and needs of each tool. The Ete’s Evol extension, is fully integrated inside ETE, and extends its documentation with specific tutorial, moreover, its source code is fully commented in order to ease the implementation of future extensions and its call from other programs.
7. Conclusions

- In the whole diversity of life, from viruses to mammals, a universal law governs the structure of genomes. Only dramatic changes—like can be polyploidization events or strong biases in nucleotidic contents— are able to lower genome entropy.

- We hypothesise that this law obligates raises in biological complexity to be preceded by genome expansions.

- The dynamics of distribution and abundances of genetic elements—either functional or repetitive— in eukaryotic genomes are following the expectations of a neutral ecological model.

- GSSA allows to test for functional biases within fast or slow evolving genes. This methodology successfully identified all previously reported functional categories candidates to be important targets of natural selection. Moreover, given that the GSSA is not limited by the compulsory presence of PSGs, it extended this set of targets to previously undetectable ones.

- We present a new argument supporting that the role of PSGs is not only the adaptive evolutionary change of phenotypes.

- The implementation of tools and programs that were used to generate the results presented in this thesis, were implemented following specific programming guidelines in order to maximize their contribution to scientific community.
Bibliography


Bibliography


[Cavalier-Smith 2005] THOMAS CAVALIER-SMITH, Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Annals of botany 95(1) (2005), 147–75. $\rightarrow$ page 73


Bibliography


[Eichinger & Noegel 2003] Ludwig Eichinger and Angelika A Noegel, Crawling into a new era—the Dictyostelium genome project. The EMBO journal 22(9) (2003), 1941–6. → page 50


Bibliography


VI


Bibliography


[He et al. 2010] Xionglei He, Wenfeng Qian, Zhi Wang, Ying Li, and Jianzhi Zhang, Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nature genetics 42(3) (2010), 272–6. → page 87


[Jabot & Chave 2011] Franck Jabot and Jérôme Chave, Analyzing Tropical Forest Tree Species Abundance Distributions Using a Nonneutral Model and through Approximate Bayesian Inference. The American naturalist 178(2) (2011), E37–47. → pages 11, 28, 32, and 35

Bibliography


Bibliography


[Massingham & Goldman 2005]  TIM MASSINGHAM AND NICK GOLDMAN, Detecting amino acid sites under positive selection and purifying selection. Genetics 169(3) (2005), 1753–62.  ⇒ pages 18, 93, and 95


Bibliography


XII
Bibliography


XIII
Bibliography


XIV


Bibliography


XVI


List of Figures

1.1. Overview of Tree of Life .................................................. 5
1.2. C-values of the main groups of life ............................... 6
1.3. Genomic components of human genome .............................. 8
1.4. Relative amount of retrotransposon and DNA transposon in diverse eukaryotic genomes ................................................. 9
1.5. Cloud of Functional categories enriched in PSGs in first genomic studies. ............................................................... 14

2.1. Maximum likelihood inference of neutral parameters. ........... 31
2.2. Comparing simulated and empirical evenness ..................... 33
2.3. Type I and Type II errors of neutral test in ranges of S and J ......................................................... 35
2.4. Mammals and Drosophila phylogeny ............................... 36
2.5. Summary of the steps developed by the GSSA. .................. 38
2.6. Randomization experiment diagram. ............................ 41
2.7. Randomization experiment results. .............................. 42
2.8. Evolutionary and statistical simulation of GSSA. ........... 43

3.1. Genome complexity value ........................................... 50
3.2. Chromosome complexity ratio ..................................... 53
3.3. Sliding window analysis in chromosomes ........................................... 57
3.4. Sliding window in a full chromosome ....................................... 57
3.5. Return to maximum complexity after polyploidization ........ 59

4.1. Proportion of major genetic species (GSs) in 31 eukaryotes. ....... 66
4.2. RSA curves. ............................................................ 68
4.3. Relative species abundance curves for human chromosome 1 and chromosome 19 ................................................. 69
4.4. Species chromosome size relationship ............................... 70

5.1. GSSA of evolutionary variables. .................................. 79
5.2. Positive selection in functional modules’ evolutionary scenarios. .. 85

6.1. Transition and transversion ........................................... 94
6.2. Simple example of allowing / disallowing different \( \omega \) rates in tree. ......................................................... 97
6.3. Branch-site tests for the detection of positive selection and relaxation. 99
List of Figures

6.4. Sample representation of the free-ratio model in an EvolTree. . . . 101
6.5. Representation of model M2a with CodeML and SLR result. . . . 102
# List of Tables

1.1. Analogies between population genetics models of TEs dynamics and inter-specific relationships in ecology. .................................................. 11  

2.1. CR explained by example ................................................................. 22  
2.2. superfamily of repetitive elements and description .............................. 26  
2.3. Biotypes and description .................................................................. 27  
2.4. Transformation of Chromosome size. .................................................. 27  
2.5. Comparison of partitioning strategies. ............................................... 40  
2.6. Number of positively selected and relaxed genes in each of the simulated evolutionary scenarios ......................................................... 44  
2.7. Proportion of significant functional categories that are still significant. 44  

3.1. Genomes Complexity. ...................................................................... 49  
3.2. Human language Complexity ............................................................ 52  
3.3. Mean complexity ratio of some genome components .......................... 55  
3.4. Complexity ratio of genome classes concatenated and shuffled ...... 55  

4.1. Result of the fit to UNTB and test of neutrality for selected chromosomes .................................................................................. 72  

5.1. Numbers and percentages of functional modules with significant results after GSSA. ................................................................. 77  
5.2. Functional enrichment results using gene-by-gene and gene-set approaches. ................................................................. 83  

6.1. $\omega$ ratios assumed in model branch-site A ...................................... 98
Glossary

**Burrows-Wheeler-transform**, also called block-sorting compression, is an algorithm used in data compression techniques such as bzip2. It is based on sorting all possible rotations of a given string, sort the result in lexicographic order and finally take the last character of each rotated string. 21

**C-value** it refers to the amount of DNA contained within haploid nucleus. Its unit is the picogram (pg). 4–7

**class** In computer science, a class is the description of the characteristics defining an object. Basically the class is what is written in the program while the object is the result of the execution of a class. XXIII, 99, 100

**de Bruijn ~-sequence** these is the mathematically defined string of characters with perfect equal frequency of sub-sequences: every possible combination of logarithmic length appears exactly once as a sequence of consecutive symbols. 60

**ecological niche** The role of a species of organisms in an ecological community, defined by the resources that the species requires from its environment. The ”competitive exclusion principle” implies that species can only stably coexist if they have different ecological niches. 10, 11

**LINE** A long interspersed element sequence - typically used for non-long terminal repeat retrotransposons. 11, 23, 26, 54–56, 66

**LTR** Long Terminal Repeat, a kind of retrotransposons with direct repeats of 300-500bp of DNA at each end of the element. These sequences resemble the integrated proviruses of retroviruses. 11, 23, 26, 54–56, 66

**object** In computer science, an object is a symbolic container with its own being defined in a class. An object can incorporate data and methods in relation to something of the real world manipulated in a computer program. XXIII, 99
patch ecological ~, a homogeneous area with a given shape and spatial configuration differing from the rest of the ecosystem. It is the lowest unit of a landscape. 11

retroposon A mobile DNA sequence that can move to new locations through an RNA intermediate. XXIV, 11

retrotransposon An autonomous transposable element that can move to a new location through an RNA intermediate. Two major classes of retrotransposons exist, with or without long terminal repeats (see LTR and non-LTR). 9, 11, 66

satellite A kind of tandem repeats, larger than minisatellites (10-60 bp) or microsatellites (2-6 bp). 11, 12, 23, 26, 54–56, 61, 66, 69

script It is a program written for a software environment that automate the execution of tasks which could alternatively be executed one by one by a human operator. 28, 106

seed -sequence of a gene or a protein, is the sequence used as starting point in the search of homologous sequences within a given set of entries. Extending this concept at genomic level, we can talk about seed-genome or seed-species. Note: in a phylome, it is expected to observe an over-representation of proteins belonging from the seed-species. -species in the case of ortholog retrieval a seed species is the equivalent of a seed sequence. 17, 34, 36

selfish DNA Sequences of DNA that accumulate in the genome through non-selective means, and which have a negative effect on the fitness of their hosts. 11

SINE A short interspersed element sequence - this is a retroposon sequence of less than 500 bp in length that does not encode the protein activities required for its movement. 11, 23, 26, 54–56, 61, 66

superfamily Transposable elements’, it corresponds to the fourth level in the classification of transposable elements according to http://www.bioinformatics.org/wikiposon/doku.php?id=main. This level will gather elements based on the structure of the internal sequence. 12, 25, 26

tandem repeat Repetitive sequence of the DNA, constituted by a pattern of 2 or more nucleotides placed one beside the other.. XXIV, 66
**transposon** A mobile DNA sequence that moves to new genomic locations through a DNA route, rather than through an RNA intermediate. This movement is catalysed by the action of a transposase protein that is encoded by an autonomous element. 9, 11, 12, 23

**trophic** Of or involving the feeding habits or food relationship of different organisms in a food chain. 9
### A. RepeatMasker summary output

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of Elements</th>
<th>Length</th>
<th>Percentage of Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SINEs</td>
<td>1658864</td>
<td>385270856</td>
<td>12.45%</td>
</tr>
<tr>
<td>ALUs</td>
<td>1136457</td>
<td>306395826</td>
<td>9.90%</td>
</tr>
<tr>
<td>MIRs</td>
<td>517233</td>
<td>78244089</td>
<td>2.53%</td>
</tr>
<tr>
<td>LINEs</td>
<td>913889</td>
<td>609952196</td>
<td>19.70%</td>
</tr>
<tr>
<td>LINE1</td>
<td>539553</td>
<td>503348534</td>
<td>16.26%</td>
</tr>
<tr>
<td>LINE2</td>
<td>319303</td>
<td>93411598</td>
<td>3.02%</td>
</tr>
<tr>
<td>L3/CR1</td>
<td>42713</td>
<td>10009516</td>
<td>0.32%</td>
</tr>
<tr>
<td>LTR elements</td>
<td>487433</td>
<td>259122242</td>
<td>8.37%</td>
</tr>
<tr>
<td>ERVL</td>
<td>108675</td>
<td>55875700</td>
<td>1.80%</td>
</tr>
<tr>
<td>ERVL-MaLRs</td>
<td>247590</td>
<td>108138874</td>
<td>3.49%</td>
</tr>
<tr>
<td>ERV_classI</td>
<td>109816</td>
<td>82706444</td>
<td>2.67%</td>
</tr>
<tr>
<td>ERV_classII</td>
<td>7480</td>
<td>8820605</td>
<td>0.28%</td>
</tr>
<tr>
<td>DNA elements</td>
<td>383832</td>
<td>95646896</td>
<td>3.09%</td>
</tr>
<tr>
<td>hAT-Charlie</td>
<td>214295</td>
<td>43419001</td>
<td>1.40%</td>
</tr>
<tr>
<td>TcMar-Tigger</td>
<td>82218</td>
<td>33550442</td>
<td>1.08%</td>
</tr>
<tr>
<td>Unclassified</td>
<td>9962</td>
<td>5418573</td>
<td>0.18%</td>
</tr>
<tr>
<td>Total interspersed repeats</td>
<td>1355410763</td>
<td>43.78%</td>
<td></td>
</tr>
</tbody>
</table>

**Small RNA:**
- **13482** 1443809 bp 0.05%
- **Satellites:** 4502 12381861 bp 0.40%
- **Simple repeats:** 403012 25937716 bp 0.84%
- **Low complexity:** 393080 17947554 bp 0.58%

*most repeats fragmented by insertions or deletions have been counted as one element*

The query species was assumed to be homo sapiens
RepeatMasker version open-3.3.0, default mode
run with rmblastn version: 2.2.23+
RepBase Update 20110419, RM database version 20110419